Spring 5-11-2019

New Methodologies for Examining and Supporting Student Reasoning in Physics

John C. Speirs
University of Maine, john.c.speirs@maine.edu

NEW METHODOLOGIES FOR EXAMINING AND SUPPORTING STUDENT REASONING IN PHYSICS

By

J. Caleb Speirs

B.S. Colorado School of Mines, 2011

M.S. Colorado School of Mines, 2012

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Physics)

The Graduate School

The University of Maine

May 2019

Advisory Committee:

MacKenzie R. Stetzer, Associate Professor of Physics, Advisor

John R. Thompson, Professor of Physics

Michael C. Wittmann, Professor of Physics

Robert W. Meulenberg, Associate Professor of Physics

Natasha M. Speer, Associate Professor of Mathematics

NEW METHODOLOGIES FOR EXAMINING AND SUPPORTING STUDENT REASONING IN PHYSICS

By J. Caleb Speirs

Thesis Advisor: Dr. MacKenzie R. Stetzer

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Physics)
May 2019

Learning how to reason productively is an essential goal of an undergraduate education in any STEM-related discipline. Many non-physics STEM majors are required to take introductory physics as part of their undergraduate programs. While certain physics concepts and principles may be of use to these students in their future academic careers and beyond, many will not. Rather, it is often expected that the most valuable and long-lasting learning outcomes from a physics course will be a repertoire of problem-solving strategies, a familiarity with mathematizing real-world situations, and the development of a strong set of qualitative inferential reasoning skills.

For more than 40 years, the physics education research community has produced many research-based instructional materials that have been shown to improve student conceptual understanding and other targeted learning outcomes (*e.g.,* problem solving). It is often tacitly assumed that such

materials also improve students' qualitative reasoning skills, but there is no documented evidence of this, to date, in the literature. Furthermore, a growing body of research has revealed that a focus on conceptual understanding does not always result in the anticipated performance outcomes. Indeed, students may demonstrate solid conceptual understanding on one physics question but fail to demonstrate that same understanding on a closely related question. This body of research suggests that reasoning processes general to all humans (*i.e., domain-general* processes) may impact how students understand and reason with physics concepts. Methodologies that separate (to the degree possible) the reasoning involved in a physics problem from the conceptual understanding necessary to correctly answer that problem are necessary for gaining insight into how conceptual understanding and domain-general reasoning processes interact.

In order to explore such research questions, new research tools and analysis methodologies are required. Physics education researchers pursuing these questions have begun to embrace data-collection methodologies outside of the written free-response questions and think-aloud interviews that are ubiquitous in discipline-based education research. Some of these researchers have also begun to utilize dual-process theories of reasoning (DPToR) as an analysis framework. Dual-process theories arise from findings in cognitive science, social psychology, and the psychology of reasoning. These theories tend to be mechanistic in nature; as such, they provide a framework that can be prescriptive rather than solely descriptive, thereby providing a theoretical basis for examining the interplay of domain-general and domain-specific reasoning.

In the work described in this thesis, we sought to gain greater insight into the nature of student reasoning in physics and the extent to which it is impacted by the domain-general phenomena explored by cognitive science. This was accomplished by developing and implementing new methodologies to examine qualitative inferential reasoning that separate reasoning skills from understanding of a particular physics concept. In this work we present two such methodologies: reasoning chain construction tasks, in which students are provided with correct reasoning elements (*i.e.,* true statements about the physical situation as well as correct concepts and mathematical relationships) and are asked to assemble them into an argument in order to answer a physics question; and possibility exploration tasks, which are designed to measure student ability to consider multiple possibilities when answering a physics problem. The overarching goal of these novel tasks is to explore mechanistic processes related to the generation of qualitative inferential reasoning chains and to uncover insight into the nature of student reasoning more generally.

The work reported in this dissertation has yielded a variety of important results. In concert with reasoning-chain construction tasks, the dual-process framework has been leveraged to provide testable hypotheses about student reasoning and to inform the design of an instructional intervention to support student reasoning. By applying network analysis approaches to data produced by reasoning chain construction tasks with network analysis, insights were uncovered regarding the structure of student reasoning in different contexts, and the development of a coherent reasoning structure over the course of a two-semester physics course was documented. Finally, students' tendency to explore possibilities has been, both in the literature and in this dissertation, found to impact performance on physics questions. This tendency is examined and a possible mechanism controlling this tendency has been proposed. Taken together, these investigations and findings constitute substantive advances in how student reasoning is studied and serve to open new doors for future research.

DEDICATION

To my wife, Ellen Rose Speirs.

ACKNOWLEDGMENTS

In addition to the many academic contributions to this work, my advisor, MacKenzie R. Stetzer, has consistently supported me professional and personally by advocating for me in many different ways. Among these, he was integral in securing the UMaine Emerging Research and Signature Areas Graduate Research fellowship, which was a great financial support; he, along with John Thompson and Pat Byard, consistently went to bat for me during various fiascos regarding paychecks and employment status, accidental travel card revocations, and more. Furthermore, he helped me meet and interact with other researchers in the field, setting me up for future collaborations as well as giving me opportunities to present my work in various places. He was always encouraging and supportive of exploration, which allowed me to grow at my own pace. Even when necessarily critical, he did so with kindness. He deserves a lot of credit.

The research described in this dissertation emerged from a collaborative project involving five different institutions. I would like to acknowledge contributions from collaborating PI's Beth A. Lindsey, Mila Kryjevskaia, Paula R. L. Heron, and Andrew Boudreaux. Beth A. Lindsey, in particular, contributed many insights into the application of dual-process theory, the use of appropriate statistical analyses of data, and general support.

I'd like to acknowledge academic contributions from my advisory committee; Natasha Speer for introducing the literature on student difficulties regarding the construction of mathematical proofs, Michael C. Wittmann for discussion regarding the resources framework, and John Thompson for insights into student difficulty reading graphs. Early conversations with William N. Ferm about student reasoning were delightful and helpful. The members of the Physics Education Research Laboratory (PERL) at UMaine, in specific Benjamin Schermerhorn, Kevin Van De Bogart, and Thanh Lê; and members of the Maine Center for Research in STEM Education (RiSE Center) also contributed thoughts, for which I am grateful. I also wish to acknowledge support from UMaine through the RiSE Center in the form of the UMaine Signature and Emerging Research Areas Graduate Fellowship, which has, in part, supported me financially for the final three years of my doctorate program.

I wish to acknowledge with appreciation conversations with Eric Brewe and Zachary Hutchinson about network analysis. Additionally, my connections to the Ell Lab at UMaine, led by Dr. Shawn Ell, have been invaluable. Finally, it must be stressed that without the honest interest and efforts of the students who participated in this research as subjects, no progress would have been possible.

I would also like to acknowledge the interpersonal support given by the members of the PERL, RiSE Center. The members of these organizations (faculty and students) as well as others in the Department of Physics and Astronomy (notably Katee Schultz, and James Deaton) have worked hard to create a supportive environment in which positive affect is easy to come by. This characteristic of the UMaine research community was an invaluable lifeline to me.

Above all, I'd like to acknowledge primary and near-constant support from Ellen Speirs and from my daughters Violette and Maeve. They've changed my life and their sustaining, grounding influence contributed immensely to this work. Additionally, they sacrificed access to my time and energy during the more consuming parts of the dissertation. They were the primary motivation for advancing my education, and they also bore the primary burden of support while I worked.

This material is based upon work supported by the National Science Foundation under Grant Nos. DUE-1431857, DUE-1431541, DUE-1431940, DUE-1432765, DUE-1432052, and DRL-0962805. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Learning how to reason is essential to a STEM education (National Research Council, 2013; N.G.S.S. Lead States, 2013). Without practice reasoning productively with science concepts, students taking a science course often struggle to develop a functional understanding of those concepts (McDermott, 2001). In addition to definitions, procedures, and strategies related to each concept, students are also often expected to learn how to apply their knowledge on new and difficult problems.

Many students take a physics course in the service of a non-physics STEM major (Conference on Introductory Physics for the Life Sciences Report, 2015; Redish & Hammer, 2009). While certain physics concepts and principles will be of use in these students' future academic careers, many will not. Instead, it is often expected that the lasting takeaways from a physics course will be a repertoire of problem-solving strategies, a familiarity with mathematizing real-world situations, and a strong set of qualitative inferential reasoning skills. These takeaways are of course important to all students taking a physics course, even those who go on to be physics majors and physicists.

Physics education research has produced many instructional materials that have been shown to bolster conceptual understanding and learning outcomes (Finkelstein & Pollock, 2005; Crouch & Mazur, 2001; Beichner R. , 2007; Saul & Redish, 1997; Sokoloff & Thornton, 1997). Many of these materials are scaffolded and step students through a qualitative chain of inferences via a series of questions (McDermott & Shaffer, 2001; McDermott, 1995; Wittmann, Steinberg, & Redish, 2004). It is often tacitly assumed that such materials also improve qualitative reasoning skills, but there is no documented evidence of this, to date, in the physics education research literature. Furthermore, a growing body of research demonstrates that attending solely to conceptual understanding may not produce satisfactory outcomes (Heckler, 2011; Heckler & Scaife, 2014; Kryjevskaia, Stetzer, & Grosz, 2014; Heron, 2017). Instead, these studies suggest that reasoning processes general to all humans (*i.e., domain-general* processes) may impact how students understand and reason with physics concepts. As a result, many researchers have begun to investigate the domain-general cognitive mechanisms that influence human reasoning and how these affect student reasoning on qualitative physics questions (Heckler & Scaife, 2014; Heckler & Bogdan, 2018; Gette, Kryjevskaia, Stetzer, & Heron, 2018; Wood, Galloway, & Hardy, 2016).

Part of the emphasis on domain-general cognitive mechanisms is driven by the observation that students often will demonstrate functional understanding on one physics question but fail to demonstrate the same understanding on a closely related question (Heckler, 2011; Kryjevskaia, Stetzer, & Le, 2015). This phenomenon highlights that conceptual understanding alone is not necessarily predictive of performance on any given task. Instead, domain-general processes may interfere with the application of conceptual understanding on specific tasks. For this reason, it is important to try to separate the reasoning about a physics problem from the conceptual understanding necessary to correctly answer the problem. Methodologies that enable this will aid in understanding how conceptual understanding and domain-general reasoning processes interact.

Understanding this interplay between domain-general reasoning skills and reasoning in a physics context is especially important to the study of how students generate qualitative inferential reasoning chains. A qualitative inferential reasoning chain is a series of inferences where the consequence of one inference becomes the premise for the next. An example would be "My dog is scratching therefore she has fleas. If my dog has fleas it needs a flea collar. These are sold at the pet store, so I need to go the pet store."

To make progress understanding the interplay between domain-general reasoning skills and the formation of qualitative inferential chains of reasoning in physics, new research tools and analysis methodologies are required. Physics education researchers have started to use methodologies that generate data outside of the written free-response questions and think-aloud interviews that are ubiquitous in discipline-based education research. For example, physics education researchers have begun to investigate cognitive processes more directly using alternative methods such as eye tracking (Rosiek & Sajka, 2016; Madsen et. al., 2013; Susac et. al., 2017), timing data (Heckler & Scaife, 2014), gesture analysis (Scherr, 2008), and even fMRI scans of brain functioning (Brewe, et al., 2018). These methodologies have given insight into the root causes of some well-known phenomena. For instance, it has long been established in the literature that students often answer according to the height of a point on a graph even when the when asked to find the slope of that point (McDermott, Rosenquist, & Zee, 1987; Beichner, 1994; Christensen & Thompson, 2012). Timing data has recently suggested that this may be due to the perceptual system taking longer to process the slope than it takes to process the height (Heckler & Scaife, 2014; Heckler, 2011).

Dual process theories of reasoning (DPToR) have played a key role in a renewed effort to understand the mechanisms behind student reasoning in physics. These theories arise from findings in cognitive science, social psychology, and the psychology of reasoning. Popularized by the book *Thinking, Fast and Slow* (Kahneman, 2013), dual-process theories model human reasoning with two types of processing: an unconscious, fast, and associative process 1; and a conscious, effortful, and typically slower process 2. These theories tend to be mechanistic in nature; as such, they provide a framework that can be prescriptive rather than solely descriptive, thereby providing a theoretical basis for the development of successful instructional interventions.

In the work described in this thesis, we sought to gain greater insight into the nature of student reasoning in physics and the extent to which it is impacted by the domain-general phenomena explored by cognitive science. Critical for this investigation were methodologies that could disentangle, to the degree possible, reasoning skills from conceptual understanding. The work presented in this dissertation was aimed at providing new methodologies to examine qualitative inferential reasoning that separate reasoning skills from understanding of a particular physics concept. Accordingly, in this work we present two such methodologies, the overarching goal of which is to explore mechanistic processes related to the generation of qualitative inferential reasoning chains and to uncover insight into the nature of student reasoning generally. In particular, we sought to answer the following research question: To what extent can additional insight into the nature of student reasoning in physics be obtained by applying results from cognitive science about the mechanisms behind human reasoning to the analysis of data from novel physics task formats or methodologies?

The first methodology, implemented in the form of *reasoning chain construction tasks*, aims to create knowledge surrounding how students construct linear chains of inferences in response to qualitative physics questions. Chapter 3 of this dissertation uses reasoning chain construction tasks to investigate the extent to which dual-process theories of reasoning can account for the observed reasoning phenomena mentioned above as well as the extent to which these theories can provide mechanistic predictions for how to improve performance on challenging physics questions. Chapter 4 describes the use of network analysis techniques to gain insight into the structure of student reasoning using the data afforded by the novel reasoning chain construction format.

The second methodology aims to examine student tendency to explore alternate possibilities and is implemented via the *possibilities tasks*. The tendency to explore alternate possibilities is associated with more productive reasoning (Johnson-Laird, 2009; Evans, 2007; Lawson, 2004; Tishman, Jay, & Perkins, 1993); indeed, in some frameworks for human reasoning, that tendency is foundational to the reasoning in general (see, for example, Johnson-Laird, 2009). Motivated by a desire to understand how this domain-general tendency might impact reasoning in physics, Chapter 4 details the *possibilities tasks* and compares data relating to the tendency to construct specific cognitive models with the ability to recognize these models as consistent with physics principles.

The core of this dissertation consists of three individual papers (in preparation for journal submission), included in Chapters 3, 4 and 5. To unify the work presented in those papers, Chapter 2 presents a literature review that establishes the narrative connecting the work described in this dissertation and the existing literature. Chapter 6 summarizes the work done, highlights the coherence of investigations documented in the three papers, and describes plans for future work.

2 REVIEW OF RELEVANT LITERATURE

In this chapter, we draw upon literature from multiple fields in order to motivate, contextualize, and establish the common threads that run through the research described in this dissertation. In physics education research, conceptual understanding and reasoning are often treated as a single thing. Moreover, little distinction is made between domain-specific and domain-general reasoning approaches. As such, this chapter first aims to clearly delineate conceptual understanding, domain-specific reasoning, and domain-general reasoning. Once these important distinctions have been established, key concepts and theories from the psychology of reasoning and decision-making that have been particularly influential on recent physics education research exploring student reasoning are discussed. The chapter then shows how the work presented in this dissertation aims to make further progress on the threads of research established in the literature through the development of new methodologies and the implementation of new analysis techniques to better understand the nature of student reasoning in physics.

2.1 Conceptual understanding, domain-specific reasoning, and domain-general reasoning

The work presented in this dissertation focuses on reasoning skills related to the development of a qualitative inferential chain of reasoning in response to a physics task. When discussing such reasoning, it is helpful to draw distinctions between different phenomena. As such, it is important to operationalize and distinguish between conceptual understanding, domain-specific reasoning, and domain-general reasoning.

Conceptual understanding and domain-specific reasoning are closely related, but this work assumes a distinction between the two on a structural level. Concepts are cognitive constructs with which one can reason. Domain-specific reasoning processes are closely tied to these constructs and comprise procedures, strategies, and rules dictating the use of specific concepts. This distinction is similar to the distinction drawn in the idea of a "coordination class" (diSessa & Sherin, 1998), in which a concept is paired with "readout strategies" and other instructions for the use of that concept in the "coordination class." Indeed, it is hard to separate conceptual knowledge from the reasoning processes most closely associated with that knowledge. McDermott and Shaffer (1992) argued that these associations are fundamental; they stressed that "a concept cannot be isolated from the reasoning process inherent in its definition and application [...]." Thus, knowledge of a concept in some cases depends on the reasoning that establishes the concept. Stanovich (2011) also places these two on a similar level with the concept of *mindware*, which includes "rules, knowledge, procedures, and strategies that a person can retrieve from memory in order to aid decision making and problem solving.” It may be hard either theoretically or empirically to distinguish between a concept and the reasoning associated with that concept, but given the nature of the current investigation, it is imperative to consider the two as separate constructs that are closely associated (in the tradition of diSessa and Sherin (1998)) rather than as a single construct. By doing so, progress can be made in attempting to isolate reasoning skills (to the degree possible) for further study.

Contrasted with domain-specific reasoning is domain-general reasoning, with the latter relying on reasoning mechanisms that may occur in any context. Examples of domain-general reasoning mechanisms are mechanisms that control the allocation of attention, the framing of a problem or task, and/or the generation of intuitive responses. Such mechanisms include the perceptual salience of a task feature and the effects of semantic priming or other priming effects (Heckler, 2011; Hammer, Elby, Scherr, & Redish, 2005, Higgins, 1996). (Section 2.3.2 describes many of these mechanisms in greater detail.) The mechanisms along with the associated reasoning can apply in any context, but they necessarily operate within a specific context (*e.g.,* the context of a physics task) and therefore can produce different results based on that context. Thus, domain-general reasoning can occur in any context and can heavily influence the domain-specific reasoning that occurs in any given context.

As a note, in the remainder of this dissertation, domain-specific reasoning is referred to as "context-specific", "content-specific", or "physics-specific" reasoning in order to contrast it with domain-general reasoning

2.2 Conceptual understanding and domain-specific reasoning in Physics Education Research (PER)

Most of the reasoning-related work conducted in the context of PER has primarily focused on student understanding of specific physics concepts and the domain-specific reasoning related to those concepts. One may consider the large body of work conducted using the framework of specific difficulties (see, for example, McDermott, 2001; McDermott, 1991; Heron, 2004) in order to gain productive insight into student thinking about physics topics using multiple tasks. In this research framework, conceptual and reasoning difficulties are identified, and research-based instructional materials are created to address them. No claims are made about the theoretical or cognitive structure of the difficulties identified; instead, the difficulties are described and their relative prevalence before and after instruction are noted. Difficulties are operationalized in a pragmatic fashion to create actionable data that may guide instructional interventions. The interventions, in turn, can be pre- and post-tested to assess their effectiveness and inform their subsequent refinement.

As a specific example of the use of the difficulties framework, it has been observed that students tend to treat momentum as a scalar quantity rather than a vector quantity when combining momenta (Close & Heron, 2010; Graham & Berry, 1996). This is typically considered to be a conceptual difficulty because it relates to a momentum knowledge construct (*i.e.,* the classification of momentum as either a vector or a scalar quantity), but it could also be seen as a "reasoning difficulty" because it may be that a student has available in memory the knowledge of momentum as a vector but has difficulty determining how vector quantities should be combined. Regardless of the exact cognitive structure of the difficulty, the insights gained informed the development of a tutorial that addressed this specific difficulty (Close & Heron, 2010). Performance on a task that probed the prevalence of this difficulty improved significantly, with the percentage of correct responses with correct reasoning increasing from 35% to 60% after tutorial instruction, indicating that the tutorial successfully addressed and resolved the difficulty for many students.

While the pragmatic specific difficulties framework makes no assumptions about the underlying structure of students' knowledge, other research paradigms expressly focus on the nature of that structure. In the *misconceptions* paradigm (McCloskey, 1983; Posner, Strike, Hewson, & Gertzog, 1982), which is extremely pervasive in the early discipline-based education research literature, once knowledge about physics is constructed, it is thought to be stable and robust. Accordingly, the same knowledge structure is used each time that a particular concept is needed for a task. From this paradigm, one would predict that student performance and reasoning should be consistent on tasks targeting the same concept. In the *knowledge in pieces* paradigm (diSessa, 1993), however, concepts are thought to be built from finer-grained and fragmentary knowledge that combine in the context of a task to produce a conception. These conceptions are inherently unstable and may change from task to task depending on how the fragments are cued and arranged. As an example of a knowledge fragment, consider a primitive conceptual construct possibly born out of experience observing the real world: “dying away”. (For instance, the wind dies down after a storm and a water puddle slowly shrinks until it is gone.) This primitive construct, by itself, is somewhat meaningless, but when combined with specific contexts, it produces emergent knowledge. For example, in a task about energy conservation regarding a gong that has been struck, “dying away” could combine with “the energy” to say that the energy in the gong slowly dies away (as does the sound) and vanishes. However, for an expert, “dying away” would be correctly combined with “kinetic energy”, and the associated construct of dissipation could be cued.

The knowledge in pieces paradigm was subsequently extended into the *resources* framework, which allows one to identify and observe the use of student resources for reasoning (Hammer, Elby, Scherr, & Redish, 2005; Hammer, 2000). Resources refer to finer-grained cognitive structure (*i.e.,* general rules, epistemological stances, the phenomenological primitives from the knowledge in pieces framework, *etc.*) that make up larger-grained cognitive structures such as concepts or skills. It is posited by this framework that the act of reasoning is an act of cognitively selecting and coordinating the use of a subset of available resources. The framework is helpful but tends to fall short of making specific predictions about which resources are activated, when/why they are activated, and how they subsequently impact reasoning. Instead, the resources framework yields compelling post-hoc explanations for reasoning phenomena.

Before moving on, it is worth mentioning that a body of literature has developed in mathematics education research that examines how students construct qualitative inferential "proofs" of mathematical principles (Selden & Selden, 2008). In a typical undergraduate mathematics program, there are specific courses that aim to teach students how to create mathematical proofs. These proofs tend to take the form of a series of deductive, qualitative inferences that are linked together as an argument in support of a specific conclusion. The research regarding student skill at constructing proofs is reminiscent of many research endeavors in physics education research.

Often, students' responses to a particular proof task are examined through various epistemological and conceptual lenses, with the emphasis placed on identifying student difficulties with constructing proofs. The data sources are similar: student written work, interviews, *etc*. Therefore, the methodologies used to study chains of reasoning in a mathematical proof are similar to those already employed in PER. Given that the goal of the work described in this dissertation is develop and apply new methodologies that yield greater insight into the interplay of domain-specific and domain-general reasoning, the specific strategies employed in the proofs literature will not be discussed in detail in this overview.

The three frameworks outlined above have been helpful in creating new knowledge around conceptual understanding and domain-specific reasoning, but recent research is revealing more about their limitations. It is often observed that students may demonstrate functional understanding on one physics question but fail to demonstrate that same understanding on a closely related question (*e.g.,* Kryjevskaia, Stetzer, & Grosz, 2015, Kryjevskaia, Stetzer, & Le, 2015; Heckler, 2011; Kryjevskaia, Stetzer, & Heron, 2012; Close & Heron, 2010; Loverude, Kautz, & Heron, 2002). Further, even after research-based instruction and a documented improvement in conceptual understanding, some physics questions remain difficult for students to answer. Additionally, the existing frameworks may provide some explanatory power in regards to describing what happens when students reason and perhaps why, but they lack predictive power regarding student behavior on novel tasks. These observations present a new challenge for all of these existing frameworks. In particular, the observations highlight that conceptual understanding alone may not be predictive of performance on any given task. Instead, domain-general processes may interfere with the application of conceptual understanding on specific tasks. For this reason, it is important to try to separate the reasoning about a physics problem from the conceptual understanding necessary to correctly answer the problem.

The work in this dissertation aims to separate, to the degree possible, reasoning skills from conceptual understanding for the reasons outlined in the previous paragraph. A method for doing so, which involves paired questions, has been reported previously in the literature (Kryjevskaia, Stetzer, & Grosz, 2014; Kryjevskaia, Stetzer, & Le, 2015). The paired-question methodology uses a screening question that requires that students generate a specific line of reasoning followed by a target question that effectively requires the same line of reasoning in a slightly different context. This then allows researchers to study the responses of those students who answer the screening question correctly but opt for other, perhaps more salient, lines of reasoning on the target question; such students have demonstrated the ability to correctly draw upon relevant concepts in the correct line of reasoning at least once, and so their opting for other lines of reasoning on the target question is likely not due solely to difficulties in conceptual understanding. This methodology is similar to the pairs of questions – developed by Elby (Elby, 2001) and known as "Elby pairs" (see Redish, 2004) – that elicit intuitive answers that are in conflict with each other. While working through an Elby pair, students are tasked with reconciling their intuition with formal physics models, ultimately aiming to refine intuition about those models. The difference in the methodologies is that the goal of the latter was to create an educational outcome while the goal of the former was to isolate and study a reasoning phenomenon. However, both essentially exploit a separation between an intuitive reasoning phenomenon and the conceptual construct associated with it on any given particular task.

2.3 Domain-general reasoning

This section provides an overview of relevant frameworks from the fields of psychology and cognitive science, and then describes PER investigations that have employed these frameworks to date. The PER work is organized around domain-general reasoning mechanisms as a way to establish greater ties between the research presented in this dissertation and the broader work occurring in the PER community.

2.3.1 Research from the fields of psychology and cognitive science

Psychologists have been studying general reasoning processes since the foundation of the field. Modern research regarding the psychology of reasoning began with an intense focus on logical reasoning, primarily with deductive tasks (*e.g.*, syllogisms (Johnson-Laird, 1983) or the Wason selection task (Wason, 1968)). This research gave rise to two competing models of human reasoning. Both theories posit domain-general frameworks for all of human reasoning, meaning that the mechanisms of reasoning are proposed to be the same for each person, in every context. One, the mental logic theory, posited formal but abstract schema for all human reasoning in any context (Braine & O'Brien, 1998), such as *"p or q; not p; therefore q"* for reasoning about logical disjunctions. The other, the mental models theory (Johnson-Laird, 2009), contends that all human reasoning is done by mentally representing the relationships between entities in the mind and then reading judgments and conclusions directly from this representation. The mental model is abstract but iconic, meaning that it represents information spatially and symbolically, even if no actual image is formed in the mind. For example, the phrases "the duck is directly above the dog" and "the dog is somewhere below the fish" would create a mental representation, such as

Fish

Duck
Dog,

from which one could immediately deduce that the fish is above the duck. The proponents of both theories were engaged in ongoing debates, while amassing evidence for both perspectives, for quite some time. However, it has since been pointed out that both theories, even though they disagree on fundamental mechanisms for reasoning, could be true in that reasoners may pick and choose which strategy to use when. “The question of what people ‘really do’ is probably the wrong one to ask,” writes Sternberg (2004), “The question to ask is who does what under what circumstances?” As such, the different theories are suited for different types of analysis. The *mental models theory* is particularly helpful in studying student exploration of alternate possibilities, which is the topic of Chapter 5. Accordingly, more will be said of the *mental models theory* in that chapter.

The context-independent nature of the two previous models of reasoning can be juxtaposed with another class of theories. These theories posit that reasoning is highly context dependent and is not derived from a single mechanism but rather a collection of processes and heuristics built into an "adaptive toolbox" (Gigerenzer, 2008), wherein one can select the best process or heuristic for the job at hand.

Dual-process theories of reasoning and decision-making fall into this view (Kahneman, 2013; Evans & Stanovich, 2013). These theories propose two separate processes in the mind by which reasoning and decision-making occur: process 1; an automatic, subconscious, and generally fast process; and process 2; an effortful, explicit, and generally slow process. Process 1 is primarily at play in decisions such as how to manipulate a steering wheel to keep a car in the center of a lane or judging someone's emotions from a glance at that person's face. Process 1 guides much of adult decision-making throughout the course of a day because it is optimized to reduce cognitive load and free up working memory for more important tasks (*i.e.,* we tend to be misers with respect to cognitive resources). When there is a reason to expend effort, process 2 recruits working memory to run simulations, test hypotheses, or execute an algorithm. This process is helpful with problems such as long division, deducing a result from first principles, or deciding which tax cut to take.

Among the general theories of reasoning that fall under the umbrella of dual-process theories, we have found the heuristic-analytic theory (Evans, 2006) to be particularly helpful in analyzing student responses to our physics tasks. While it is general to any process of reasoning, the heuristic-analytic theory was developed in the context of the psychology of logical reasoning, wherein participants were asked to make judgments about syllogisms or solve logic puzzles such as the Wason selection task (Wason, 1968). The heuristic-analytic theory, shown diagrammatically in Figure 2-1, is therefore particularly suitable for providing detailed roles for process 1 and process 2 in the context of physics. The heuristic-analytic theory of reasoning is especially helpful because it rests on three main principles that describe the mechanisms by which models are selected and/or abandoned. These principles are the relevance principle, the singularity principle, and the satisficing principle (Evans, 2006), and are described below along with the theory itself.

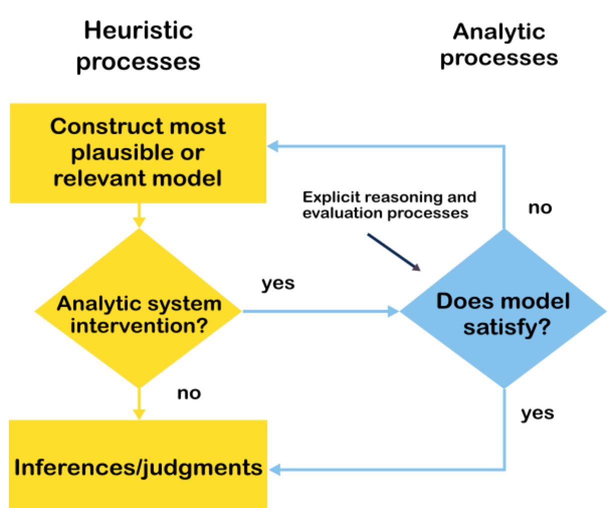


Figure 2-1. Diagram showing the separate roles of the heuristic (type 1) and analytic (type 2) processes, taken from Evans (2006).

In the heuristic-analytic theory, process 1, the heuristic process, is responsible for generating a mental model that is perceived to be the most plausible or relevant given the task features, the goals of the task, and the reasoner's prior knowledge. In this context, a mental model is a hypothetical mental representation of the structure or relationships between given entities. For instance, it may be a schematic of a car engine, a proposition such as "the bigger the coefficient of friction, the bigger the frictional force", or indeed a judgment such as "that person is happy". The singularity principle states that only one mental model is considered at a time. Which model is chosen for consideration is based on the perceived relevance of the model to the current task, which is a statement of the relevance principle. One key aspect of this default model is that it is accompanied with a value judgment about how plausible the model is. This is referred to elsewhere in the literature as a "feeling of rightness" (Thompson, 2009), a measure of how confident a reasoner is that the model is the correct and appropriate one for the task at hand. If the feeling of rightness is strong, process 2 may only be engaged superficially, if at all, before a final judgment is made. If the feeling of rightness is not strong, however, an analytic intervention is triggered and only then does process 2 comes into play in a non-superficial way.

Process 2, the analytic process, is responsible for running mental simulations (explicit reasoning) using the model, and it primarily attempts to ascertain whether the model truly is satisfactory for the task at hand. This point is called the satisficing principle. Thus, process 2 becomes mostly a hypothetical or reflective process with an aim of validating, if possible, the process 1 model. As a result, reasoning biases such as confirmation bias (Nickerson, 1998) can enter into the reasoner's thinking and decision-making. Because process 2 utilizes working memory and is effortful, it is also susceptible to errors in reasoning such as performing an algorithm incorrectly. If the analytic process determines that the initial model is insufficient to the task, the process searches for alternate models and possibilities, and the process is repeated.

Evan's original heuristic-analytic theory (among the first dual-process theories put forward in modern times) had the motivation "to show why reasoning errors are both common and inconsistent across situations" (Evans, 1984). Thus, the intent was to produce a model of reasoning such that the general process described would be able to adapt to context sufficiently to make reasoning itself context-specific. That is, the procedure by which type 1 processes construct a model can differ based on the context, and the type 2 processes employed can also differ from task to task. Thus, the heuristic-analytic theory ensures that there is no need to restrict analysis to a single framework of mental modeling or mental logic. Instead, a wide variety of reasoning phenomena can occur within the basic flow of the heuristic-analytic theory.

Alongside the development of dual-process theories is research regarding "fast and frugal" heuristics for reasoning (Gigerenzer, 2008). It is important to note that Evans' "heuristic" refers to the process that selects models for reasoning, while the "fast and frugal" heuristics explicitly refer to "rules of thumb" for reasoning. These heuristics are thought to have emerged evolutionarily out of a need for reasoners to create good conclusions despite the impossibly complex problems presented by the real world. For instance, the "gaze heuristic" (McLeod, Reed, & Dienes, 2003; see also Shaffer, Krauchunas, Eddy, & McBeath, 2004) is a cognitive heuristic that allows a baseball player to position him- or herself directly under a ball undergoing projectile motion without having to compute differential equations or gather data about initial velocity, wind speed, and other complexities. Instead, the players, utilizing the gaze heuristic, maintain eye contact with the ball and positions themselves such that the angle of their gaze is always constant. Using heuristics, computationally intractable problems (for humans and for computers) can become solvable with a high degree of accuracy.

Heuristics also cause systemic errors, however. For instance, one heuristic proposed by Kahneman and Tversky (1973) is the "availability heuristic", which substitutes an unanswerable question pertaining to the frequency of an event with an answerable question pertaining to the availability of examples of the event. The classic example of this heuristic is to ask the question: "Are there more words that start with 'K' or that have 'K' as the third letter?" The common (incorrect) answer is that there are more words that begin with "K", even though there are, in fact, more words with "K" as the third letter. Kahneman and Tversky demonstrated that because a search of memory likely produces more examples of words that begin with "K" and that these examples come more readily to mind, we assume that the "availability" of examples is proportional to the frequency of the occurrence. While this may be true in many cases, in some cases, it is not, and yet reasoners still make the same error.

Heuristics provide a variety of domain-general reasoning mechanisms that can interact with and interfere with domain-specific reasoning processes. They also fit cleanly into the dual-process perspective but are somewhat incompatible with the two views of *mental models* and *mental logic*.

In the following section, we describe how recent research in physics education has utilized these findings from cognitive science and the psychology of reasoning to advance the community's understanding of how students' reason in a physics context.

2.3.2 Research in field of physics education

While, as discussed in Section 2.2, the reasoning-focused work in PER historically was integrated into topical, concept-focused investigations, the focus of this section is on more recent research on domain-general reasoning in the context of physics education. First, the studies that motivated much of the recent research on domain-general reasoning in physics education are summarized. Then, research regarding known domain-general reasoning mechanisms are detailed, organized by mechanism. The purpose of this section is to introduce and outline what has been done in a physics context so far, illustrating the context and motivation for the current work.

As has been said before, students may demonstrate functional conceptual understanding in one setting and fail to demonstrate it in another setting (Loverude, Kautz, & Heron, 2002; Close & Heron, 2010; Kryjevskaia, Stetzer, & Heron, 2012). Heckler, applying a dual-process framework, argued that patterns of incorrect responses could be explained without referencing an incorrect concept at all; instead, he illustrated how observed patterns could be due to lower-level cognitive factors alone, upon which process 1 draws (Heckler, 2011). Once an answer is obtained, the student might *perhaps* justify using higher-level conceptions and type 2 processes. Thus, the student may answer not from an incorrect physics conception but from no conception at all. In this paper, Heckler also called for new methodologies that use domain-general mechanisms to make and test predictions about answering patterns. In particular, he proposed two such these mechanisms: the time it takes to cognitively process task features and the allocation of attention given to salient distracting cues. The current work documented in this dissertation is in large part a direct response to that call.

The rest of the section is organized around specific reasoning mechanisms and the work that has been done surrounding these mechanisms. This discussion is important because it sets up the context in which the current work is taking place and serves as an introduction to some of the mechanisms that will be in play in the tasks described later in this dissertation.

2.3.2.1 Processing time

In order for a task feature to cue a specific resource in the course of reasoning, it must be processed by the brain. Thus, the time it takes to process a certain feature represents a control mechanism that may predict which resources are cued and when. To show the impact of processing time on answering patterns, Heckler and Scaife measured the approximate processing time of finding either the slope or the height of a particular point on a graph and determined that processing the slope took a longer time than processing the height (Heckler & Scaife, 2014). The researchers then demonstrated that applying a time delay on answering in order to guarantee that the brain had time to process the slope improved performance on graph-based questions in which the slope and the height were in competition. They framed this mechanism as a version of the fluency heuristic (Schooler & Hertwig, 2005) wherein process 1 gathers information about the two dimensions available in the question (height and slope) and responds based on the dimension processed first (*i.e.,* the height) or most fluently.

2.3.2.2 Allocation of Attention

Heckler (2011) proposed[1] that salient features (Elby, 2000; Heckler, 2011; Kryjevskaia, Stetzer, & Le, 2015; Le, 2017) control the allocation of attention and can be used to make predictions about student behavior. *Salient distracting features* (SDFs) are features of a task that draw immediate attention away from other task features, are processed easily, and cue incorrect lines of reasoning. The salience of a feature can be empirically measured by using eye-tracking techniques to determine where attention is being placed. For questions in which high-salience information is irrelevant and low-salience information is relevant, it can be expected that the competition between these relevant and irrelevant features will lead to most students generating an incorrect default model based on the high-salience of

[1] It should be mentioned that a similar argument was put forward by Elby in 2000 (Elby, 2000).

the irrelevant feature. Thus, in salient distracting features, we have a predictive factor that can provide insight into student answering patterns.

Heckler demonstrated the impact of salient distracting features on physics questions by providing students with a plot of two position *vs.* time graphs representing the motion of two cars, shown in Figure 2-2. In each question, the students were asked to find the time at which the cars have the same speed. In one question (shown in Figure 2-2.a), the two graphs were parallel lines, and 90% of students chose the correct answer (“At all times”). In another question (shown in Figure 2-2.b), the two graphs intersected at time B while the slopes of the graphs were the same at time A; in this question, the intersection serves as the salient distracting feature. Sixty percent of students answered correctly (time A), while 40% answered incorrectly by picking the intersection (time B). This tendency to focus on and incorrectly interpret intersections on graphs is reported extensively in the literature (McDermott, Rosenquist, & Zee, 1987; Beichner, 1994; Christensen & Thompson, 2012; Elby, 2000; Heckler, 2011; Speirs, Ferm Jr., Stetzer, & Lindsey, 2016). Notably, students may utilize physics concepts in order to rationalize an incorrect time B answer, highlighting the interplay between low-level factors and higher-level reasoning structures, as discussed by Heckler (see Heckler, 2011).

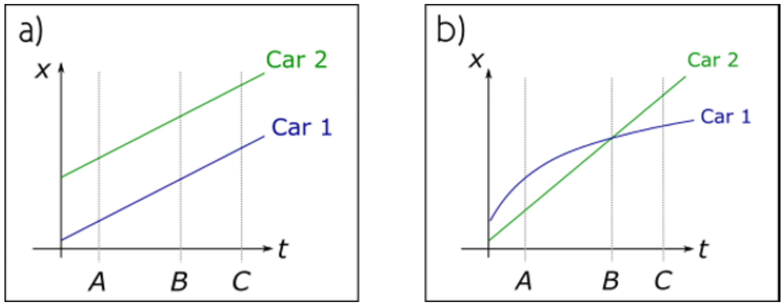


Figure 2-2. Diagrams given to students as part of a study reported in (Heckler, 2011). The graph shown in (b) was used in the kinematics graph task (Experiments 1A and 1B) for the current work.

The effect of non-science-graph related salient distracting features on inconsistencies in student reasoning was also explored using the paired question methodology. Kryjevskaia *et. al.* (2015) studied a physical context in which a box remains at rest when a known force is applied, and the student must reason with Newton's 2nd Law to infer the magnitude of the static friction force. In the screening question (see Figure 2-3.a), a single box is shown, and students are told that the box remains at rest when an applied force of 30 N is acting on the box. Students are asked to compare the magnitude of the applied force with the magnitude of the friction force. The correct line of reasoning is that the box remains at rest and, by Newton's 2nd Law, this requires that the net force on the box must be zero and therefore the magnitudes of the two forces must be equal to each other. In the target question, students are asked to compare the forces of friction on two identical boxes on two different surfaces with identical applied forces exerted on both boxes (Fig. 2b). In the diagram, the coefficient of static friction for each box-surface pair is shown next to each box. These coefficients appear to elicit a common but incorrect comparison that the friction force on box A is less than the friction force on box B because the coefficient for box A is less than the coefficient for box B. Typically, 50% of students will answer this way, and 50% will answer correctly (Kryjevskaia, Stetzer, & Le, 2015).

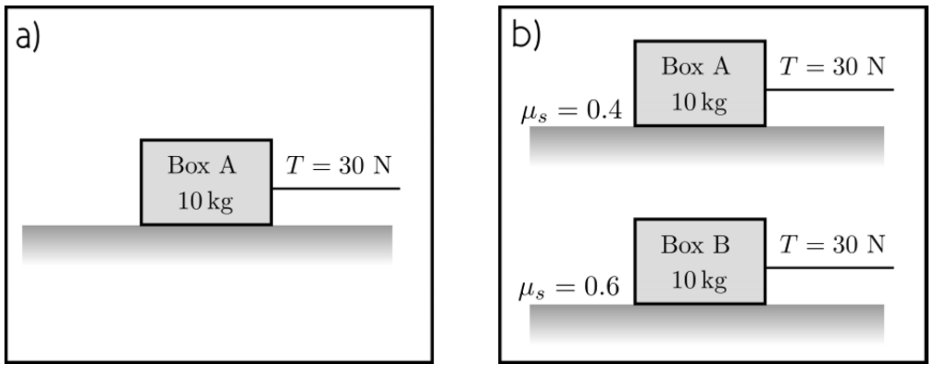


Figure 2-3. Diagrams given to students for (a) the screening question and (b) the target question of the two-box friction task.

If, instead, one was to reason from Newton's second law and the observation that both boxes remained at rest, the (correct) conclusion would be that the friction force on box A is equal to the friction force on box B. Of those who answered the screening question correctly (demonstrating the relevant conceptual understanding) 35% employed an incorrect line of reasoning on the target question (Kryjevskaia, Stetzer, & Le, 2015). This result was interpreted as a failure to engage the analytical process 2 in a productive manner due to the salient distracting features. Instead, students appeared to rely on process 1 first impressions for reasoning; despite the fact that they demonstrated the ability to step through a correct line of reasoning, they abruptly abandoned that line of reasoning on the target question. This abrupt abandonment was further supported in Kryjevskaia, Stetzer, & Le (2015) using a transcript from an interview in which a pair of students worked through both parts of the static friction task consecutively. This study provided additional evidence that low-level cognitive influences can have an impact on the use of higher-level mental structures, but it was unclear as to how exactly this impact could be mitigated. The work described in Chapter 3 makes progress on cognitive-science based efforts to mitigate such effects.

2.3.2.3 Reasoning Heuristics: Compensation Reasoning

A related study is of note because it utilizes dual-process theory as well as the paired-question methodology to study a commonly cued incorrect line of reasoning not necessarily associated with a salient distracting feature. Kryjevskaia *et al.* (2014) reported on a physics task that was known to cue a common incorrect line of reasoning involving *compensation reasoning*. In the capacitor question (diagram shown in Figure 2-4), two identical capacitors are each fully charged across an identical battery and then placed in series such that they didn't discharge. The left capacitor is then modified by increasing the distance between its plates. The screening questions asks students to determine whether, for the modified (left) capacitor, the charge on the plates and the potential difference between the plates *increases*, *decreases*, or *remains the same* after the modification. The target question asks the student to determine if the potential difference across the right (unmodified) capacitor *increases*, *decreases,* or *remains the same* after the modification.

The correct answer to the first screening question is that because charge is conserved and the capacitors are not connected to a battery, the charge remains the same on all plates. Then, the distance has increased between the left capacitor plates, so the capacitance has decreased ($C = \frac{\varepsilon A}{d}$, where d is the plate separation), in turn causing the potential difference between the plates of the left capacitor to increase, because $\Delta V = \frac{Q}{C}$. Since the charge on the plates and the capacitance of the right capacitor remain the same, the potential difference across the plates of the right capacitor also remains the same.

On the target question, about half of students are reported to have answered incorrectly that since the potential difference across the left capacitor has increased, the potential difference across the right capacitor must decrease to keep the total potential difference constant. This reasoning was identified as "compensation reasoning", which has been reported in the literature in a variety of contexts (Lindsey, Heron, & Shaffer, 2009; Kautz, Heron, Shaffer, & McDermott, 2005; Loverude, Kautz, & Heron, 2003). It was suspected by Kryjevskaia, Stetzer and Grosz (2014) that the frequent use of "equilibrium" and "conservation" ideas in the physics classroom made those ideas more readily accessible to students on this task, and thus the default model selected by process 1 would be related to conserving the total potential difference. Because process 2 only considers other alternatives when the default model is rendered unsatisfactory for some reason, students would not tend to consider the reasoning they used on the screening questions, they surmised. Approximately 50% of students who answered both screening questions correctly used compensation reasoning on the target question, thereby arriving at an incorrect response.

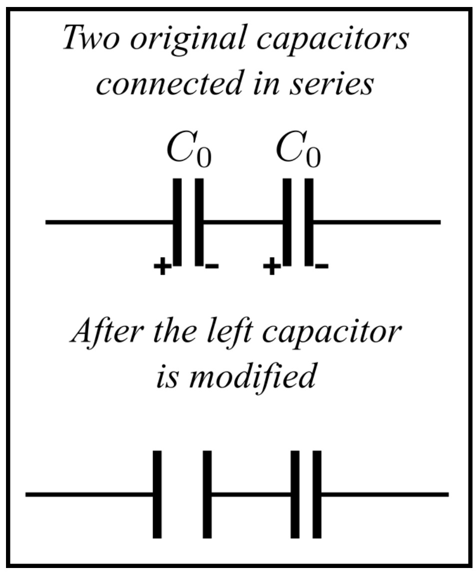


Figure 2-4. Diagram given to students on a capacitor task administered by Kryjevskaia et. al. (Kryjevskaia, Stetzer, & Grosz, 2014).

In a second experiment, students were given an alternate version of the target question that asked them to justify why the potential difference across the right capacitor remained the same. Of those who answered both screening questions correctly, almost all (86%) answered this alternate version with correct reasoning.

The results from the second experiment were also interpreted from a dual-process theory perspective. When given the correct answer, students were able to reason effectively either because (1) being cued by the correct answer, process 1 was used to construct a correct model or (2) process 2 was effectively engaged to aid the student in abandoning an incorrect default model because it was not satisfactory in arriving at the stated answer. In either case, it was clear that students did, in fact, have correct and relevant mindware (in the sense used in Section 2.3.1) available to them, even if they did not use it when the target question was posed originally.

2.3.2.4 Cognitive Accessibility

The cognitive accessibility of an initial idea can impact a student's tendency to explore alternate possibilities if the accessibility of the initial idea is much higher than the other possibilities (Quinn & Markovits, 1998). Cognitive accessibility is a measure of how easily a concept or model is retrieved from memory (Higgins, 1996). Heckler and Bogdan (2018) investigated the effects of accessibility on physics questions. They first measured the relative cognitive accessibility of causal factors associated with different physics contexts, such as how the length and mass affect the period of a pendulum. They then found that when a highly accessible factor was offered in a problem statement, students tended not to explore alternate factors, even when the factor offered was causally irrelevant to the physics scenario (*e.g.,* the mass of a pendulum). Furthermore, when the less accessible factor was offered students did explore alternate factors, namely the highly accessible factor. They surmised that accessibility could thus represent a "soft contour" (*i.e.,* a control mechanism) that influences the trajectory of a reasoning process.

Importantly, the general effects of relative cognitive accessibility were demonstrated in the contexts of forces/friction, simple harmonic motion, kinematics, potential energy, and mass density (Heckler & Bogdan, 2018). This is particularly relevant to the current work in that their findings demonstrate how low-level factors such as how closely two ideas are associated can be domain-general in that they impact performance in predictable ways across context.

2.3.2.5 Cognitive Reflection

When a question has a strong intuitive but incorrect response (for instance, "Which weighs more? A pound of feathers or a pound of rocks?"), a reasoner must suppress or otherwise reason through that strong intuitive response in order to arrive at a correct answer. Frederick (2005) introduced a test, known as the "cognitive reflection test" or CRT, to measure this tendency to suppress such incorrect responses. The CRT consists of seven questions, each of which cues a strong intuitive yet incorrect answer, but which are relatively easy to solve otherwise. For instance, one question poses the following problem, “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?” The intuitive answer is “ten cents”, but a quick calculation shows that this would imply that the bat costs $1.10 for a total of $1.20 for both the ball and bat. Therefore, the ball must cost five cents. Performance on this and similar questions serves as a proxy for the skill of reasoning past an intuitive response.

A recent study (Wood, Galloway, & Hardy, 2016) in a physics context examined the relationship between students’ scores on the Cognitive Reflection Test and their performance on the Force Concept Inventory, a survey designed to assess student understanding of Newtonian concepts of force and motion. Wood *et al.* report that students who scored higher on the CRT have higher scores on the FCI, both pre- and post-instruction. An unstated implication of Wood *et al.*’s work is that the skill of productively navigating intuitive responses is required to answer FCI questions correctly. Indeed, with distractors built into many FCI questions, this would be expected since intuitive responses are in part based on distracting features. However, their findings also highlight that a domain-general reasoning skill (that of productively navigating intuitive responses) may also have an impact into the formation of correct physics concepts for students. One implication of such work is that attending to these domain-general skills may lead to better outcomes for students.

2.3.2.6 Summary of work in PER

The low-level, domain-general influences described in the previous subsections represent mechanisms from which predictions about student performance patterns can be made; as such, understanding their impact on reasoning can provide guides and leverage for improving student performance and reasoning skills overall. Some early efforts have been made to draw upon these mechanisms in order to improve student performance, (see, for example, (Gette, Kryjevskaia, Stetzer, & Heron, 2018)), and the closely related investigations described in Chapter 3 of this dissertation represent another attempt to leverage the ongoing research on cognitive mechanisms to improve student performance.

2.4 Connections to the current work

The work documented in this dissertation attempts to explore the impact of domain-general reasoning phenomena on student reasoning and performance in physics. This research was motivated by the emerging body of work in PER that draws upon the findings of cognitive science, particularly the recent work that investigated student reasoning inconsistencies in detail and highlighted mechanistic pathways for progress (via dual-process theories and domain-general reasoning mechanisms).

The research described in this dissertation capitalizes on and leverages the current literature to gain deeper insight into the nature of student reasoning in physics. The existing research base has not provided an actionable position from which we can use our current understanding of student reasoning to help improve student performance. Instead, it provides primarily descriptive, post hoc accounts of student reasoning. The methodologies presented in this work are a direct response to the call in Heckler (2011), and others, for mechanistic theories of student inconsistency.

In particular, the reasoning chain construction tasks and alternate means of analyzing data generated from such tasks (Chapters 3 and 4) serve as the foundation for a comprehensive new methodology that can be used to examine the structure of student qualitative inferential reasoning chains and has the ability to study concept-specific reasoning as well as the effect of cognitive mechanisms on that content-specific reasoning. The effectiveness of this methodology stems from the disentangling of conceptual understanding and reasoning skills that is expressly built into the reasoning chain construction task format (as highlighted in Chapter 3).

The possibilities tasks (Chapter 5) provide the basis of another methodology that examines the tendency of students to search for alternate possibilities and is directly related to the domain-general mechanism of knowledge accessibility (in contrast to knowledge availability). This methodology allows the researcher to examine the impact of knowledge accessibility on reasoning in the context of a physics problem.

The experiments described in the following chapters were designed expressly through the lenses of cognitive science frameworks of reasoning and decision-making. Dual-process theories of reasoning (highlighted in Section 2.3.1) guided the majority of the research and research design, but Johnson Laird's *mental models* framework was also used in order to explore the accessibility and availability of knowledge via the possibilities tasks.

Because of the foundations of the theories of reasoning utilized, the current work stands to advance our field's understanding of the interplay between domain-general reasoning and physics content-specific reasoning and to leverage that increased understanding to establish a foundation for future research-based instructional materials capable of improving student performance in physics more broadly.

3 EXPLORING AND SUPPORTING STUDENT REASONING IN PHYSICS BY LEVERAGING DUAL-PROCESS THEORIES OF REASONING AND DECISION-MAKING

3.1 Abstract

A major goal of a typical physics course is to improve student reasoning skills. As such, there has been attention placed on developing theoretical frameworks in a physics context for how students reason through physics problems. Many theories of student reasoning focus on the cueing and structure of the mental model(s) the student uses when reasoning through a physics question but are vague with regards to questions about why a particular model is cued instead of another or the circumstances under which one is abandoned in favor of another. In other words, they tend to be explanatory rather than predictive. Dual-process theories of human reasoning, established outside of a physics context in the fields of cognitive science and psychology, have recently been applied in a physics context and allow for more mechanistic predictions of student reasoning. However, new methodologies are needed to study in greater detail the effects predicted by dual-process theories of reasoning, and to study reasoning in a physics context from other frameworks as well. Here, we present a novel methodology, the reasoning chain construction task, for studying student inferential reasoning in a physics context. In a reasoning chain construction task, or simply chaining task, a student is given a list of reasoning elements

(such as statements of physics concepts) and is asked to assemble a chain of reasoning leading to an answer from the elements. In this paper, we draw upon dual-process theories specifically to make predictions for student behavior on chaining tasks and demonstrate a successful intervention based on these theories. Our findings support the mechanisms put forward by many dual-process theories, namely that reasoners consider only one model at a time, that the first model considered is selected based on salient problem features, and that students only abandon a first-impression model if that model is directly challenged by new information.

3.2 Introduction

Many students take introductory physics courses in service of other majors in a variety of different STEM fields. It is often expected that these students will take the knowledge gained and, perhaps more importantly, the reasoning skills acquired in the course for use in their respective fields of study. Research-based instructional materials and approaches have been demonstrated to increase student conceptual understanding of core physics concepts (Finkelstein & Pollock, 2005; Freeman, et al., 2014), but little of this work attends to the process of reasoning itself. Additionally, even after instruction using these approaches it remains difficult to increase student performance on certain qualitative physics questions (Kryjevskaia, Stetzer, & Grosz, 2014; Kryjevskaia, Stetzer, & Le, 2015). More detailed research into these questions has led physics education researchers to believe that processes generic to all human reasoning – not necessarily associated with physics content - may be impacting the way students answer these questions (Kryjevskaia, Stetzer, & Grosz, 2014; Kryjevskaia, Stetzer, & Le, 2015; Heckler, 2011). As a result, many researchers have begun to investigate the cognitive mechanisms that influence human reasoning and how these affect student reasoning on qualitative physics questions (Heckler & Scaife, 2014; Heckler & Bogdan, 2018; Gette, Kryjevskaia, Stetzer, & Heron, 2018; Wood, Galloway, & Hardy, 2016).

For example, physics education researchers have begun using alternative methods such as eye tracking (Rosiek & Sajka, 2016; Madsen, Rouinfar, Larson, Loschky, & Rebello, 2013; Susac, Bubic, Martinjak, Planinic, & Palmovic, 2017), timing data (Heckler & Scaife, 2014), gesture analysis (Scherr, 2008) and even fMRI scans of brain functioning (Brewe, et al., 2018) to investigate cognitive processes directly. These methodologies have given insight into the root causes of some well-known phenomena. For instance, it is established in the literature that students often answer according to the height of a point on a graph even when the when asked to find the slope of that point. Timing data suggested that this may be due to the perceptual system taking longer to process the slope than it takes to process the height.

Dual process theories of reasoning (DPToR) have played a key role in a renewed effort to understand the mechanisms behind student reasoning.

These theories arise from findings in cognitive science, social psychology, and the psychology of reasoning. Popularized by Daniel Kahneman's book *Thinking, Fast and Slow* (Kahneman, 2013), DPToR models human reasoning with two types of processing: an unconscious, fast, and associative process 1; and a conscious, effortful, and typically slower process 2. These theories tend to be mechanistic in nature; as such, they provide a framework that can be prescriptive rather than solely descriptive and therefore provide a basis for progress in developing successful instructional interventions.

While dual-process theories are useful for understanding domain-general cognitive mechanisms and their impact on student use of conceptual understanding on a given physics problem, new research methodologies that can disentangle student reasoning skills from conceptual understanding are also needed. Our collaboration has sought to develop and refine such methodologies, and this paper presents one of these novel methodologies, the *reasoning chain construction task*. This methodology has been useful in studying explicit process 2 reasoning, especially the formation and structure of student's qualitative inferential reasoning chains. However, it has also proven useful in investigating the extent to which DPToR can account for observed patterns in student reasoning. Accordingly, in this paper, we draw upon dual-process theories to make predictions for student behavior on chaining tasks and demonstrate a successful intervention based on these theories. This provides additional support for the mechanisms put forward by many dual-process theories and has implications for future curricular materials.

3.3 Background/Motivation

When a student answers a qualitative physics question incorrectly, it is often assumed that the student did not possess a robust conception of the accurate physics involved. Instead, the student presumably reasoned from an incorrect or incomplete conception of the physics. There are differing perspectives as to the structure of these conceptions. One perspective is that physics (mis)conceptions, once learned, are stable and robust and the same conception would be applied in every instance in which they are needed (McCloskey, 1983; Posner, Strike, Hewson, & Gertzog, 1982), much like a tractor, once manufactured, is used whenever one perceives that a tractor is needed. Another perspective (diSessa, 1993; Hammer, Elby, Scherr, & & Redish, 2005; Hammer, 2000) holds that physics conceptions are built from fragmentary knowledge and resources assembled at the time the task is being performed, much like a toy tractor assembled from toy construction bricks; as such, each conception is inherently unstable and can appear different based on the task. The former is generally referred to as the "misconceptions" framework, while the latter is referred to as the "resources" perspective. A third, alternate way of modeling student reasoning is to search for student "difficulties"; in this perspective, the emphasis is not on the structure of the knowledge or its stability, but rather the frequency of its occurrence among a population of students (Heron, 2004; McDermott, 1991; McDermott, 2001).

In both of the misconceptions and resources perspectives, it is assumed (Heckler, 2011) that some form of higher-level cognitive construct, such as a concept or a particular type of mental model (*e.g.,* Gentner & Stevens, 1983), is being used to answer physics questions even if the model was constructed from lower-level knowledge pieces. A growing body of recent research is challenging this view. Much of this research utilizes dual-process theories of reasoning (Evans, 2006; Evans & Stanovich, 2013; Kahneman, 2013) which posit two types of reasoning processes in the mind; one is automatic, subconscious (intuitive), and generally fast; the other is effortful, reflective, and generally comparatively slower. These two processes are referred to as process 1 and process 2, respectively[2]. Process 1 is responsible for giving a first impression response that process 2 then follows up on (if necessary) using explicit reasoning, most commonly in the form of mental simulation and hypothetical thinking. From a dual-process theory perspective, Heckler argued in 2011 that incorrect responses could be explained without reference

[2] There has been an evolution of terms in the literature regarding dual-process theories. In some cases, the terms "system 1" and "system 2" are used, as in Kahneman (2013); wishing to not implicate specific biological or neurological systems in dual-process theory, the terminology now preferred by Evans and Stanovich (Evans & Stanovich, 2013) is "type 1 processes" and "type 2 processes". This manuscript uses primarily uses "process *x*" to refer to "type *x* processes".

to an incorrect conception; instead, the pattern could be due to lower-level cognitive factors alone, which process 1 uses to determine an answer that the student might – perhaps - *justify* using the higher-level conceptions and type 2 processes. Thus, the student may be answering not from an incorrect physics conception but from no conception at all.

Heckler's argument brings into focus the need for research regarding the reasoning processes that might be impacting how students think about and answer qualitative physics questions. More specifically, the interplay between the lower-level factors and the higher-level mental constructs needs to be understood in greater detail. Along these lines, recent research has investigated the role of processing time in questions where there are two competing dimensions (such as the slope and the height of a point on a graph) (Heckler & Scaife, 2014), the impact of perception-based bias in determining the center-of-mass (Heron, 2017), how the relative cognitive accessibility of certain ideas can influence student's performance on a wide range of tasks (Heckler & Bogdan, 2018), and how the cognitive skill of suppressing an intuitive, process 1 response impacts student performance in the course overall (Wood, Galloway, & Hardy, 2016).

The presence of a salient distracting feature (SDF) (Elby, 2000; Heckler, 2011; Kryjevskaia, Stetzer, & Le, 2015; Le, 2017) is another of these factors. They have special relevance to the current work and will therefore be explained in greater detail. Salient distracting features are features of a task that draw immediate attention away from other task features, are processed easily, and cue incorrect lines of reasoning. The salience of a feature can be operationalized by using eye tracking techniques to determine where attention is being placed. For questions in which high-salience information is irrelevant and low-salience information is relevant, it can be expected that the competition between these relevant and irrelevant features will lead to most students generating an incorrect default model based on the high-salience of the irrelevant feature. Thus, in salient distracting features we have a predictive factor that, if harnessed, can provide insight into student answering patterns.

Heckler demonstrated the impact of salient distracting features on physics questions by providing students with a plot of two position vs. time graphs representing the motion of two cars, shown in Figure 3-1. In each question, the students were asked to find the time where the cars had the same speed. In one question (shown in Figure 3-1.a), the two graphs were parallel lines, and 90% of students chose the correct answer ("At all times"). In another question (shown in Figure 3-1.b), the two graphs intersected at time B while the slopes of the graphs were the same at a time labelled "A"; 60% of students answered time A (correct), and 40% answered time B. This difficulty with intersection points on graphs is also reported in other studies (McDermott, Rosenquist, & Zee, 1987; Beichner R. J., 1994; Elby, 2000; Heckler, 2011; Christensen & Thompson, 2012; Speirs, Ferm Jr., Stetzer, &

Lindsey, 2016). Notably, students may utilize physics concepts in defense of a time B answer, highlighting the interplay between low-level factors and higher-level reasoning structures.

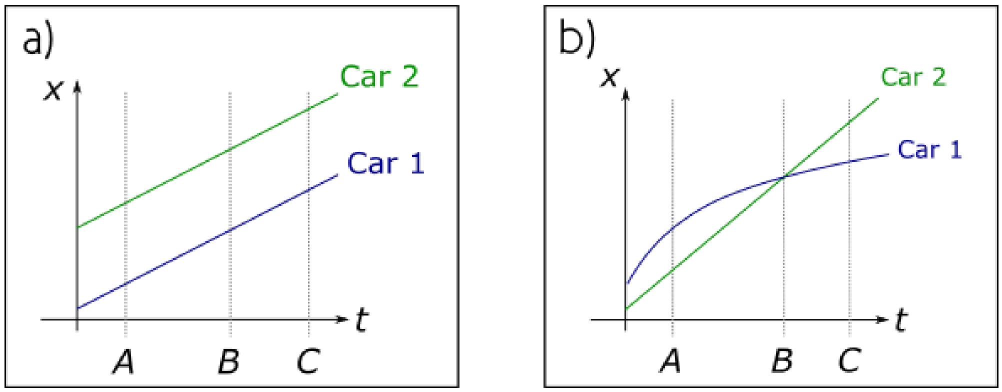


Figure 3-1. Diagrams given to students as part of a study reported in (Heckler, 2011). The graph shown in (b) was used in the kinematics graph task (Experiments 1A and 1B) for the current work.

To better understand these factors and their interplay with higher-level knowledge, there is a need for methodologies that separate, to the degree possible, student reasoning skills from conceptual understanding. A method for doing this, which involves paired questions, has been reported on previously (Kryjevskaia, Stetzer, & Grosz, 2014; Kryjevskaia, Stetzer, & Le, 2015). The paired-question methodology uses a screening question which requires the student to generate a specific line of reasoning followed by a target question that effectively requires the same line of reasoning in a slightly different context. This then allows one to study those students who answer the screening question correctly but opt for other, perhaps more salient, lines of reasoning on the target question; such students have demonstrated the ability to correctly draw upon relevant concepts in the correct line of reasoning at least once, and so their opting for other lines of reasoning on the target question is likely not due completely to difficulties in conceptual understanding. This methodology is similar to so-called "Elby Pairs" (Elby, 2001; Redish E. F., 2004) which are pairs of questions that elicit intuitive answers which are in conflict with each other; the task for the student became reconciling their intuition with the formal physics with the aim of refining intuition. The difference in the methodologies is that the goal of the latter was to create an educational outcome while the goal of the former was to isolate and study a reasoning phenomenon.

The paired question methodology was used to study a static friction task in which the student is expected to reason with Newton's 2nd Law to determine the magnitude of a friction force for a box that remains at rest. In the screening question (see Figure 3-2.a), a single box is shown and students are told that the box remains at rest when an applied force of 30 N is acting on the box. Students are asked to compare the magnitude of the applied force with the magnitude of the friction force. The correct line of reasoning is that the box remains at rest and, by Newton's 2nd Law, this requires that the net force on the box must be zero and therefore the magnitudes of the two forces must be equal to each other. In the target question, students are asked to compare the forces of friction on two separate, identical boxes on different surfaces with identical applied forces exerted on both boxes. (see Figure

3-2.b). In the diagram, the coefficient of static friction for each box-surface pair is shown next to each box. These coefficients appear to elicit a common but incorrect comparison that the friction force on box A is less than the friction force on box B because the coefficient for box A is less than the coefficient for box B. Typically, 50% of students will answer this way, and 50% will answer correctly (Kryjevskaia, Stetzer, & Le, 2015).

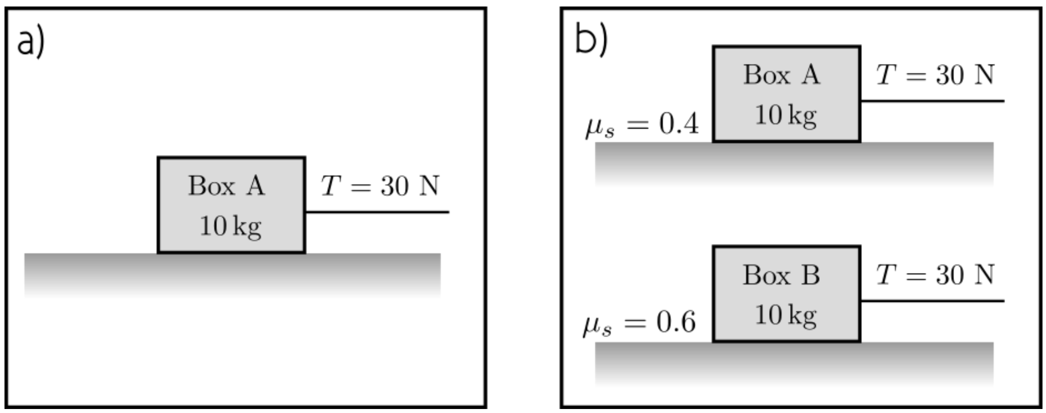


Figure 3-2. Diagrams given to students for (a) the screening question and (b) the target question of the two-box friction task.

If, instead, one was to reason from Newton's second law and the observation that both boxes remained at rest, the (correct) conclusion would be that the friction force on box A is equal to the friction force on box B. Of those who answered the screening question according to the correct line of reasoning, 35% opted to use the incorrect line of reasoning on the target question (Kryjevskaia, Stetzer, & Le, 2015). This result was interpreted as a failure to engage the analytical process 2 in a productive manner. Instead, students appeared to rely on process 1 first impressions for reasoning cued by the salience of the coefficients; despite the fact that they demonstrated the ability to step through a correct line of reasoning, they abruptly abandoned that line of reasoning on the target question. This abrupt abandonment was highlighted in Kryjevskaia, Stetzer, & Le (2015) using a transcript from an interview in which a pair of students worked through both parts of the static friction task consecutively. This study provided further evidence that low-level cognitive influences can have an impact on the use of higher-level mental structures, but it was unclear as to how exactly this impact could be mitigated.

Low-level factors such as the salience of a specific feature can be domain-general in that they impact performance in predictable ways across context. For instance, the general effects of relative cognitive accessibility (Heckler & Bogdan, 2018) were demonstrated in the contexts of forces/friction, simple harmonic motion, kinematics, potential energy, and mass density. These low-level, domain-general influences represent mechanisms from which predictions about student performance patterns can be made; as such, understanding their impact on reasoning can provide guides and leverage for improving student performance and reasoning skill overall. Some early efforts have been made to draw upon these mechanisms in order to improve student performance, *e.g.,* Gette, Kryjevskaia, Stetzer, & Heron (2018), and the closely related investigations described in this article represent another attempt to leverage the ongoing research on cognitive mechanisms to improve student performance.

3.4 Theoretical framework

This work utilizes dual-process theories of reasoning as a theoretical framework. These theories propose two separate processes in the mind by which reasoning and decision making occur: process 1, an automatic, subconscious, and generally fast process, and process 2, an effortful, explicit, and generally slow process. Process 1 is primarily at play in decisions such as how to manipulate a steering wheel to keep a car in the center of a lane or judging someone's emotions from a glance at that person's face. Process 1 guides much of adult decision making throughout the course of a day because it is optimized to reduce cognitive load and free up working memory for more important tasks (*i.e.,* we tend to be misers with regards to cognitive resources). When there is a reason to expend effort, process 2 comes into play recruiting working memory to run simulations, test hypotheses, or execute an algorithm. This process is helpful with problems such as long division, deducing a result from first principles, or deciding which tax cut to take.

Because dual-process theories of reasoning originated outside the field of physics education research, it is helpful to situate them within the context of the frameworks utilized by physics education researchers. Dual-process theories fit cleanly into the resources perspective. This point is illustrated by

Elby (2000). In this paper, he posited a fine-grained cognitive structure that promotes a "same means same" resource which he named the "WYSIWYG intuitive knowledge element" (*"what you see is what you get"*). He used this knowledge element to predict that students would be stymied by a graph task with an intersection such as the one shown in Figure 3-1.b because of activation of this knowledge element. Critically, he argued that activation of the knowledge element is based upon the perceptual salience of the intersection because "the human visual system [is] hardwired to 'see' certain features such as edges, corners, and motion." In this paper, he put forward salient distracting features as a *control mechanism* by which resources are activated or remain unactivated.

The resources framework offers post-hoc explanatory power for understanding how our students may be thinking, but it falls short in offering the mechanisms by which predictions could be made (aside from the paper mentioned above). Specifically, the framework falls short in answering the questions of which models are activated when there are competing models and when models are abandoned in favor of other models. Dual-process theories of reasoning offer these mechanisms, and as such can provide predictions for student performance on physics questions.

Among the general theories of reasoning that fall under the umbrella of dual-process theories, we have found the heuristic-analytic theory (Evans, 2006) to be particularly helpful in analyzing student responses to our physics tasks. While it is general to any process of reasoning, the heuristic-analytic theory was developed in the context of the psychology of logical reasoning, wherein participants were asked to make judgements about syllogisms or solve logic puzzles such as the Wason selection task (Wason, 1968). The heuristic-analytic theory, shown diagrammatically in Figure 3-3, is therefore able to provide detailed roles for process 1 and process 2 in the context of physics. The heuristic-analytic theory of reasoning is especially helpful because it rests on three main principles that describe the mechanisms by which models are selected and/or abandoned. These principles are *the relevance principle, the singularity principle,* and *the satisficing principle* (Evans, 2006), and are described below along with the theory.

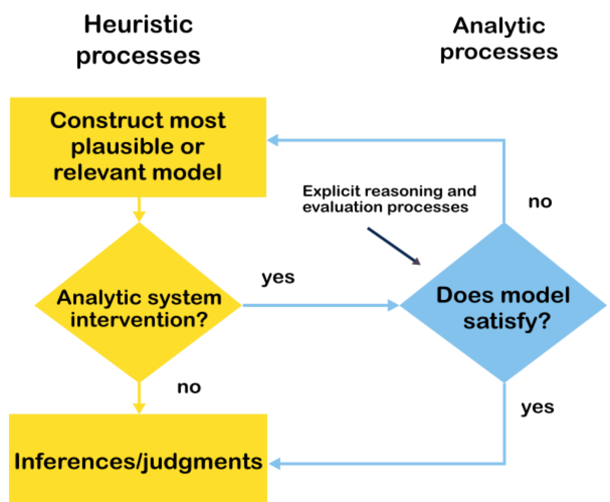


Figure 3-3. Diagram showing the separate roles of the heuristic (type 1) and analytic (type 2) processes, taken from (Evans, 2006).

In this theory, process 1, the heuristic process, is responsible for generating a mental model that is perceived to be the most plausible or relevant model given the task features, the goals of the task, and the reasoner's prior knowledge. In this context, a mental model is a hypothetical mental representation of the structure or relationships between given entities. For instance, it may be a sketch schematic of a car engine, a proposition such as "the bigger the coefficient of friction, the bigger the frictional force", or indeed a judgement such as "that person is happy". The *singularity principle* states that only one mental model is considered at a time. Which model is chosen for consideration is based on the perceived relevance of the model to the current task. This is a statement of the *relevance principle*. One key aspect of this default model is that it is accompanied with a value judgement about how plausible the model is. This is referred to elsewhere in the literature as a "feeling of rightness" (Thompson, 2009), a measure of how confident a reasoner is that the model is the correct model appropriate for the task at hand. If the feeling of rightness is strong, process 2 may only be engaged superficially, if at all, before a final judgement is made. If the feeling of rightness is not strong, however, an analytic intervention is triggered and only then does process 2 comes into play in a non-superficial way.

Process 2, the analytic process, is responsible for running mental simulations (explicit reasoning) using the model, and it primarily attempts to ascertain whether the model truly is satisfactory for the task at hand. This point is called the *satisficing principle.* Thus, process 2 becomes mostly a hypothetical or reflective process with an aim of validating, if possible, the process 1 model. As a result, reasoning biases such as confirmation bias (Nickerson, 1998) can enter into the reasoner's thinking and decision-making. Because process 2 utilizes working memory and is effortful, it is also susceptible to errors in reasoning such as performing an algorithm incorrectly. If the analytic process determines that the initial model is insufficient to the task, the process searches for alternate models and possibilities, and the process is repeated.

Using this theory we can derive implications for student behavior on a qualitative inferential reasoning task in physics. One implication is that task features (such as intersection points on graphs) and task goals (such as "speed over accuracy" (Heckler & Scaife, 2014)) have a large impact on which model becomes the default model in a given context because of the relevancy principle.

Since reasoning occurs using one model at a time (the singularity principle) and alternate models are considered only if the initial default model is deemed unsatisfactory by explicit reasoning (the satisficing principle), an analytic intervention is unlikely to be triggered in a meaningful way without either (1) a general disposition of reflectiveness (Tishman, Jay, & Perkins, 1993) which engages the analytic system out of habit or (2) sufficient evidence to question the relevance of the default model (*i.e.,* a decreased feeling of rightness in the initial model). Studies such as (Wood, Galloway, & Hardy, 2016) that correlate proficiency at reflecting on intuitive responses (*i.e.,* the skill of "cognitive reflection" (Frederick, 2005)) with course success are addressing the first issue. This work addresses the second issue.

For reasoners whose default model is incorrect, the intervention of the analytic process is necessary but not sufficient; they must also have the relevant conceptual knowledge to correctly solve the problem, otherwise there will not be an adequate alternate model for consideration. Thus, a productive analytic intervention requires both that the analytic intervention be triggered in a meaningful way *and* that the student possesses the relevant conceptual knowledge.

At this point, we wish to bring greater definition to some of the terms we have been using. We understand relevant conceptual knowledge to be more than a single mental model. Instead, we view conceptual knowledge as, in the words of Stanovich, "mindware" - a collection of "rules, knowledge, procedures, and strategies that a person can retrieve from memory in order to aid decision making and problem solving." (Stanovich, 2010, pg. 40). Additionally, we wish to draw a distinction between automatic, bottom-up processes that influence type 1 reasoning and the reasoning strategies and procedures used by process 2. The former have domain-general impact, meaning that they influence regardless of context (though to varying degrees based on how context mediates the effect); the latter, however, are explicit and tied closely to specific conceptual models and are therefore included in the “mindware” associated with the model.

We now summarize these points as a working hypothesis for this paper:

> *An analytic intervention that results in abandoning the default model is more likely to occur in the presence of both (1) information that refutes the default model as opposed to information that promotes alternate models and (2) a satisfactory alternate model associated with correct mindware.*

A corollary to this hypothesis is the following:

> *Information that promotes alternate models is likely to be incorporated into reasoning based on the default model (even if that information is inconsistent with that model) rather than causing a student to abandon the default model.*

Together, this working hypothesis and corollary provide the theoretical basis for the experiments described in this article.

Several research questions, both general and specific, guided this investigation. Can reasoning chain construction tasks be used in order to explore the extent to which dual-process theories of reasoning can successfully predict student reasoning and performance on certain physics questions? In particular, can reasoning chain construction tasks be used to examine previously untested aspects of these dual-process frameworks for reasoning? The specific research questions that emerged from these overarching research questions are listed below.

1. How, if at all, does providing students with statements of relevant and correct conceptual understanding impact student performance on a physics question containing one or more salient distracting features?

2. How, if at all, does providing students with a statement that refutes the common incorrect model impact student performance on a physics question containing one or more salient distracting features? Does the impact of this statement on student performance depend on whether or not students possess the relevant mindware?

3.5 Methodology and experimental design

In order to make progress in developing instructional materials that support students in the development of expert-like reasoning strategies, it is first necessary to better understand the interplay between domain-general and domain-specific processes. As such, new methodologies that help both disentangle reasoning approaches from conceptual understanding and foreground domain-general reasoning phenomena are critical for advancing our understanding of the role of these phenomena in physics reasoning. In this section, we present one such methodology and describe two experiments that highlight the affordances of the methodology in service of probing the extent to which dual process theories of reasoning can describe student reasoning in physics.

3.5.1 A new methodology: The reasoning chain construction task

The methodology we developed and employed involves what we call a *reasoning chain construction task*, or simply a *chaining task*, which allows students to focus on arranging conceptual knowledge into a logical progression of inferences. We accomplish this using a modified card sorting task in which we: (1) provide the student with a list of reasoning elements; (2) indicate that all of the statements within these elements are true and correct; and (3) ask the student to construct a solution to a physics problem by selecting elements from the list, ordering them, and, as needed, incorporating provided connecting words (“and", “so", “because", “but"). The reasoning elements primarily consist of observations about the problem setup, statements of physical principles, and qualitative comparisons of quantities relevant to the problem, all of which are true. Everything the student needs to produce a complete chain of reasoning is present in the elements; the student’s task is then to pick from given conceptual pieces and directly assemble a reasoning chain.

We have found the reasoning chain construction task to be useful in providing insight into the processes by which students generate a chain of qualitative inferences in a variety of ways. For instance, some physics tasks require only a few steps to arrive at a correct answer (*e.g.,* a qualitative question that can be solved via a short, linear chain of elements like the one shown in Figure 3-1.b), while others require the student to combine two independent lines of reasoning (*e.g.,* synthesis problems such as those reported by (Ibrahim, Ding, Heckler, White, & Badeau, 2017)); by casting each of these types of questions as a chaining task, we can obtain information about how students approach these different scenarios. We have previously interpreted results from chaining tasks through a dual-process perspective (Speirs, Ferm Jr., Stetzer, & Lindsey, 2016), and here we utilize dual-process theories of reasoning to make and test predictions about student behavior on chaining tasks. Additionally, Chapter 4 will report on the utility of network analysis techniques on data derived from chaining tasks.

Reasoning chain construction tasks have primarily been implemented online using Qualtrics' "Pick/Group/Rank" question format. This online format is illustrated in the context of a graph task and is shown in Figure 3-4. Reasoning elements from the "Items" column, connecting words, and final conclusions can be dragged and dropped into the "Reasoning Space" box; the box increases in size vertically as elements are added.

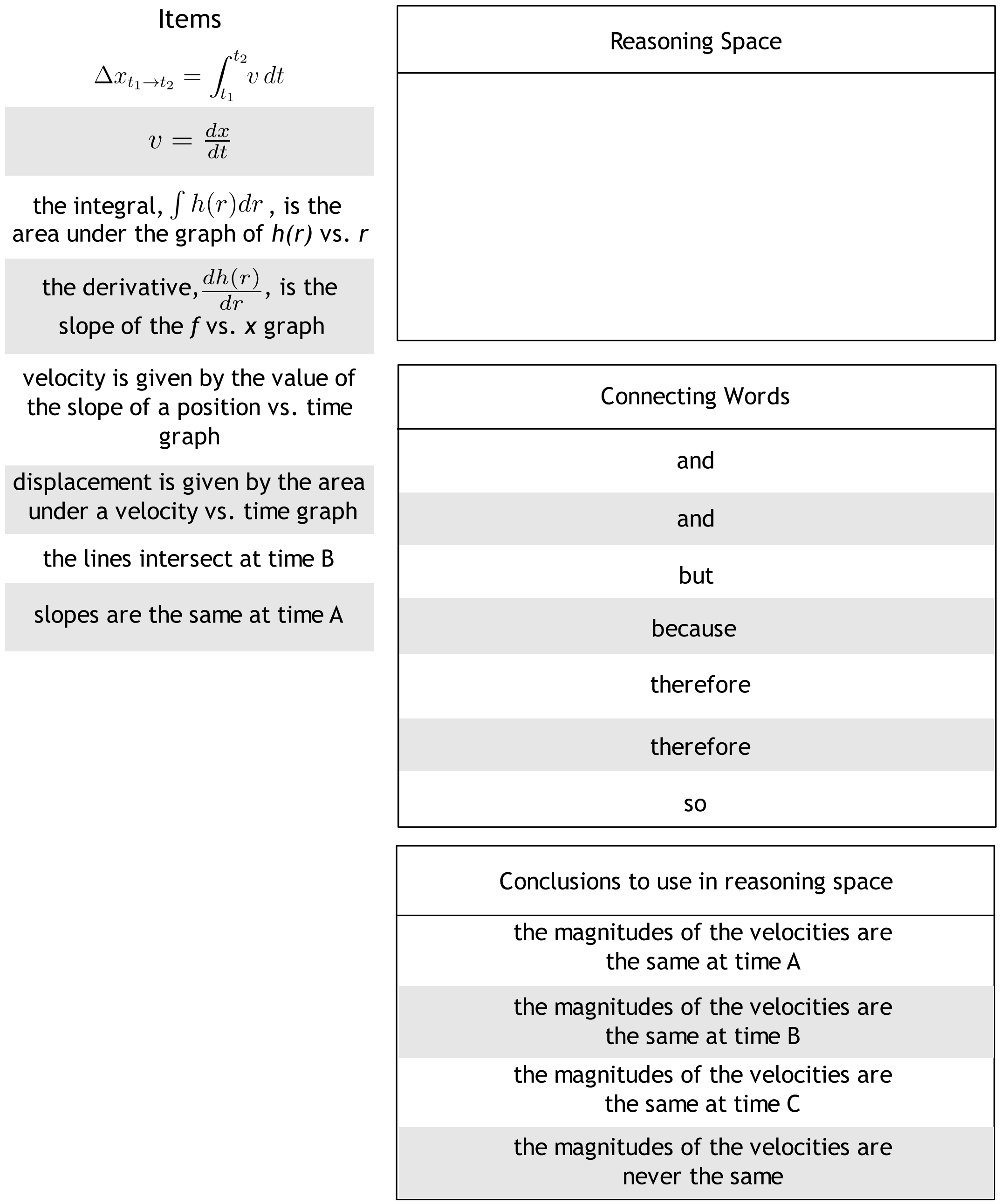


Figure 3-4. Example of how a chaining task appears to the student using the online survey platform Qualtrics' "Pick/Group/Rank" question format.

These tasks were administered on homework assignments or exam reviews for students enrolled in an introductory calculus-based physics sequence, along with other questions relevant to the course but not relevant to the content found in the research task. These assignments counted for participation credit in most cases, though in some cases extra credit was awarded. In all cases, the tasks were administered after relevant lecture, laboratory, and small-group recitation instruction at a research university in New England. Research-based materials from *Tutorials in Introductory Physics* (McDermott & Shaffer, 2001) were used in the recitation section.

The reasoning elements provided to the student are based on previously obtained student responses to open-ended, free-response versions of the task. Elements consisted of statements of first principles, observations about the task, and statements that are derived from first principles and observations. Some were productive to the correct line of reasoning, and some were not. Among the unproductive elements were elements which, while true, were useful primarily in constructing the common incorrect line of reasoning. In addition, the extent to which students selected unproductive elements not associated with the common incorrect line of reasoning could help us gauge the likelihood that students were simply inserting elements at random. Three blank elements labeled "Custom:" were provided, with instructions that if students felt they wanted to add something not represented among the given reasoning elements, they could use the text box attached to the custom element to create their own reasoning elements.

An affordance of an online chaining task is the ability to track the progression of a students' work in the reasoning space. Using JavaScript, we added functionality to Qualtrics to capture the contents of the reasoning space whenever there was a “mouse up” event as the students engaged with the task. A mouse up event is a construct within the JavaScript language that triggers when a pointing device button is released within the window of the webpage. If a mouse up event occurred, but the reasoning space had not changed (*i.e.,* if there was nothing added or rearranged in the space), we did not record the contents.

The experiments will be now briefly summarized, and then greater detail and results will be given in a following section.

3.5.2 Experiment 1A and 1B: Providing information that promotes alternate models

Experiments 1A and 1B test the hypothesis that information that promotes alternate models is not enough to productively help students disengage from a default model. These experiments also test the corollary that if a default model is not abandoned, the information would instead be used to justify the default model, even if it appears inconsistent to an expert.

For Experiment 1A, we cast the kinematics graph task (KGT) used by Heckler, 2011 (see Figure 3-1.b) as a reasoning chain construction task. We also developed two screening questions that were meant to gauge whether a student possessed an ability to determine the magnitude of an object’s velocity from a position vs. time graph. These two screening questions are shown in Figure 3-5.

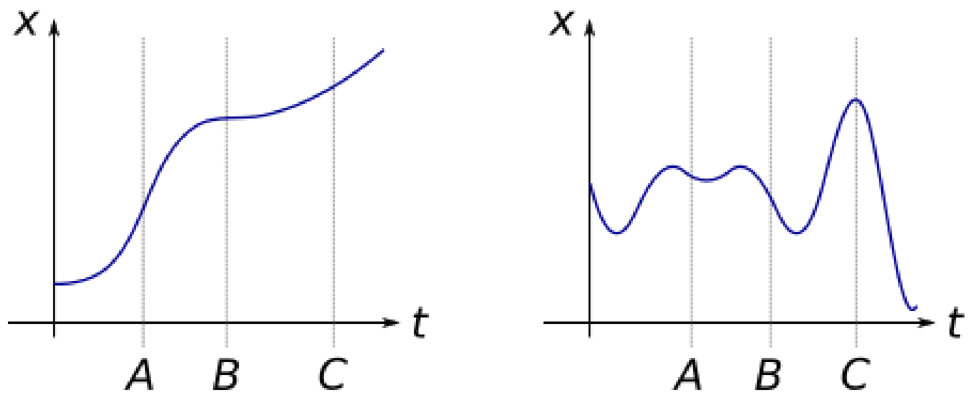


Figure 3-5. Screening questions used to gauge ability to determine the magnitude of velocity from a position vs. time graph. Each graph was shown to the student along with the prompt, “At which of the three labeled times is the magnitude of the velocity (i.e., the speed) of the car the greatest?”

In the design of the experiment, students who participated in the online exam review were randomly placed in a treatment or control condition. In the treatment condition, students were given the chaining task version of the kinematics graph task; in the control condition, students were given the kinematics graph task in a more standard multiple-choice format followed by a prompt to explain the reasoning they used to arrive at an answer. All students were given the screening questions in the multiple choice with explanation format. Since we wanted to ensure that the act of doing the screening questions would not interfere with student performance on the kinematics graph task, thereby ensuring that student performance data on the KGT in the experiment could be compared with KGT data from previous semesters, the screening questions were placed after the kinematics graph task and separated from it by a several exam review questions on unrelated topics.

Experiment 1B tested the domain-general nature of the salient distracting feature and was meant to further examine the hypothesis that information that promotes alternate models would not cause students to abandon the default model. In Experiment 1B, three graph tasks isomorphic to the kinematics graph task were devised in the contexts of mechanical potential energy, electric potential, and magnetic flux. Each task uses the same plot with the intersecting graphs, and the wording in the plots was kept as parallel as possible while reflecting the new contexts. Additionally, the reasoning elements provided on the kinematics graph task were altered to reflect the new context but were otherwise parallel and isomorphic in structure to those on the kinematics task. The problem statements and reasoning elements for these three tasks are provided in the appendix. Isomorphic screening questions were similarly constructed.

The design for Experiment 1B was the same as that for Experiment 1A: students were randomly placed in a treatment (chaining task) condition or a control (regular format) condition. In each case, the screening questions were placed after the graph task and separated from it by multiple questions on unrelated topics. Given that the four graph tasks were all administered across a single academic year, most students who completed the introductory calculus-based sequence would have seen and completed multiple, and likely all four, tasks.

3.5.3 Experiment 2A and 2B: Providing information that refutes the default model

Experiment 2A was designed to test the main working hypothesis that providing information that refutes the default model will be more productive than information that supports alternate models. In this experiment, we cast the two-box friction task from Kryjevskaia, Stetzer, and Lê (2015) (see Fig. 2_b) as a reasoning chain construction task and randomly split the students into treatment and control conditions. Both conditions utilized the chaining task version of the friction task, but in the treatment condition, a single element was added to the list of reasoning elements provided to the student. This element indicated that "the coefficient of static friction is not relevant to this problem" and was designed to call into question student satisfaction with the common, incorrect default model.

In experiment 2B, we included the screening question (in regular format) reported by Kryjevskaia, Stetzer, and Lê (2015) and shown in Figure 3-2.a before the chaining task. This allowed us to test the hypothesis that correct mindware is required for a productive engagement of the analytic process that leads to the selection of an appropriate alternate model. In the experiment, we operationalized possessing the correct mindware as answering the screening question correctly with correct reasoning; namely, such students demonstrated in at least one context that they were able to generate the correct line of reasoning needed to answer the target question. Examining the impact of the analytic intervention element in both the presence and absence of requisite mindware (as indicated by performance on the screening question) will allow us to determine the impact (or lack thereof) of possessing relevant mindware.

3.6 Experiments 1A and 1B: Graph tasks, predictions and results

In Experiment 1A, we cast the kinematics graph task (KGT, shown in Figure 3-1.b) as a chaining task, with the reasoning elements shown in Table 3-1 provided to the students. As previously stated, elements consisted of statements of first principles (such as "$v = dx/dt$"), observations about the task (such as "the slopes are the same at time A"), and statements that are derived from first principles or observations, such as "velocity is given by the value of the slope of a position vs. time graph". In the list, there are elements productive to the correct line of reasoning as well as elements that are true but (logically) irrelevant to that line of reasoning.

$\Delta x_{t_1 \to t_2} = \int_{t_1}^{t_2} v(t)dt$
$v = dx/dt$
the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*
the derivative, $dh(r)/dr$, at a specific point is the slope of the tangent line of the *h(r)* vs. *r* graph at that point
velocity is given by the value of the slope of a position vs. time graph
displacement is given by the area under a velocity vs. time graph
the lines intersect at time B
the slopes are the same at time A
the magnitudes of the velocities are the same at time A
the magnitudes of the velocities are the same at time B
the magnitudes of the velocities are the same at time C
the magnitudes of the velocities are never the same

Table 3-1. Reasoning elements provided to the students on the kinematics graph task. Elements productive to the correct line of reasoning are shaded.

There is a logical structure inherent among the productive elements provided (shaded gray in Table 3-1). While at first glance, it may appear that the elements "$v = dx/dt$", "*the derivative,* $dh(r)/dr$*, at a specific point is the slope of the tangent line of the h(r) vs. r graph at that point*", and "*velocity is given by the value of the slope of a position vs. time graph*" are equivalent and interchangeable statements, they actually constitute a logical argument justifying why the slope is the velocity: the two elements "$v = dx/dt$" and "*the derivative[…] is the slope…*" combine to imply the third element. We refer to the element, "*velocity is given by the value of the slope of a position vs. time graph*", as a derived heuristic because it represents a chunked knowledge piece (National Research Council, 2000) that is derived from two independent principles. While it would be acceptable to many instructors if students were to simply use the "slope is velocity" heuristic, all three elements are needed to provide a logically sound argument. Their inclusion, then, provided an opportunity for additional insight into student reasoning.

Both screening questions asked students to determine the time at which the magnitude of velocity was the greatest. The screening questions, shown in Fig. 5, contained distractors that tend to elicit slope/height confusion and difficulty interpreting a negative vs. a positive slope. We operationalized understanding how to obtain the velocity from a position vs. time graph as answering both screening questions correctly. Indeed, students who answered both questions correctly demonstrated that they possessed a functional understanding of velocity sufficient to compare velocities by finding and comparing slopes on position vs. time graphs.

3.6.1 Predictions

We hypothesized that a student will not abandon a default model unless there is sufficient reason to question the satisfaction of that model, and as a corollary, that information promoting models other than the default model would be recruited to defend the default model rather than change it. This hypothesis leads to specific predictions for student behavior on the chaining tasks in experiment 1A.

The high-salience intersection point results in many students in the population to embrace a default, intersection-cued model, leading to an answer of time B (the time at which the two graphs intersect). For such students, the elements productive to the correct line of reasoning are in promotion of an alternate model, and there are no elements that explicitly refute the default model of the intersection point. Thus, one prediction drawn from our hypothesis is that explicitly providing the reasoning elements associated with a correct line of reasoning will not largely increase performance on the task.

Because the high-salience of the intersection point affects process 1 reasoning and is not necessarily connected with models based in physics content, we would expect the default model to be associated with the intersection regardless of whether or not someone possessed a robust understanding of how to obtain the velocity from a position vs. time graph. Because that understanding will not likely be explored without dissatisfaction with the default model, we would expect that a lack of shift in performance will also hold among the subset of students who correctly answer both screening questions.

Finally, because of the satisficing principle, if the default model is not abandoned, process 2 will likely utilize formal reasoning to justify the default model – even if that reasoning is logically flawed or inconsistent with other reasoning provided by the student elsewhere. Thus, elements productive to the correct line of reasoning would likely be incorporated into the reasoning chains of students who answer incorrectly.

In summary, on the basis of our hypotheses, we made the following predictions for Experiment 1a:

> Prediction 1) The reasoning elements provided will not be sufficient to produce a large increase in student performance on the kinematics graph task.
>
> Prediction 2) Prediction 1 will hold even in the case of those who demonstrate relevant prior knowledge by answering both screening questions correctly.
>
> Prediction 3) Productive reasoning elements will be endorsed by students who select time B, the answer associated with the default, intersection-based model.

3.6.2 Results and discussion

In this section we review results from Experiment 1(A). We first examine and discuss the general performance on the graph task and then consider the results from the screening questions.

3.6.2.1 Performance

Student performance data on the chaining version of the kinematics graph task from a single semester is shown in Table 3-2, along with data from the multiple choice with explanation version of the task administered to the same class. As can be seen in Table 3-2, there is a statistically significant but small positive shift in the performance ($p = 0.03$, $V=0.1$), suggesting that the presence of correct, relevant reasoning elements alone is not enough to produce a large shift in performance. For reference, Heckler (2011) reported that 60% of students gave the correct response, whereas 40% of students chose time B (the intersection point). In Heckler's study, students did not have the option of indicating that the slopes were never the same.

	Percentage of total responses	
	KGT MC w/explanation format (N=158)	KGT chaining format (N=149)
Time A (correct)	**44%**	**57%**
Time B (intersection)	30%	29%
Time C	1%	0%
Never	24%	14%

Table 3-2. Student performance data from two versions of the kinematics graph task (KGT) administered as part of Experiment 1A. The task itself is shown in Figure 3-1.b. There is a small increase in performance on the chaining format in comparison with the multiple-choice with explanation format ($\chi^2 = 7.31, df = 2, p = 0.03, V = 0.1$).

3.6.2.2 Discussion of performance results

Student response data shows that the presence of relevant, correct information was not enough to produce a large, positive shift in performance on this task. This may be hard to explain from a perspective that highlights the construction or possession of incorrect models as the primary reason for incorrect answers.

Indeed, taking the perspective that students who answer the physics questions incorrectly are utilizing an incorrect model of a physics concept, one might predict that giving students statements of relevant knowledge would increase performance. For instance, it has been argued that students who select the intersection in this kinematics question lack a conceptual understanding of velocity, are drawing upon incorrect ideas about velocity or are cued to construct incorrect knowledge around p-prims such as “same is same”. By providing the relevant conceptual elements, one might predict that performance should increase substantively because students may now draw upon these elements, which might help them refine their understanding of velocity, address the incorrect concept, or give them an alternate cue around which they can construct their knowledge and argument. However, because there are not well-defined mechanisms for what specific knowledge is constructed in any of these cases, no firm prediction can be made.

Dual-process theories of reasoning, however, make a firm prediction because they give more definition to the control mechanisms by which models are chosen for consideration as well as the conditions under which they would be abandoned in favor of alternate models. In this case, an incorrect model based on the intersection point drew some students to the time B answer. In order for students to switch away from this default answer, an analytic intervention would need to be triggered (*i.e.,* a productive engagement of process 2) resulting in a loss of confidence in this answer. However, the analytic system is primarily concerned with running simulations based on the original model; thus, it is more likely that a student would come up with physics-like justifications of an incorrect answer than that they would change the model itself to arrive at a different answer. The presence of correct information alone, then, would not be expected to produce the level of dissatisfaction required to prompt an exploration of alternate models. This is consistent with prediction 1 articulated in section 3.6.1.

3.6.2.3 Results from Screening Questions

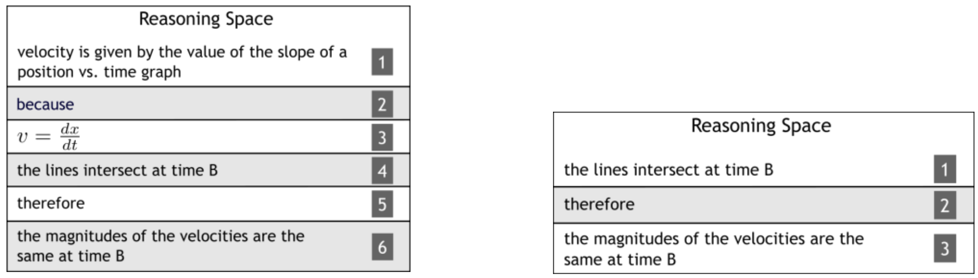
According to prediction 2 from section 3.6.1, we would expect that the even among those students who demonstrate functional knowledge of how to obtain the magnitude of velocity from a position vs. time graph on the screening questions, their performance on the KGT would not largely improve upon increased access to relevant conceptual knowledge. We would thus expect that the intersection point would still be a prevalent incorrect answer among those who have previously demonstrated the requisite knowledge.

Overall, student performance on the screening questions (see Figure 3-5) was rather strong. Ninety-six percent of students correctly answered screening question 1, 83% of students correctly answered screening question 2, and 82% correctly answered both. It is worth noting that the screening questions included a distractor consistent with slope-height confusion. In both questions, time C had the greatest height. This answer was not prevalent in screening question 1 but comprised 17% of student responses to screening question 2. It is surmised that the shape of the graph contributed to this difference in prevalence of responses indicative of slope-height confusion, with the sharpness of the curve at time C in question 2 possibly being more salient than the smooth curve at time C in question 1. This speculated difference in salience is consistent with previous research on salient distracting features in graphs (Elby, 2011).

For those students who answered both screening questions correctly, the observed increase in performance was statistically significant but with a small effect size ($p = 0.03$, $V = 0.1$). Additionally, as shown in column 1 of Table 3-4, 22% of students who answered both screening questions correctly ultimately chose time B on the KGT, which corresponds to the intersection point. This is consistent with prediction 2 described earlier.

3.6.2.4 Analysis of reasoning chains

The chaining format affords students an opportunity to employ reasoning elements that they otherwise might not consider using. According to the dual-process framework, we predicted that such reasoning elements would likely be used in conjunction with the default answer put forward by process 1, even if the element itself was inconsistent with the default answer (prediction 3). This prediction proved to be correct; many students who chose the common incorrect answer used elements in their chain that represented reasoning that, to an expert, is more closely related to the correct line of reasoning.



(a) (b)

Figure 3-6. (a) A student endorses information more closely associated with the correct answer in the process of justifying the common incorrect answer. (b) Another student answers with only the observation that the lines intersect at time B (a "canonical response").

As a specific example, consider the student response shown in Figure 3-6. The first three elements, "velocity is given by the value of the slope of a position vs. time graph / because / $v = dx/dt$" are logically connected in a way that, to an expert, suggests an understanding of the underlying physics. With that point of view, this student is clearly endorsing correct conceptual information before abruptly shifting toward the incorrect answer associated with the salient distracting feature.

To study this phenomenon in greater detail, criteria were developed to gauge the extent to which students who both chose the intersection (time B) and endorsed productive elements were demonstrating understanding of the underlying physics. In doing so, we rely on the assumption that including elements in the reasoning space is a tacit endorsement from the student of the usefulness or relevancy of that element.

The most rigorous criterion required the student to use 2 or more of the 3 elements that comprise the "velocity is slope" triad explained above. An example of this is shown in Figure 3-6. In all cases in which a student satisfied this criterion, it was clear that they were linking the elements together logically. Of those students who answered time B, 7% of their reasoning chains met this requirement.

The second, more relaxed criterion contends that any student who uses at least one of the three elements (without using irrelevant elements) is endorsing correct conceptual information. This is appropriate given that the derived heuristic, "velocity is slope", element is commonly the only element used in supporting a correct answer. It also represents correct information that is likely to occur to a student because of possible repetition during normal classroom instruction. Relaxing the requirements to this level of constraint raises the proportion of students who chose time B and also certified correct information to over 50%. These results are summarized in the chart shown in Figure 3-8. Thus, we were able to generate both upper (50%) and lower (10%) bounds on the extent to which students who chose the intersection and also endorsed productive elements were demonstrating some level of understanding of the underlying physics.

The fact that between 10% and 50% of students supposed the common incorrect answer by endorsing information more closely aligned with the correct line of reasoning indicates a sort of cognitive conflict between learned information and an intuitive answer generated by process 1. Our prediction above was that students who chose time B, when confronted with improved access to knowledge relevant to the correct line of reasoning, would choose to incorporate that very knowledge into a reasoning chain in support of the incorrect answer. This prediction (prediction 3) proved to be correct.

Students who did not endorse these productive elements typically responded by only using the elements “the lines intersect at time B” and “the magnitudes of the velocities are the same at time B” (44% of those students who selected time B as their answer) or else they endorsed elements that were irrelevant (< 1% of those students who selected time B). Given its ubiquity in the chaining format versions of the KGT as well as its prevalence in the explanations in the multiple-choice with explanation version of the KGT, we refer to the former class of responses as the “canonical incorrect answer”, illustrated in Figure 3-6.b and reported in Figure 3-8.

3.6.2.5 Reasoning chain analysis using move tracking

Using the added functionality described above to capture the contents of the reasoning space anytime there was a “mouse up” event, we were able to obtain data about which elements were placed in the reasoning space and at what times they were inserted or moved as each student worked through the chaining task. In the remainder of this article, we call this “move tracking”.

The move tracking data revealed another pattern in student responses: some of those students who answered correctly placed the “time B” answer element (the answer associated with the intersection point) into the reasoning space before changing their answer to another option.

Figure 3-7 shows a representation of this behavior. The column on the left shows the *answer option* that was placed in the reasoning space first (regardless of whether another element was already in the box). An arrow connects this answer option to the answer option that was in the reasoning space when the student completed the task (the right column). (This figure is similar to an alluvial diagram, which shows how different entities flow and transform with time.) Each arrow represents one student, and students who did not switch their answer options are not shown in the diagram. Thus, if student A initially thought the answer was “never the same” and put that element in the reasoning space, but then while looking at the other reasoning elements decided that time A was the answer and replaced “never the same” with the time A element, that student would be represented in Figure 3-7 as one of the arrows going from “never the same” in the left column to “time A” in the right column.

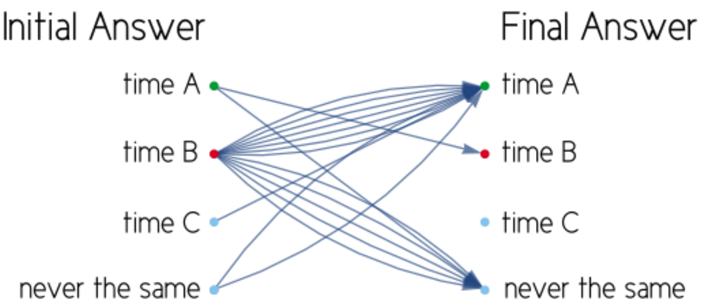


Figure 3-7. Answer switching measured via tracking the movements of the reasoning elements while students completed the task.

This “switching graph” reveals a tendency on the part of students to select time B as an initial answer, and to then shift away from this answer, suggesting that even those who answer correctly may initially be taken in by the salience of the distracting feature. The 12 students shown switching from “time B” to some other answer represent 8% of the total population. Given the manner in which we capture this answer switching data, this number most likely underrepresents the actual percentage of students who are, at least initially, cognitively drawn toward “time B” as an answer before pivoting away from it mentally while thinking about the task. Our analysis only captures those students who provide explicit evidence of this switch in the reasoning space.

It is tempting to think that this 8% could account for the increase in performance of about 10% on the graph task and that the answer switching seen above was catalyzed by the presence of the reasoning elements in the chaining format. There are two reasons that make this less unlikely. First, there are about the same number of students shifting from “time B” (the intersection point) to the “time A” answer (the correct answer) as there are shifting from “time B” to “never the same”. Secondly, the overall increase in the performance didn’t alter the percentage of students choosing “time B”, as shown in Table 3-2.

Our data corpus does not provide any explicit connection between this switching behavior and the reasoning elements provided. However, even if it were catalyzed by the presence of the reasoning elements, the point still stands that there was, in fact, a subset of students who ultimately arrived at the correct answer but who were originally invested in the “time B” intersection answer, as predicted by our dual-process framework.

This phenomenon as well as others described above suggest a spectrum of impact of the intersection point, possibly based on each student’s “feeling of rightness”. For some, the feeling of rightness about the intersection point answer is low, and these students abandon that model with little or no prompting. For others, the feeling of rightness may be moderate, and so process 2 engages presumably relevant mindware. Some of these students have correct physics conceptions that they struggle to reconcile with their intuitive answer. Others may construct incorrect conceptions that they use to justify their response. Still others may have such a strong feeling of rightness that they engage process 2 only on a surface level to ratify the process 1 answer. With a more refined methodology utilizing chaining tasks, it may be possible to tease out the relationship between behavior on a chaining task and the feeling of rightness in the initial model.

3.6.3 Experiment 1B: Isomorphic graph tasks

Based on dual-process theories of reasoning, the intersection point present in the KGT should cue the same default judgment even in contexts outside of the kinematics context due to process 1 relying on the salient features of a task when selecting an initial model. Indeed, Heckler & Scaife (2014) used math graphs, kinematics graphs, and graphs of electric potential to demonstrate that processing time had an effect on answer patterns for questions regarding the slope of a graph independent of context. While context and content mediate the effects of domain-general factors, these factors are still at play. For instance, in Heckler and Scaife's work (2014), the effects of processing time were less pronounced in more familiar contexts but were still present. Likewise, the working hypothesis of this paper (*i.e.,* that access to relevant conceptual information would not be sufficient to abandon a default model) should be operative regardless of specific physics content.

To test this hypothesis, three additional chaining tasks were devised. These tasks were structurally parallel to the kinematics graph task to the greatest extent possible but were in the contexts of potential energy, electric potential, and magnetic flux. Each context has a correct line of reasoning that relies on an understanding that the desired quantity can be obtained from the derivative of the known quantity, and thus the slopes of the graphs at the point of interest ought to be compared. We constructed screening questions that would indicate the extent to which the students possessed an ability to determine the desired quantity from the slope in the absence of the salient distracting feature. The reasoning elements provided to the student in each task were modified to fit the context but remained isomorphic in their structure. All four graph tasks (including the KGT), the reasoning elements provided on the chaining version of each task, and the screening questions are all included in Appendix A.

All tasks were administered after relevant course instruction was completed in class. Given the contexts associated with these isomorphic tasks, data were collected in both semesters (fall and spring) of the on-sequence calculus based introductory physics course. The experimental design was the same as for the kinematics graph task in that a between-student design was employed with the treatment condition corresponding to the chaining version of the graph task, and the control condition corresponding to a multiple-choice with explanation version of the graph task.

Given the similarity in experimental design, we expected all three predictions made for experiment 1A to hold for experiment 1B as well. Namely, we predicted

Prediction 1) The reasoning elements provided will not be sufficient to produce a large increase in student performance on any graph task.

Prediction 2) Prediction 1 will hold even in the case of those who demonstrate relevant prior knowledge by answering both screening questions correctly.

Prediction 3) Productive reasoning elements will be endorsed by students who select the answer associated with the default, intersection-based model.

The three additional graph tasks serve the purpose of generalizing results. If the predictions held across all three additional contexts, our results would provide further evidence that the observed phenomena on the KGT are truly driven by domain-general reasoning phenomena.

3.6.4 Experiment 1B: results and discussion

In this section we review results from Experiment 1(B). We first examine and discuss the general performance on the graph task and then consider the results from the screening questions.

3.6.4.1 Performance

The results from all four isomorphic graph tasks are summarized in Table 3-3. There is little or no statistically significant improvement in student performance (*i.e.,* more correct time A responses and fewer intersection or time B responses) on the chaining version in comparison to that on the multiple-choice with explanation version for three of the four graph tasks. These results suggest that, in general, providing greater access to relevant physics concepts does not increase performance. It is important to note, however, that the electric potential graph task exhibited a positive, medium effect-size improvement in performance on the chaining version in comparison to the control version. We discuss this discrepancy in the next section.

Context	Kinematics		Potential Energy		Electric Potential		Magnetic Flux*	
Format	CG	MC + Exp.	CG	MC + Exp.	CG	MC + Exp.	CG	MC + Exp.
N:	149	158	76	80	97	121	88	83
Time A	**57%**	**44%**	**43%**	**38%**	**73%**	**44%**	**66%**	**59%**
Time B	29%	30%	51%	58%	21%	45%	28%	40%
Time C	0%	1%	1%	0%	1%	3%	5%	0%
Never	14%	24%	4%	5%	5%	8%	1%	1%
(p,V)	(0.03,0.1)		(0.75,0.04)		(0.001,0.21)		(0.34,0.07)	

Table 3-3. Performance comparison between control (multiple choice with explanation) and treatment (chaining format) for each graph task. *Data collected from the previous year for Magnetic Flux task. See text for details.

Given a different experiment we were conducting as part of our broader investigation, it was not possible to collect truly analogous multiple-choice with explanation data for the magnetic flux task. As such, data collected the previous year from both versions (treatment and control) of the isomorphic flux graph task are included in Table 3-3. However, the results are similar to those collected for the flux task administered in the same year as the other three tasks.

Chaining format results for those students who answered both screening question correctly are shown in Table 3-4. The intersection point still tends to be a common incorrect answer, even in the electric potential task, with around 25% of the population picking “time B”.

Context	Kinematics	Potential Energy	Electric Potential	Magnetic Flux
N:	122	38	76	90
Time A (Correct)	**63%**	**58%**	**75%**	**73%**
Time B (Intersection)	22%	34%	17%	22%
Time C	0%	3%	1%	0%
Never	15%	5%	7%	4%

Table 3-4. Performance data for the isomorphic graph tasks in chaining format for those students who answered both of the corresponding screening questions correctly. Data from magnetic flux graph task are drawn from the same year as the other three tasks.

3.6.4.2 Discussion of performance results

With the exception of the electric potential graph task, there is little to no positive shift in performance from the control to the experimental condition, with the only statistically significant improvement being of small effect size. The lack of sizable performance shift among three of the four graph tasks strengthens the claim that improved access to relevant conceptual information does not automatically improve performance. The impact of the reasoning elements on student performance on the electric potential graph task is of note in that the increase in performance is of medium effect size. The impact of the reasoning elements appears to be specific to the topic of electric potential, but we are unsure of the specific cause. However, the domain-general nature of the salient intersection is still apparent in the control (MC with explanation) condition, and, to a lesser extent, in the treatment (chaining format) condition, as evidenced by the prevalence of time B answers for that task.

Through the use of the screening questions, in combination with the chaining versions of the isomorphic graph tasks, we were able to ascertain that the predicted process 1 default answer was still present even among those who answer both screening questions correctly and are given the relevant conceptual information in the chaining task. In other words, students who previously demonstrated the functional knowledge needed to obtain the relevant quantities from a graph and who were provided reasoning elements that might cue them toward another model still answered consistent with a model based on the salient distracting feature. Since this occurs across all four different contexts, it is unlikely that this pattern stems from either student difficulties with the relevant concepts or topic-specific misconceptions. Instead, it is more likely the result of a process 1 response that is not followed up with a productive analytic intervention.

3.6.4.3 Analysis of reasoning chains: cross-task comparison

Because the element structure on each task was identical, comparison between tasks is made possible. To analyze the reasoning chains of those students who selected the common incorrect answer, we apply the same criteria discussed in Section 3.6.2.4. The result is shown in Figure 3-8. As described in Section 3.6.2.4, the "canonical" category is defined as those responses that only include the elements "the lines intersect at time B" and the "time B" answer. The other two categories give two levels of constraint regarding the usage of productive reasoning elements. In the most rigorous, a student is required to have used 2 or 3 of the 3 conceptual elements productive to the correct line of reasoning. In the more relaxed constraint, only one of the three elements is required. The percentage of these students who only used the derived heuristic is indicated by crosshatching placed over this latter, relaxed constraint. The "other" category contains students who utilized irrelevant elements, either in conjunction with productive elements or alone, or were otherwise uninterpretable.

Across all four tasks, there is a tendency for those students who answered time B on the chaining versions to endorse elements that were productive to the correct line of reasoning. Between a half and a quarter of students answering incorrectly endorsed at least one element associated with the correct line of reasoning. Interestingly, the prevalence of the "derived heuristic only" is larger in the kinematics context compared to the other three tasks. Instead, students seem to favor either listing two or three of the three triad elements or using one of the two independent principles only. This is likely related to instruction. The heuristic of finding the velocity from the slope of a position graph is more common in instruction than finding the induced EMF from the slope of flux graph; instead, when teaching flux, the emphasis is typically on the mathematical relation of Faraday's law (*i.e.,* $\varepsilon = -\frac{d\Phi_B}{dt}$).

In summary, analysis of the incorrect reasoning chains produced by students on the isomorphic chaining tasks provide further support for the prediction that the productive elements, if used at all, will be incorporated into incorrect answers despite their logical inconsistency from the perspective of an expert.

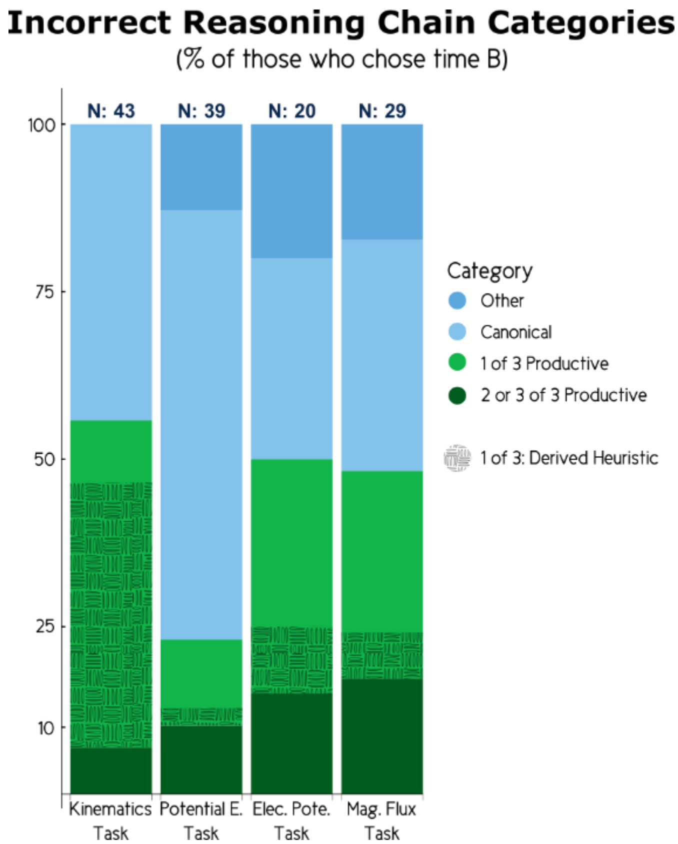


Figure 3-8. Incorrect reasoning chains categorized. Values shown are percentages of those students who selected time B as their answer. Total number of students who selected time B is indicated for each task.

3.6.5 Summary of Experiment 1B

In experiment 1A, we utilized the kinematics graph task to investigate the working hypothesis that providing improved access to relevant conceptual information would not cause students to abandon an initial incorrect model. A variety of measures on that task provided evidence for this hypothesis. The isomorphic graph tasks employed in experiment 1B resulted in student performance data and analysis of reasoning chains that support the proposed mechanisms driving the selection and abandonment of mental models. These data also establish that these mechanisms are at play in contexts outside of kinematics. The predictions drawn from the working hypothesis about student performance and behavior on reasoning chain construction tasks were shown to be correct not just in the kinematics context, but across four different physics contexts.

3.7 Experiment 2A and 2B: Friction task with added "Analytic Intervention Element"

Experiments 1A and 1B demonstrated that providing relevant conceptual information to students was not helpful in improving performance on physics graph tasks. This supported the working hypothesis that an incorrect default model would only be abandoned in the presence of information that refutes this model. In Experiments 2A and 2B we wish to provide this refutery information and determine whether a productive engagement of the analytic system occurs.

In Experiment 2A, we explored the working hypothesis in a direct way by providing an element which was intended to stimulate a more productive process 2 intervention. Process 2 reasoning is initiated by an analytic intervention triggered by a low feeling of rightness (Thompson, 2009) in the initial model and is primarily concerned with evaluating the satisfaction of the initial model. If the feeling of rightness is strong, the analytic process may not be engaged, or may be engaged only superficially. To induce a more productive analytic intervention in the context of having an incorrect default model, the feeling of rightness needs to be lowered to a point where the default model becomes unsatisfactory. In Experiment 2A, we attempted to decrease the feeling of rightness and to promote a productive analytic intervention in the context of the chaining format via a relatively modest intervention; namely, we inserted a single reasoning element into the list that explicitly refuted the common incorrect default model.

To do this, we utilized the two-box friction task described in the introduction (shown in Figure 3-2.b) and cast it into the chaining format. In the two-box friction task, students are asked to compare the magnitudes of the friction forces on two identical boxes on different surfaces. Next to each box is indicated the coefficient of friction for the box-surface pair; these coefficients are a salient distracting feature for students, resulting in a common incorrect answer based on reasoning from the coefficients alone.

The reasoning elements used in this task are shown in Table 3-5. To test the effect of an element that would attack the satisfaction of the common incorrect default model, the population was split into treatment and control groups. The treatment group received the chaining version of the friction task with the element "the coefficients of friction are not relevant to this problem" included. In this manuscript, we refer to this element as the "analytic intervention element", or AIE, because it was designed to stimulate a more productive analytic intervention by reducing the satisfaction with the model that the coefficients of static friction determine the magnitude of the static friction. The control group received a chaining version of the friction task that did not include the AIE.

$F_{net} = ma$
Both boxes have the same mass
The tension force on box A is equal to the tension force on box B
Both boxes remain at rest
Coefficient of friction for A is smaller than the coefficient of friction for B
Both boxes have the same weight
The normal force on box A is equal to the normal force on box B
Neither box is accelerating
The horizontal forces are balanced
The vertical forces are balanced
The net force on both boxes is zero
The friction force and the applied force are the only horizontal forces acting on the box
The coefficient of static friction is not relevant to this problem*
$F_{\text{frct on A}}$ is [insert relationship here] $F_{\text{app on A}}$
$F_{\text{frct on B}}$ is [insert relationship here] $F_{\text{app on B}}$

Table 3-5. Reasoning elements provided to the students on the chaining version of the two-box friction task. Elements productive to the correct line of reasoning are shaded. The final two elements had a text box where students could indicate whether the friction force was *greater than*, *less than*, or *equal to* the applied force for each box. *denotes the analytic intervention element, which was present only in the treatment condition.

To ensure that we would have sufficient statistical power to compare the experimental and control groups described above and because our intervention required the chaining format of the two-box friction task, we did not attempt to randomly assign any students to a more traditional multiple-choice with explanation format version of the two-box friction task. In section 3.7.2, however, we will compare our control results with published results on the two-box friction task (Kryjevskaia, Stetzer, & Le, 2015).

3.7.1 Predictions

In experiments 1A and 1B, we saw that the presence of elements that support a correct line of reasoning was not enough to stimulate a productive analytic search for alternate possibilities. Indeed, a significant percentage of those students who demonstrated the relevant mindware to construct a correct line of reasoning on the screening questions still drew upon a default, incorrect model of the intersection when answering the kinematics graph task and the other three isomorphic graph tasks. Moreover, of those students giving incorrect responses consistent with the default model, many incorporated productive reasoning elements into an erroneous chain.

Similarly, in Experiment 2A, we expected that providing reasoning elements productive to a correct line of reasoning on the two-box friction task would not increase performance substantially. However, we expected that the inclusion of the analytic intervention element would reduce the satisfaction of the default model and would, by implication, improve performance by causing students to switch from an incorrect default model to the correct model. Thus, our prediction for Experiment 2A is that student performance would be stronger in the treatment condition than in the control condition for the chaining version of the two-box friction task.

3.7.2 Results and discussion

3.7.2.1 Performance

Table 3-6 shows the student performance on both versions (experiment and control) of the chaining version of the two-box friction task collected in two different semesters (both on sequence and off sequence) of the introductory calculus-based mechanics course.

PHY121 Semester	Control (without AIE) % Correct	Treatment (with AIE) % Correct
On-sequence N=119/120 ($p = 0.02, V = 0.2$)	55%	74%
Off-sequence N=64/66 ($p = 0.03, V = 0.2$)	27%	38%

Table 3-6. Student performance on both versions (experiment and control) of the chaining version of the two-box friction task. The task itself is shown in Figure 3-2.b. The off-sequence course was without a fully implemented tutorial instruction. See note in text about how p-values are calculated.

While the overall performance in the on-sequence and off-sequence courses differed substantively, in both trials there was a statistically significant, medium-effect size improvement in performance in the treatment condition with respect to the control condition. This suggests that the AIE had an impact on performance over all. While only the percentages correct and incorrect are shown in the table, the p-values were derived from a chi-squared test of independence comparing the distributions of all answer choices from the treatment (AIE) condition and those from the control (non-AIE) condition. It is worth noting that the overall performance difference between the on- and off-sequence courses may possibly stem from differences in instruction (*e.g.,* differences in the implementation of *Tutorials in Introductory Physics*) and/or differences in participation rate and participation incentives among the two courses; for the purpose of our investigation, the absolute performance was of less interest than the shift in performance between treatment and control.

3.7.2.2 Discussion of Performance Results

Table 3-6 demonstrates that the AIE impacted student performance, regardless of the baseline level of understanding demonstrated by the performance of the control group from each population. Indeed, the performance of students from the off-sequence course was considerably lower, suggesting that the population differed somehow from that in the on-sequence course. Even in the off-sequence population, however, the AIE still produced a medium effect-size positive shift in performance despite the overall lower performance. The fact that we observed improved performance by the treatment group in both courses provides further evidence for the generalizability of the AIE result.

It may be surmised that the answer choice "equal" could also be arrived at using solely perceptual (non-physics) cues, especially once the coefficients are eliminated as a relevant factor. It may therefore be tempting to think that the AIE simply redirects students from the default, coefficient-based model to a purely perception-based based approach, as opposed to our interpretation that the AIE stimulates deeper examination of physics principles via a productive analytic intervention. This alternative explanation for our results will be explored more fully in Experiment 2B, in which we investigated the impact of the AIE while controlling for performance on a screening question.

3.7.2.3 Switching behavior on the two-box friction task

As on the kinematics graph task, we inserted JavaScript into the Qualtrics platform in order to capture the reasoning space every time a "mouse-up" event was triggered (as described in Section 3.5.1). Using these data, we determined when students initially put an answer element into the reasoning space that differed from the final answer element in the reasoning space when they moved to the next page. Graphs of the documented switching behavior are presented separated for treatment and control groups in Figure 3-9.

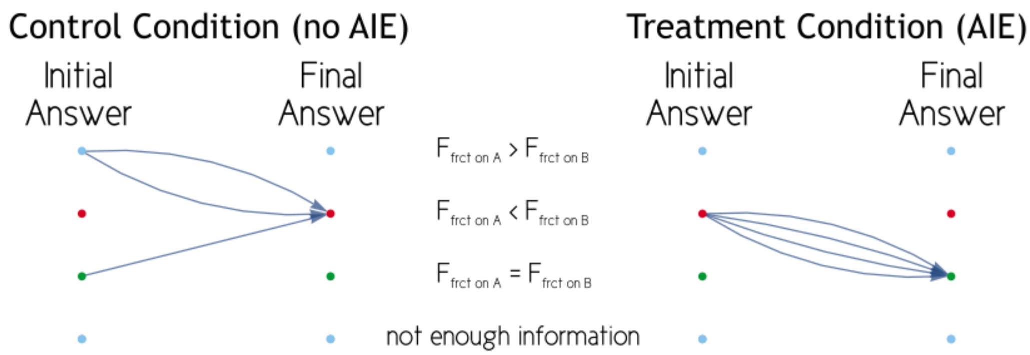


Figure 3-9. Answer switching measured via tracking the movements of the reasoning elements while students completed the two-box friction task. The switching behavior are presented separated for treatment (AIE) and control (non-AIE) groups.

As indicated in Fig. 3-9, only a few students switched their answers in the treatment and control groups. However, although the data presented in Figure 3-9 are sparse, one can see that the trend for answer switching between treatment and control groups is different. In the treatment (AIE) condition, the trend shows a shifting from the common incorrect answer (driven by the coefficients of static friction) to the correct answer, whereas in the control (non-AIE) condition, the shift is to the common incorrect. This suggests the important role that a salient distracting feature may play in impacting student reasoning as well as the apparent impact of an element that attacks reasoner satisfaction with the default model in moving students away from the model cued by the salient distracting feature. Of course, these data don't carry enough statistical power to make a solid claim (as only about 5% of the population explicitly changed their answers), so additional research would need to be conducted in a more rigorous experiment to ascertain whether or not the switching trend observed here is both reproducible and statistically significant. However, we do suspect that these graphs underreport the actual amount of switching that is occurring because many students likely switch answers without explicitly documenting these switches in the reasoning space. In any case, our data indicate that the analytic intervention element seems to influence student reasoning in a manner that helps students arrive at a correct answer – a phenomenon observed in a statistically significant manner in our experimental and control group comparisons in Table 3-6 and in a much less rigorous manner in the switching diagrams in Fig. 3-9.

3.7.2.4 Analysis of reasoning chains

It seems likely that the analytic intervention element causes students to abandon an incorrect default model in favor of a correct model. However, it could be that little to no physics knowledge is utilized when students make judgments (*i.e.,* ratify a final answer) based on the new, correct model. Instead, once the coefficients are ruled out, they may be basing their answer on the perceptual cue that everything else in the problem is *equal* for both cases: the mass, the weight, the tension, etc. (Indeed, we cannot preclude the possibility that those who were cued on the correct model initially were backfilling their formal reasoning in support of a process 1 answer derived solely by the perceptual cues.)

To further investigate whether students who are seemingly affected by the AIE are employing formal physics knowledge when answering correctly, we examined the chains that the students who gave correct responses used in the treatment condition. In both semesters, 80% of correct responses in the treatment condition exhibited chains that clearly indicated correct reasoning. Generally, these responses included an indication of Newton's 2nd law being utilized to determine that the horizontal forces are balanced on both boxes. An example is given below:

"*both boxes have the same mass / and / the normal force on box A is equal to the normal force on box B / so / because /* $F_{net} = ma$ */ and / both boxes remain at rest / the horizontal forces are balanced / and / the net force on both boxes is zero / because / the friction force and the applied force are the only horizontal forces acting on the box /* $F_{frct\ on\ A}$ *is equal to* $F_{frct\ on\ B}$"

The other 20% of responses were ambiguous; they could easily be seen as indicating correct reasoning but could also be construed as rationalization based on the features of the problem that are equal. One example is a student who responded in the following manner:

"$F_{frct\ on\ A}$ *is equal to* $F_{frct\ on\ B}$ */ because / both boxes remain at rest / and / the tension force on box A is equal to the tension force on box B*".

It is important to keep in mind that only 20% of responses were ambiguous, the rest were unambiguous. These results indicate that those who answered correctly in the treatment condition were able to answer with correct reasoning, that is, by engaging with some version of Newton's 2nd law. This suggests that for the students in the treatment condition, the answer choice "equal" was not arrived at solely by perceptual cues for most students; if it were, we would expect that a larger percentage of students would answer "equal" but lack the ability to chain together a response that indicated a complete and correct line of reasoning. Perhaps such perceptual cues served as the origin of their answer, but there is no evidence that, for any of these students, perceptual cues are the sole factor behind their reasoning and conclusion.

Our analysis revealed interesting features in the incorrect responses as well. As a specific example of a common class of incorrect responses, consider the following student argument:

"both boxes have the same mass / but / coefficient of friction for A is smaller than the coefficient of friction for B / so / $F_{frct\ on\ A}$ *is less than* $F_{frct\ on\ B}$*"*

This student responded with what would be considered a canonical incorrect answer – an answer that primarily relies on a direct judgment based on the coefficients of friction or the equation $f = \mu N$ without reference to other physics principles. Approximately 60% of incorrect responses fell into this category in each semester, in both the treatment and control conditions.

Another 30% utilized the coefficient reasoning but included other pieces of relevant information such as the observation that the boxes remained at rest. For example, one student argued:

"both boxes have the same weight / and / the normal force on box A is equal to the normal force on box B / but / neither box is accelerating / because / both boxes remain at rest / $F_{frct\ on\ A}$ *is less than* $F_{frct\ on\ B}$*"*

This response is consistent with an incorrect conception in which friction is greater than the applied force until the applied force is big enough to *overcome* that friction force. Thus, the student didn't answer purely based on the coefficients alone, but likely had some form of intervention of process 2, though one that resulted in an erroneous justification for their answer.

Other students gave responses similar to the following:

"the normal force on box A is equal to the normal force on box B / and / both boxes have the same weight / but / coefficient of friction for A is smaller than the coefficient of friction for B / so / Custom: "B needs more force to

move" / but / Custom: "since neither of them moved" / the horizontal forces are balanced / and / neither box is accelerating / and / the net force on both boxes is zero / therefore / both boxes remain at rest / but / Custom: "since the coefficient of friction for B is greater" / $F_{frct\ on\ A}$ *is less than* $F_{frct\ on\ B}$*"*

This response shows a student who appears to struggle between a desire to articulate correct knowledge and a strong default model, similar to the incorrect responses we saw on the isomorphic graph tasks (including, for example, the kinematics graph task). These responses were not prevalent (less than 2% percent of all responses) in the two semesters in which Experiment 2A was implemented. For this reason, we did not attempt to establish and evaluate such responses according to rigorous criteria in order to determine upper and lower bounds on the extent to which this type of struggle was occurring for students (as we did for the graph tasks).

Overall, the findings from our analysis of the incorrect responses fall in line with dual-process theories. In the context of our framework, those who are attracted to the salient distracting feature likely have a strong feeling of rightness. We would expect, therefore, that there would not be motivation to search for alternate models, and this seems to be reflected in the reasoning chains leading to an incorrect answer; indeed, most of them do not indicate any reflection on the answer beyond a single model built around the coefficients.

3.7.3 Experiment 2B: Description of the experiment and predictions

3.7.3.1 Description of Experiment 2B

In our working hypothesis, we stated that a productive analytic intervention would require an alternate model that was more satisfactory, and that this model would need to be associated with relevant and productive mindware. In Experiment 2A, it was demonstrated than an element that attacked student satisfaction with the default, common incorrect model successfully increased performance on the two-box friction task. In Experiment 2B, we modify Experiment 2A to test the full extent of the working hypothesis with a focus on the need for this alternate model and requisite mindware.

To gauge the effect of having, or not having, this model and associated mindware, we repeated Experiment 2A again with a single modification: the screening question originally used before the two-box friction task by Kryjevskaia *et al.* (Kryjevskaia, Stetzer, & Le, 2015) was administered to students in both conditions before they were given the chaining version of the two-box friction task. We thus operationalized student possession of the requisite model/mindware as demonstrating that knowledge on the screening task. We were then able to control for performance on the screening question and to probe the impact on the analytic intervention element on students who did and did not possess the requisite mindware.

Additionally, it was argued in a previous section that the increase in performance caused by the analytic intervention element could be due solely to redirecting students' attention to alternate, less salient features of the task. These features may lead students to a correct answer even in the absence of any reasoning directly connected to correct physics models. By including the screening question, we could determine the extent to which such a phenomenon is happening, if at all. If we found that the analytic intervention element had roughly equal impact on those students who answer the screening question correctly versus those who do so incorrectly, our study would be inconclusive with regards to whether correct physics models necessarily played a role in the documented increase in correct answers. However, if the impact of the AIE was found to be greater among students who demonstrated that they had the requisite mindware, we could conclude that those who switched did so because of relevant mindware, not solely because the default model involving the coefficients was ruled out and they were thus led to choose the next best answer based solely on task features.

3.7.3.2 Predictions

As the only change in Experiment 2B was the inclusion of a screening question prior to the chaining version of the friction task, we expected that the inclusion of the analytic intervention element would incite a strong positive performance shift, consistent with the prediction made in Experiment 2A.

Based on the criteria from the working hypothesis that a more satisfactory alternate model is necessary for a productive analytic intervention, we expected that a performance shift would occur most prevalently for those who possessed the relevant “mindware” to replace the default model with something more satisfactory. Without a more satisfactory model to replace the default model, the default model would be ratified by process 2 because of its initial salience (Johnson & Raab, 2003; Tversky & Kahneman, 1973; Hertwig, Herzog, Schooler, & Reimer, 2008). Thus, by controlling for performance on the screening question, we expected that any shift caused by the analytic intervention element would primarily manifest itself in the responses of those students who answered the screening question correctly. Thus, for Experiment 2B, we made the following two predictions:

> Prediction 1) There will be an improvement in performance for the treatment (AIE) condition compared to the control (non-AIE) condition on the two-box friction task.

Prediction 2) The improvement in performance caused by the AIE will occur more predominately among those who demonstrate relevant prior knowledge by answering the screening question correctly with correct reasoning.

3.7.4 Experiment 2B: Results and discussion

3.7.4.1 Performance results

Of those students who participated in Experiment 2B (N = 153), 52% arrived at a correct answer on the screening question supported by correct reasoning. In the control condition, 49% of students (N=81) correctly answered the target question. In the treatment condition, 64% of students (N=85) answered correctly. This improvement in performance of the treatment group with respect to the control group is not statistically distinguishable ($p = 0.13, V = 0.11$). Results controlling for the screening question are shown in Table 3-7.

	Screening Correct (with correct reasoning)		Screening Incorrect	
Task Version	Treatment (AIE) (N = 39)	Control (No AIE) (N = 40)	Treatment (AIE) (N = 39)	Control (No AIE) (N = 35)
$F_{frct\ on\ A} = F_{frct\ on\ B}$ (Correct)	90%	60%	41%	40%
$F_{frct\ on\ A} < F_{frct\ on\ B}$	8%	40%	54%	57%
$F_{frct\ on\ A} > F_{frct\ on\ B}$	2%	0%	5%	3%
Not enough info	0%	0%	0%	0%
	$\chi 2 = 9.24, p = 0.002, V = 0.34$		$\chi 2 = 0.22, p = 0.897, V = 0.04$	

Table 3-7. Performance data for the two-box friction task separated into treatment (AIE) and control (non AIE) groups while controlling for performance on the screening question.

3.7.4.2 Discussion of performance results

The lack of statistical difference and the small effect size observed in the overall performance improvement could have arisen from a statistical type-II error (*i.e.,* an outlier or false negative) or, alternatively, it could have stemmed from the presence of the screening question itself and its impact on student thinking; thus, at the present time, it is not possible for us to identify the source of the weaker signal in Experiment 2B compared to Experiment 2A. Given that the goal of Experiment 2B was to split both the treatment and control groups into sub-populations based on their performance on the screening question, the weaker signal is not necessarily problematic for the purposes of our intended analysis.

From Table 3-7, one can see that there was a statistically significant increase in performance with a medium-to-large effect size for the treatment group in comparison to the control group for students who answer the screening question correctly using the normative reasoning pathway; for students who did not answer the screening question correctly, no shift in performance was observed for the treatment group in comparison to the control group. Our operational definition of possession of relevant mindware was answering the screening question correctly with correct reasoning, so we see that our second prediction proved to be correct in that the performance increase was more predominate among those students who demonstrated that they possessed the relevant mindware, and that there was no improvement in performance among those students who did not demonstrate that they possess relevant mindware.

These results suggest that some students who had the requisite mindware available to them may have been prevented from applying that knowledge on the target question. We propose that they were prevented from applying that knowledge because of a strong feeling of rightness about an incorrect default model. When a challenge to that feeling of rightness is available to them in the form of the AIE, these students are then able to arrive at a correct answer using the appropriate mindware. Similarly, we propose that students who do not have the requisite mindware available to them are unaffected by a challenge to the feeling of rightness via the AIE because they do not have a more satisfactory alternative model to reason with.

In Experiment 2A, we argued that the answer choice “equal” could also be arrived at using solely perceptual (non-physics) cues once the coefficients are eliminated as a relevant factor, and that the analytic intervention element does not necessarily induce reflection on physics principles. Our results allow us to address this issue as well. If, after ruling out the common incorrect answer, the correct answer choice (“equal to”) was arrived at solely by perceptual cues and not with reference to relevant physics, we would expect the analytic intervention element to have been effective regardless of whether relevant knowledge was demonstrated on the screening question. Since the AIE had no impact on those who did not demonstrate relevant knowledge on the screening question, we are led to believe that the jettisoning of the default model is only useful when there is relevant conceptual knowledge at hand that can bolster confidence in the new model. Thus, students who were impacted by the AIE and subsequently answered correctly were likely considering physics principles and not simply answering according to perceptual cues based on task features.

3.7.4.3 Analysis of reasoning chains

Table 3-8 shows an analysis of reasoning chains while controlling for performance on the screening question. Note that Table 3-8 includes percentages based on the respective column. Each response was categorized based on the nature of the reasoning presented and is consistent with the categories described in Section 3.7.2.4. To summarize that discussion, the correct line of reasoning was typically given with either clear evidence of correct reasoning or else reasoning that was ambiguous as to whether Newton's 2^{nd} law was considered fully. (When it was ambiguous, it was regarded as possible that students were being cued directly on task features, which were equal for both boxes, and answering correctly without formal physics reasoning.) There were also a small amount of responses with no evidence of correct reasoning – these were either uninterpretable or contained only the answer element. The reasoning chains for those students who selected the common incorrect answer were marked either "canonical", wherein a student only endorsed elements which were directly related to the $f = \mu N$ model of friction, or "conceptual difficulty", wherein the student endorsed elements that indicated consideration of alternate, incorrect models of friction, or finally "struggle" reasoning, wherein the student incorporated reasoning consistent with the correct line of reasoning while ultimately selecting the common incorrect answer.

The nature of the reasoning chains in experiment 2B was similar to those for experiment 2A, namely, most correct answers were accompanied with correct reasoning and about 60% of students who chose the common incorrect answer (or, as shown in Table 3-8, about 30% of all students) employed reasoning that only referenced the single model based on the coefficients (and thus were categorized as “canonical”).

A striking difference between the reasoning chains in experiment 2A and experiment 2B is that in experiment 2B there is a greater number of students who appeared to struggle with a desire to reconcile correct knowledge and a strong default model inconsistent with that knowledge. (For an example of this type of response, see Section 3.7.2.4.) In Experiment 2A, these types of responses were not prevalent (less than 2% of responses), but in experiment 2B 30% of incorrect responses (or, as shown in Table 8, 14% of all students in the control condition) exhibited this “struggle” behavior. We surmise that asking the screening question primed these students to consider correct mindware while reasoning with an incorrect default model. Importantly, these “struggle” responses only occurred in the control condition, suggesting that similar students “struggling” to incorporate relevant conceptual information in the treatment condition were impacted by the AIE in such a way as to either push them towards a correct answer or to decide against including those considerations into their final reasoning chain. Table 3-8 provides support for this interpretation; while the percentage of students in the “struggle” category decreased from control to treatment regardless of performance on screening question, the only observed increase in performance was for those who answered the screening question correctly with correct reasoning. For those who did not answer the screening question correctly, it appears the effect of the AIE was to push them out of the “struggle” category and into other incorrect reasoning pathways.

Furthermore, from Table 3-8, one can see that in the control group, 38% of students who selected the correct answer on the screening question selected the common incorrect answer on the target question and cited either canonical reasoning or reasoning that suggests a struggle between the intuitive answer and the correct line of reasoning. With the inclusion of the AIE, however, the proportion of these responses appear to vanish while proportion of responses in the unambiguous correct line of reasoning category increases. On the basis of these results, we submit the following argument: students in the control condition who used correct reasoning on the screening question and responded to the target question incorrectly with chains that fall into the canonical incorrect category or the struggle incorrect category were blocked from using the requisite mindware by the cueing of an incorrect default model by process 1. Furthermore, we argue that if these students had access to the AIE in their reasoning elements, they would have overcome the feeling of rightness in this incorrect default model and responded with correct reasoning via a productive process 2 intervention.

	Screening: Yes AIE: Yes (N=38)	Screening: Yes AIE: No (N=42)	Screening: No AIE: Yes (N=41)	Screening: No AIE: No (N=35)
Correct w Correct Reasoning	84% (32)	48% (20)	22% (9)	17% (6)
Ambiguous Correct Reasoning	8% (3)	10% (4)	12% (5)	20% (7)
Correct w no evidence of correct reasoning	0% (0)	0% (0)	5% (2)	3% (1)
Canonical Incorrect Reasoning	3% (1)	24% (10)	39% (16)	37% (13)
Conceptual Difficulty Incorrect Reasoning	5% (2)	5% (2)	15% (6)	6% (2)
Struggle Reasoning	0% (0)	14% (6)	0% (0)	14% (5)
Other	0% (0)	0% (0)	7% (3)	3% (1)

Table 3-8. Comparison of reasoning chains in Experiment 2B controlling for performance on the screening question shown in Figure 3-2.a.

3.8 Conclusions and next steps

The overarching aim of this investigation was to study the extent that dual-process theories of reasoning could account for reasoning phenomena on qualitative physics questions using a new methodology, the reasoning chain construction task. In particular, we wished to draw upon dual-process theories of reasoning to make and test predictions about student behavior on these chaining tasks. From Evans' heuristic-analytic theory, we developed a working hypothesis that stated that students would be unlikely to shift away from an incorrect default model cued by process 1 unless they were provided with information that explicitly refuted the satisfactoriness of that model. Two sets of experiments built on the chaining task methodology were devised to test this hypothesis. In the first, students were given graph tasks with a known salient distracting feature (the intersection point, see Figure 3-1.b) which had been cast into a chaining format; the reasoning elements in the chaining task version of the graph task functioned to give students access to relevant conceptual information, thus testing whether or not this improved access would be sufficient to increase performance. In the second set of experiments, we gave students access to information (via the analytic intervention element, or AIE) that refuted a common incorrect default model about static friction in order to determine whether the presence of this information improved performance, as suggested by our working hypothesis.

Several important lessons came out of this work. Experiment 1A showed that providing increased access to relevant, correct information was not enough to produce a large shift in performance on a kinematics question with a salient distracting feature. Instead, that information was used by many students to justify an incorrect (and therefore inconsistent) answer. Experiment 1B showed that the salient distracting feature had a recognizable effect in three other content domains as well, and that, generally, the reasoning elements provided in each domain were not enough to negate the effects of the salient distracting feature on the reasoning process. Experiment

2A showed that a large increase in performance could in fact be realized by providing access to information (via the AIE) that refuted a common incorrect default model cued by the salient distracting feature on the static friction task. Experiment 2B revealed that the AIE had a greater impact on students who had previously demonstrated relevant mindware (*i.e.,* answered a screening question correctly with correct reasoning) and that there was no statistically discernible change in performance for those students who had not demonstrated relevant mindware. Together, these results provide support for the use of dual-process theories as a mechanistic framework for making and testing predictions about student performance and behavior, particularly about which models are selected and why, in turn, some are abandoned.

This work also has some broader implications related to the interplay between conceptual understanding and reasoning skills. This work strongly suggests that those students who possess the relevant mindware to answer a problem correctly may not use that mindware because of an undeveloped ability to critically reflect on an intuitive answer cued by process 1. It may also be possible that those who do have this domain-general reflective skill may answer a specific question incorrectly because they possess no relevant mindware in the specific context of that question (suggested by those students in the treatment condition (AIE) who answered the screening question incorrectly and answered the target question incorrectly as well). Alternatively, it may also be that students need a certain amount of mindware regarding a topic before being able to fully develop or employ the reflective reasoning skill. At any rate, it is clear from the current work that domain-general reasoning skills affect the process of content-specific reasoning, and that there is a need to develop both domain-general reasoning skills and conceptual understanding if increased performance is a goal. More work is needed to characterize with greater resolving power the interplay between reasoning skills and conceptual understanding in order to provide detailed research-based approaches for supporting reasoning skills and conceptual understanding in a more integrated fashion.

However, the successful leveraging of dual-process mechanisms in this work suggests a possible pathway to develop the skills needed to overcome an incorrect default model cued by a salient distracting feature. Giving the student access to a refutation of the default model apparently caused students to recognize and evaluate other relevant physics models. If this scaffolded prompting to search for other models could be repeated on many tasks with salient distracting features, students may begin to internalize a prompting to reflect on intuitive answers. This scaffolding could be provided directly by a line of questioning on a specific tutorial worksheet, but it may also be that more "hidden" scaffolding such as that provided by the AIE may be more effective in that, by interacting with the AIE, students are recognizing and modifying their answer without explicitly being prompted to do so. At some point, however, we suspect that students should be explicitly instructed about the impact of salient distracting features and how reflective thinking and searching for alternate answers can improve decision-making when these features are present, perhaps by having them reflect on their interaction with an AIE after the fact. We believe that instruction of this sort may aid students in developing the reflective skill necessary to effectively navigate qualitative physics questions with salient distracting features. More research, of course, is needed to gain insight into specific pedagogical approaches.

Finally, our work suggests that other domain-general reasoning effects can be studied through the lens of dual-process theories of reasoning, and that the mechanisms put forward by these theories can be used to make and test predictions about student performance and behavior. The results of such studies can then be leveraged to improve the teaching and learning of physics more broadly.

4 UTILIZING NETWORK ANALYSIS TO EXPLORE STUDENT QUALITATIVE INFERENTIAL REASONING CHAINS

4.1 Abstract:

Physics education research has produced instructional materials aimed at improving conceptual understanding, problem solving skills, and the skill of mathematizing real-world situations. Students are often expected to complete an introductory calculus-based physics course with these skills as well as a strong set of critical thinking skills related to qualitative inferential reasoning. Many of the research-based materials developed over the past 30 years are scaffolded and step students through a qualitative chain of inferences via a series of questions, and it is often tacitly assumed that such materials improve qualitative reasoning skills. There is, however, no real documentation of improvements in qualitative reasoning skills in the literature. Additionally, a growing body of research related to reasoning in physics highlights that general reasoning processes not tied to physics content may be responsible, in part, for the errors students make on some physics questions. New methodologies are needed to better study reasoning processes and to disentangle, to the extent possible, processes related to physics content from processes general to all human reasoning.

In our investigation, we employed network analysis methodologies to examine student data from reasoning chain construction tasks in order to gain deeper insight into the nature of student reasoning in physics. In a

reasoning chain construction task, or simply chaining task, students are given a list of reasoning elements (such as statements of physics concepts) and are asked to assemble a chain of reasoning from the elements leading to an answer. In this paper, we show that network analysis metrics are both interpretable and valuable when applied to student reasoning data generated from reasoning chain construction tasks and illustrate how network analysis is useful for both studying known inferential reasoning phenomena and for uncovering new phenomena for further investigation.

4.2 Introduction

Students pursuing undergraduate STEM majors are often expected to take one or more physics courses as part of their degree programs, even when they are not physics majors. While certain physics concepts and principles will be of use in these students' future academic careers, many will not. Instead, it is often expected that the lasting takeaways from a physics course will be a repertoire of problem-solving strategies, a familiarity with mathematizing real-world situations, and a strong set of critical thinking skills related to qualitative inferential reasoning. Furthermore, these takeaways are important to all students taking a physics course, including those who go on to be physics majors and physicists.

Physics education research has produced many instructional materials that have been demonstrated to improve conceptual understanding and other learning outcomes (Finkelstein & Pollock, 2005; Saul & Redish, 1997;

Sokoloff & Thornton, 1997; Beichner R. , 2007; Crouch & Mazur, 2001). Many of these materials are scaffolded and step students through qualitative chains of inferences via a series of questions (McDermott & Shaffer, 2001; Lillian C. McDermott, 1995; Wittmann, Steinberg, & Redish, 2004). It is often tacitly assumed that such materials also improve qualitative reasoning skills, but there is no documentation of such improvements in the PER literature. Furthermore, it has been observed that despite overall conceptual gains after research-based instruction, there are still certain physics questions for which it is difficult to improve student performance (Heckler, 2011; Kryjevskaia, Stetzer, & Grosz, 2014; Heron, 2017). Instead, these studies suggest that reasoning processes general to all humans may impact how students understand and reason in a physics context.

There is thus a need to investigate how students generate qualitative inferential chains of reasoning. To do so, new methodologies need to be explored, particularly those that can separate, to the degree possible, reasoning skills from conceptual understanding. Some methodologies have approached this goal. For instance, eye tracking methodologies seek to determine where attention is being placed while working through a physics problem and can be used to gain insight into domain-general reasoning processes that apply in many different contexts (Rosiek & Sajka, 2016; Heron, 2017; Sattizahn et. al., 2015). Additionally, methodologies that seek to find and document a particular reasoning-related phenomenon across multiple different contexts also separate, to a degree, reasoning patterns from particular physics concepts (*e.g.,* those methodologies employed in Heckler & Bogdan, 2018 and Heckler & Scaife, 2014). But these methodologies don't necessarily separate students' knowledge of a concept required for a particular problem from their ability to reason through that problem; rather, the methodologies are examining reasoning phenomena that occur outside of a given physics context.

A methodology that comes close to the goal of separating reasoning skills (in particular, the skill of productively navigating an intuitive response when it is in conflict with the correct response) from conceptual understanding on a given problem is the paired question methodology reported in (Kryjevskaia, Stetzer, & Grosz, 2014). This methodology has provided further evidence that many students possess an ability to reason correctly through a physics problem but opt for other, more salient lines of reasoning on closely related questions.

In connection with a similar project, we have developed a new methodology centered around *reasoning chain construction tasks,* or *chaining tasks*, that have been designed to separate reasoning skills from understanding of a particular physics concept. This methodology was initially reported in Speirs, Ferm Jr., Stetzer, & Lindsey (2016) and has since been used to leverage results from cognitive science to improve student performance on qualitative physics questions. In this companion paper, we describe a method for exploring chaining task data using network analysis and present four examples that demonstrate the utility of network analysis methods for gaining insight into the structure of student reasoning via chaining tasks. The overarching goal of this manuscript is to highlight the affordances of network analysis approaches to generate knowledge about how students reasoning on physics questions, particularly when they are responding to questions requiring a series of inferences. In combination with reasoning chain construction tasks, network analysis generates novel data and findings related to the content and structure of student arguments. These data and findings will support further research exploring the mechanisms behind student reasoning in physics and the development of reasoning skills over time. Indeed, the groundwork for such research is laid out in the final discussion section.

4.3 Background

In this section, we review pertinent literature that both makes the case for the need for more sophisticated analyses of student reasoning and highlights the unique affordances of network analysis of chaining task data to meet this need.

4.3.1 Research directly related to qualitative inferential reasoning in physics education

Understanding student reasoning on physics problems has long been a goal of physics education research. Early investigations of student conceptual understanding identified specific reasoning difficulties as well as conceptual difficulties. This long tradition of more than 30 years unearthed similar reasoning difficulties in many different places. One such difficulty could be referred to as *compensation reasoning*, in which two physical quantities that change in opposite ways were assumed to cancel (Lawson & McDermott, 1987; Loverude, Kautz, & Heron, 2003; Kautz, Heron, Shaffer, & McDermott, 2005; Lindsey, Heron, & Shaffer, 2009). The focus of these early investigations was to identify the prevalence of such difficulties and to address them in a non-general, content-specific way. In this research tradition, no claims were made as to the cognitive structure or composition of the difficulties; rather, the difficulties were described as observed and the empirical findings were used to guide the development of content-specific, research-based instructional materials (McDermott, 2001; McDermott, 1991; Heron, 2004).

Other early investigations sought to understand the composition of student conceptions of physics and to explain how or why certain conceptions were formed, cued, and used for reasoning (diSessa, 1993; diSessa & Sherin, 1998; Hammer, 1996; Redish E. F., 2004; Elby, 2000; Hammer, Elby, Scherr, & Redish, 2005). These investigations created a framework that allows one to identify and observe the use of student "resources" for reasoning. "Resource" is a general term for fine-grain cognitive structures (*i.e.,* general rules, epistemological stances, phenomenological primitives) that make up larger-grain cognitive structures such as concepts or skills. It is posited by this framework that the act of reasoning is an act of cognitively selecting and coordinating the use of a subset of available resources. While the resources framework is useful, it falls short of making specific predications about which resources are activated when and how they impact reasoning. Instead, the framework provides compelling post-hoc explanations for reasoning phenomena.

A growing body of research is investigating predictive control mechanisms that govern reasoning in a physics context. For example, in order for a task feature to cue a specific resource in the course of reasoning, that feature must be processed by the brain. Thus, the time it takes to process a certain feature represents a control mechanism that may predict which resources are cued and when. To show the impact of processing time on answering patterns, Heckler and Scaife (2014) measured the approximate processing time of finding either the slope or the height of a particular point on a graph and found that processing the slope took a longer time than processing the height. Applying an enforced time-delay on student answers guaranteed that the students' brains had time to process the slope and resulted in improved performance on questions in which the slope and the height of a particular point were in competition (*i.e.,* that the two quantities led to different answers).

This strand of research has called for new methodologies to be employed in physics education research that would allow for the collection of data not normally accessible from a written response or think-aloud interview alone (Heckler, 2011; Sattizahn et al., 2015). Methodologies that can separate reasoning skills from conceptual understanding are particularly useful. One methodology that represents a step in this direction is a paired question methodology reported in (Kryjevskaia, Stetzer, & Grosz, 2014; Kryjevskaia, Stetzer, & Le, 2015). This methodology aims to gain insight into the impact of intuitive responses on the formation of reasoning chains. This is accomplished by first asking a "screening question" that requires a student to step through a specific line of reasoning and then immediately asking a "target question" that requires that same line of reasoning. The target question is similar to the screening question but is typically designed or selected to elicit an intuitive, incorrect response. This methodology was used to examine "compensation reasoning" in the context of capacitors and demonstrated that even those students who articulated the correct line of reasoning on the screening question abandoned that reasoning in favor of the intuitive incorrect reasoning on the target question. To provide further evidence that students did in fact possess the ability to correctly reason through the problem, the target question was administered in two formats. In one, the student was given the question and asked to answer it. In the other, the student was given the question along with the answer and asked to justify that answer. Those students in the "justify" condition who answered the screening question with correct reasoning gave the correct justification, while some among those in the "answer" condition who answered the screening question correctly still employed the compensation argument.

4.3.2 Other discipline-specific, reasoning-related research

The list of reasoning-related research can be rightfully extended to the expansive research on student problem solving (Hsu, Brewe, Foster, & Harper, 2004). Research on student problem solving emphasizes traditional quantitative problems that typically require manipulation of multiple equations and quantities and seeks to understand and improve the strategies students employ while working through these problems. It has been pointed out that the list of skills and strategies that a student has to employ while problem solving is extensive and somewhat overwhelming. Notably, rubrics for assessing problem solving skills continue to be developed (Docktor, et al., 2016). Likewise, there has been research related to scientific reasoning skills such as control of variables, conservation of volume, and proportional reasoning, and assessments have been used to study differences in proficiency with these skills between populations before and after instruction (Lawson, 1978; Coletta et al., 2009; Bao, et al., 2009; Ding, 2014).

However, while quantitative problems and scientific reasoning are essential to a physics curriculum, the focus of this manuscript is on the structure of qualitative inferential reasoning patterns more akin to the reasoning difficulties identified in specific content areas of physics. Additionally, many of the research-based instructional materials expect students to engage in qualitative inferential reasoning in order to deepen conceptual understanding (*e.g.,* McDermott & Shaffer, 2001; Wittmann, Steinberg, & Redish, 2004). Instructors often have this expectation as well.

The *proofs literature* in mathematics education research is somewhat more closely aligned to the specific goals of the investigation described in this manuscript. Selden and Selden provide a wonderful review of this literature in a 2008 paper (Selden & Selden, 2008). In a typical undergraduate mathematics program, there are specific courses that aim to teach student how to create mathematical proofs. These proofs tend to take the form of a series of deductive, qualitative inferences that are linked together as an argument in support of a specific conclusion. The research regarding student skill at constructing proofs is reminiscent of many research endeavors in physics education. Often, students' responses to a particular proof task are examined through various epistemological and conceptual lenses, with an emphasis placed on the identification of student difficulties with constructing proofs. While the nature of the reasoning chains examined in the "proofs" literature is very closely related to those considered in this manuscript, the current work takes a different approach. Instead of examining possible causes for a particular reasoning difficulty, the current work aims to identify patterns in the structure of the reasoning chain itself; our goal is to provide new forms of data that can be utilized by future researchers investigating the mechanisms behind student construction of reasoning chains.

4.3.3 Network Analysis in Physics Education Research

Network analysis is fairly new to physics education research but has recently been seeing a dramatic increase in use, mostly in social network analysis characterizing social dynamics within a physics community (*i.e.,* a classroom, department, or university) and relating these dynamics to performance and learning gains within a physics course (Spillane & Kim, 2012; Brewe, Kramer, & Sawtelle, 2012; Bruun & Brewe, 2013; Wolf, Sault, & Close, 2018; Vargas, *et al.*, 2018). However, network analysis has also been used to study epistemological shifts in conversations as a result of instruction (Bodin, 2012), to model differentiation of concepts (Koponen, 2013), to assess patterns in representation use throughout a course employing modeling instruction (McPadden, 2018), and to gain insight the structure of answer patterns on a conceptual inventory (Brewe, Bruun, & Bearden, 2016). The current work utilizes network analysis to study the structure of student reasoning chains, which we believe is a novel pursuit.

4.3.4 Resource Graphs as Network Analysis

Returning to resources, the coordination of resources has been studied using network-like representations, sometimes called "resource graphs" (Wittmann, 2006; Sabella & Redish, 2007; Smith & Wittmann, 2008; Black & Wittmann, 2009). Resource graphs offer a view of the theoretical constructs within the resources framework by highlighting the structural topology of these constructs. One of these views is that some concepts share a similar sub-set of resources, with only one or two resources making the difference between a productive, correct conception for the context and an unproductive conception (Smith & Wittmann, 2008), and evidence has been presented for the reification of particular procedural resources from smaller grained resources (Black & Wittmann, 2009; Wittmann & Black, 2015). Another insight put forward in these studies is that conceptual change can be represented as the rearrangement or addition/deletion of connections among specific resources. Finally, Sabella and Redish (2007) modeled the flow of a student's inferential reasoning using a network-like representation called a "reasoning map". In that paper, they modeled a student's knowledge structure as brief statements of the student’s reasoning and showed that there were differences in students' knowledge structures based on the reasoning maps constructed from their think-aloud reasoning.

While resource graphs could, in principle, offer a more detailed view of student reasoning, the match between a resource graph and experimental data is challenging due to some level of ambiguity in terms of what constitutes a resource when coding experimental data. In addition, another challenge appears to be ascertaining what exactly counts as a connection between resources. For instance, a resource could be represented as a collection of smaller-grained constructs or as a reified object. Which is it to the particular student? Differentiating between the two can be hard from think-aloud data alone, unless the student is particularly loquacious. The current work side-steps this issue by providing a pre-defined statement of knowledge to the student and seeks to investigate the structures that emerge from student use of these pre-defined statements. Thus, network analysis of chaining task data may provide a methodology through which the theoretical constructs inherent in resource graphs can be studied in a systematic way.

4.3.5 Summary

The data collection and analysis methodology presented in this manuscript is designed to create a separation between reasoning skills and conceptual understanding and to provide data not normally accessible from written responses and think-aloud interviews. We aim to create a tool that can be used to study specific reasoning difficulties, to provide insight into the development of specific reasoning abilities, and to serve as a venue in which to test predictions made by mechanistic theories from cognitive science. The main goal of this paper is to demonstrate how network analysis of reasoning chain construction tasks may be used in order to accomplish all three objectives.

4.4 Methodology

This section is broken into two main parts. In the first, we describe the reasoning chain construction task, which underlies the methodology employed here. In the second, we describe the network analysis methods that are of use in this manuscript.

4.4.1 Reasoning Chain Construction Tasks

A reasoning chain construction task, or *chaining task*, is a modified card-sorting task in which we: (1) provide the student with a list of reasoning elements; (2) indicate that all of the statements within these elements are true and correct; and (3) ask the student to construct a solution to a physics problem by selecting elements from the list, ordering them, and, as needed, incorporating provided connecting words (“and", “so", “because", “but"). The reasoning elements primarily consist of observations about the problem setup, statements of physical principles, and qualitative comparisons of quantities relevant to the problem; all of which are true. Everything the student needs to produce a complete chain of reasoning is present in the elements; the student's task is then to pick from the given conceptual pieces and directly assemble a reasoning chain.

Reasoning chain construction tasks have primarily been implemented online using Qualtrics' "Pick/Group/Rank" question format. This online format is illustrated in the context of a graph task and is shown in Figure 4-1. Reasoning elements from the "Items" column, connecting words, and final conclusions can be dragged and dropped into the "Reasoning Space" box; the box increases in size vertically as elements are added.

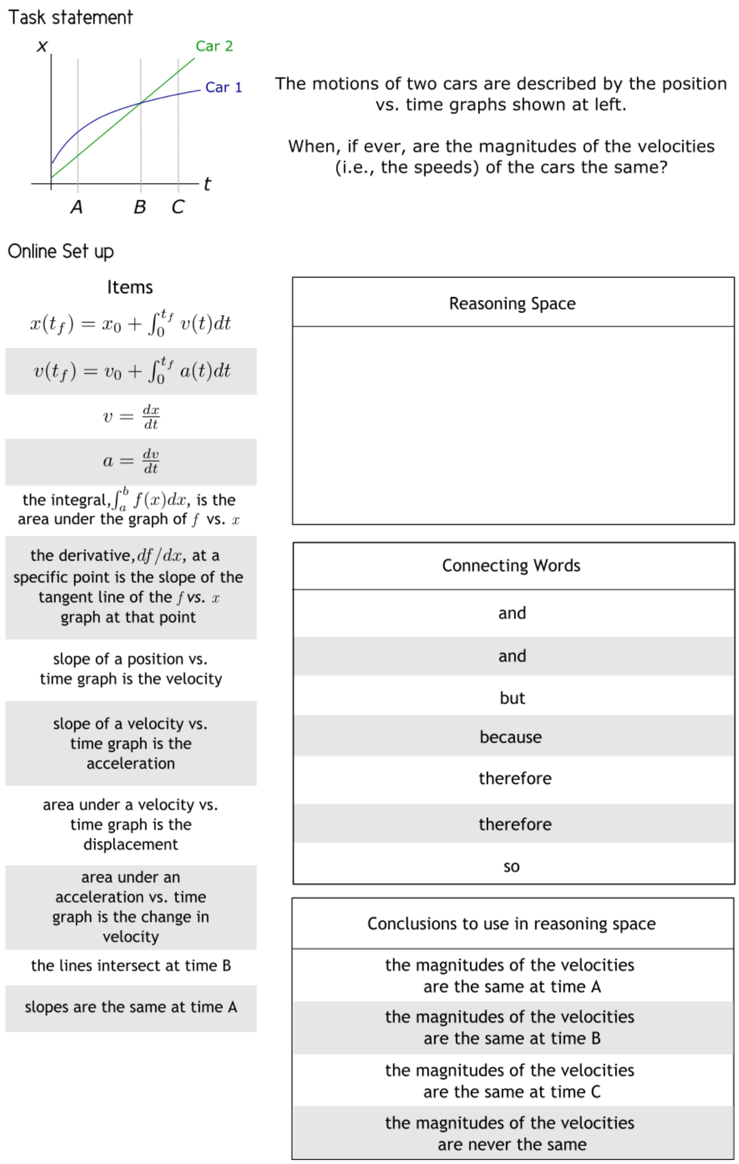


Figure 4-1. An example of a reasoning chain construction task implemented online using Qualtrics' "Pick/Group/Rank" question format.

These tasks were administered on homework assignments or exam reviews for students enrolled in an introductory calculus-based physics sequence, along with other questions relevant to the course but not relevant to the content found in the research task. These assignments counted for participation credit in most cases, although extra credit was awarded in some cases. In all cases, the tasks were administered after relevant lecture, laboratory, and small-group recitation instruction at a research-intensive university in New England. Research-based materials from *Tutorials in Introductory Physics* (McDermott & Shaffer, 2001) were used in the course recitations.

The reasoning elements provided to the student were typically based on previously obtained student responses to open-ended, free-response versions of the task. Elements consisted of statements of first principles, observations about the task, and statements derived from first principles and observations. Some were productive to the correct line of reasoning, and some were not. Among the unproductive elements were elements that, while true, were useful primarily in constructing a common incorrect line of reasoning, if there was one associated with the task. In addition, the extent to which students selected unproductive elements not associated with the correct or common incorrect line of reasoning could help us gauge the likelihood that students were simply inserting elements at random. Three blank elements labeled “Custom:” were provided, with instructions that students could use the text box attached to the custom element to create their own reasoning elements is they felt they wanted to add something not represented among the given reasoning elements.

An important aspect of a chaining task is the intended logical connections between the provided reasoning elements – that is, the logical topology of the elements. For instance, some physics tasks require only a few steps to arrive at a correct answer (*e.g.,* a qualitative question that can be solved via a short, linear chain of elements like the task shown in Figure 4-1), while others require the student to combine two independent lines of reasoning (*e.g.,* synthesis problems such as those reported by (Ibrahim, Ding, Heckler, White, & Badeau, 2017)); by casting each of these types of questions as a chaining task, we can obtain information about how students approach these different scenarios. In particular, by manipulating the logical topology of the task, we can introduce experimental conditions that can provide deeper insight into student ability to generate inferential chains.

When considering what can be learned from student responses to a chaining task, there are a few important points to remember. The first is that the provided reasoning elements determine to a large extent how students interact with the task. The elements were written by researchers (*i.e.,* the author of this work) who likely have a specific epistemological stance in mind, as well as a particular pedagogical perspective. The elements and especially the wording of the elements reflect the researchers' values about such ideas as what constitutes reasoning, a reasoning element, and the size of logical steps. For instance, an element corresponding to Newton's second law could read, among other things, "$F_{net} = ma$", "the net force is equivalent to the mass times the acceleration", or "an acceleration is caused by a net force." Each of these may convey a different meaning to the student, may interact differently with the context of the problem, and may differently represent what a "first principle" is and looks like. Thus, when interpreting responses to a chaining task, the main research endeavor is to ascertain not how students' reason generally about the problem, but how students engage in the specific types and lines of reasoning supported by the elements. In most of the tasks presented in this manuscript, attempts were made to make the reasoning space topology as close to the observed student reasoning topology by drawing upon student written explanations of reasoning, but there were some intentional exceptions (which will be discussed later).

A second point worth mentioning is that the chaining task (especially when implemented online) creates an environment in which students are required to present their argument in a linear progression of inferences, and this presentation of reasoning is separate from the process of reasoning that occurs in the mind. For instance, a student may consider a lengthy line of reasoning, but feel that simplicity and elegance are valued in the sciences and therefore seek to construct the most concise argument possible in the elements; another student, though, may report a short chain out of a desire to get through the task quickly, without deep study of the elements provided. Regardless of these differences, there is still something valuable to be gained from analyzing patterns in the reasoning chains constructed by students. For example, suppose students don't endorse first principles in their chains. We can't assume that they did not consider first principles, but we can assume that if they did consider first principles, they made a decision (whether conscious or not) to exclude those considerations in the presentation of their reasoning.

4.4.1.1 Chaining task data as networks of associations

Chaining task data can be cast as a network for quantitative analysis. To accomplish this, the reasoning elements can be represented as nodes in a network and associations made by the student between the elements can be represented as links. We considered two main methods for establishing associations (links) between reasoning elements (nodes). In the first, a connection is said to exist between two elements if the two elements are placed consecutively in a student's chain or on either side of a connecting word; a network created using this definition of association is referred to in this paper as a *direct association* network. In the second method, a connection exists between two elements if they appear together in the same student response; a network constructed in this way is referred to as an *indirect association* network. Individual student response networks are summed to create the full network for all responses in a given data set.

In both methods, we remove connecting words from the data and use undirected links to form our networks. The connecting words, while serving in many cases to clarify the logic of a student's argument, posed a challenge for network analysis for two reasons. Initially, it was hoped that the connecting words could be used to define different types of links between elements (some causal, some associative). This hope was diminished when it was observed that students often used connecting words intermittently and inconsistently. For instance, a few students placed an answer element followed by "therefore" and then elements that justified their answer, effectively reversing the inherent logic between the answer and the argument. This may have been a simple oversight or error in meaning (like a typo) or it may have reflected a deeper misunderstanding of logical connectives. At any rate, it was unclear in some cases that the connecting words were being used according to a normative understanding of logic. The second difficulty was that even when connecting words were used consistent with normal rules of logic, there remained ambiguity in the components that were intended to be associated with the connective, particularly when a task required multiple inferences. For instance, consider the phrase "A because B and C therefore D". This phrase could be parsed logically as "A because (B and C)" or it could be parsed as "(A because B) and C". (Similar ambiguity exists regarding the parsing of the "therefore" connective.) For these two reasons, we felt uncomfortable attributing representational meaning to the connecting words when constructing the networks.

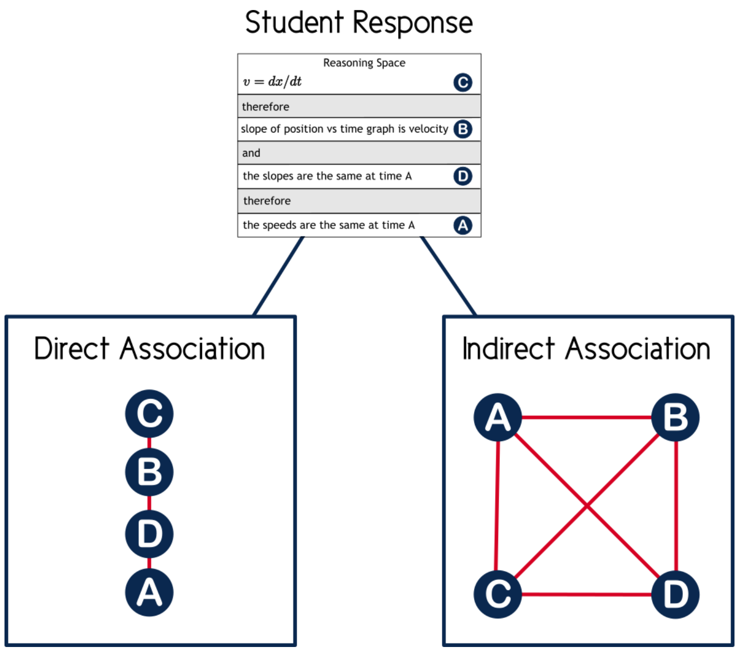


Figure 4-2. An example of two methods for constructing an individual-student network from an individual student's response.

Because we removed the connecting words from students' responses when constructing a network, we also opted to make the links undirected. One could imagine, alternatively, a scheme that encodes either (1) the ordering of the elements by placing a directed link (*i.e.,* an arrow) from an element to the element that comes next in the chain, or (2) the logical associations implied by the connecting words, using undirected links for elements connected by "and" as well as directed links for elements connected by "therefore" or "because". A network constructed according to the latter scheme would be problematic for the reasons outlined in the previous paragraph. However, a network constructed using the former scheme would also be problematic because a directed link would imply a causal direction in the association between elements. This implication would be misleading because the directionality of the association is made ambiguous when removing the connecting words. For instance, the phrase “A therefore B” could equivalently be written “B because A”. When constructing a directed network, both cases would be represented differently in the network but actually correspond to the same type of logical causality. We wished to respect this limitation by not representing the ordering of the elements in students’ responses, instead opting to represent the proximity. By doing so, we interpret a link between reasoning elements as simply a general “association” between those elements rather than interpreting any sort of logical meaning from the link. However, we find that this method of constructing networks does yield interpretable results, and we view this decision as a ground-level analysis of reasoning chains. Future analyses may be performed in order to investigate the usefulness of directed networks.

In some cases, directed networks were constructed to better interpret the undirected networks, primarily in measuring which elements were likely to be first or last in a chain. We measured this by creating a directed network according to scheme 1 explained above, in which there is a directed link from an element in a chain to the subsequent element used in that same chain. Using this network, we calculate the ratio of out-degree (number of links pointing away from the node) to in-degree (number of links pointing toward the node). Elements for which this ratio is much greater than one are considered to be likely starting points while elements with a ratio less than one are considered to be ending points. It has been observed that, in most chaining tasks, the answer elements tend to be ending points.

Note that in the work presented in this dissertation, undirected indirect- and direct-association networks are both used in the main analysis, whereas directed direct association networks are only used in certain places where useful.

4.4.2 Network analysis

In this section, we present an overview of the network analysis techniques employed in this work. Later sections will describe in detail how to interpret the results of these methods in the context of reasoning chain construction tasks.

4.4.2.1 Locally Adaptive Network Sparsification

Network sparsification aims to uncover the "backbone" structure of a large network by deleting links (sometimes called edges) that are unimportant to that structure (Foti, Hughes, & Rockmore, 2011). One simple method for achieving this is to establish a threshold value for a link's weight and delete all links that fall below this threshold. For instance, one might decide a connection is only relevant if more than 5% of students made the connection, and so we would delete any link that had a weight less than the value of $0.05 * N$, where N represents the population size. However, this method does not preserve some structures that may be of interest. Perhaps a small group of students decided to be detailed in their reasoning chains, and so they added structure to the network that is relevant to overall patterns of reasoning but, due to their low prevalence among the whole population, this structure might get cut from the network by an arbitrarily set threshold weight. Additionally, it may be hard to guess, *a priori*, a threshold weight that preserves these structures and still reduces the complexity of the network.

Another, more sophisticated, method of sparsification is Locally Adaptive Network Sparsification (LANS) (Foti, Hughes, & Rockmore, 2011). In LANS, the statistical significance of each link is calculated for the two nodes locally and a link is deleted only when it is found to be below a threshold value of significance to both nodes. This preserves local structure that would be dismantled using a threshold link weight method. The LANS method is implemented by first calculating the fractional link weight of a link connecting nodes i and j, as

$$p_{ij} = \frac{w_{ij}}{\sum_{k=1}^{N_i} w_{ik}},$$

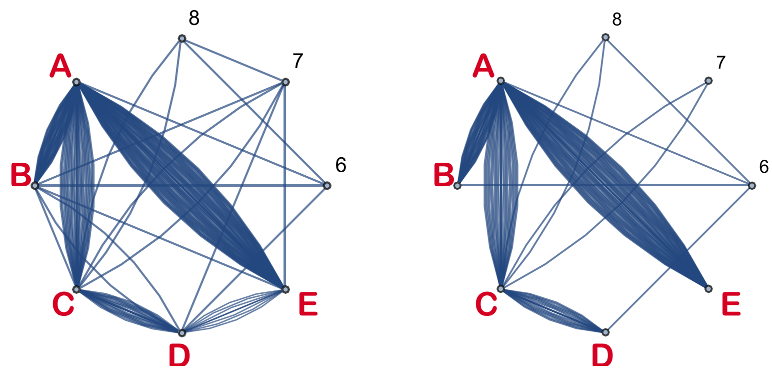
where w_{ij} is the weight of the link, and the sum in the denominator is over all the nearest neighbors of the node i. Then, the cumulative distribution function (CDF) is computed as

$$F_{ij} = \frac{1}{N_i}\sum_{k=1}^{N_i} \hat{1}\{p_{ij} < p_{ik}\},$$

and the link is retained if $F_{ij} > \alpha$, where α is the pre-determined significance threshold. These same calculations are, of course, completed for every link in the network.

To give an example of how this method works, a sample network (Figure 4-3.a) was constructed, and the technique applied. The main structure of the original network is represented by the lettered nodes. The link between nodes D and E is 7 times weaker than the link between nodes D and C; all other links between lettered nodes are roughly equivalent in strength. The added nodes 6-8 were given random connections to each other and the other nodes in the network to simulate smaller structures that may be of interest and generate "noise". The sparsified network is shown in Figure 4-3.b. One can see that the smaller structures have been retained even after the network has been simplified via the LANS technique. Importantly, the connection between nodes D and E has been severed. Thus, this technique is able to preserve small structures while still detecting and removing weaker connections among the larger structures.

Note that the four connections to node 6 remain. This is because those four connections are equally significant to node 6; more generally, anytime a node has only edges of weight one, all of those links will be preserved due to the nature of the algorithm. Because of the tendency to automatically preserve nodes such as node 6, we "prune" sparsified networks by removing all links of weight 1 *after* sparsification to make the network more readable.



a b

Figure 4-3. Example network illustrating Locally Adaptive Network Sparsification (Foti, Hughes, & Rockmore, 2011). (a) The base network. (b) The same network after sparsification at $\alpha = 0.1$.

For the work presented here, the threshold α was chosen by lowering the theshold as much as possible before either nodes or collections of nodes began to be separated from the network. For instance, in some networks, there are elements that are more tightly associated with each other than with the rest of the network, and these may break off during sparsification when the threshold is too low. We wished to preserve the structure of the network to the extent possible while still simplifying it, so we felt uncomfortable breaking the network into separate pieces. Typical values of α for this work ranged from 0.1 to 0.2. These values ended up being consistent with other studies using LANS (Foti, Hughes, & Rockmore, 2011).

4.4.2.2 Community Detection

The techniques of network analysis allow us to quantitatively determine groupings of elements, or *communities*, which are more tightly associated with each other than with the rest of the network. There are many methods of community detection available, and there is no single “best” method (Fortunato, 2010). The method used in this work is termed optimum modularity community detection (Newman, 2006). This method of community detection was chosen based on its potential for interpretability of results and also because the underlying statistical nature of the method allowed it to be useful for a broad range of network types. It was also selected because the method allowed for a rigorous definition of a community as an indivisible sub-graph of the network.

Network modularity is proportional to the number of links between a pre-defined group of elements minus the number of expected links in an equivalent network (*i.e.,* one with the same nodes) in which the links are placed at random. The expected number of links is $k_i k_j / 2m$, where k_i and k_j are the degrees of node i and node j, and m is the total number of links in the network and is given by $m = \frac{1}{2}\sum_i k_i$. Thus, the expected number of links is related to the degree of the node: the higher the degree, the more likely it is to have links in a network in which the links are placed at random.

The modularity is maximized by dividing the network into two subgraphs of maximum modularity and then repeating this process for each of the two parts. If any proposed division causes the total modularity to decrease, the corresponding subgraph is preserved and considered a community, and the algorithm moves on to the next subgraph until all communities are found. Thus, a community is defined as an indivisible subgraph of the network.

Before relying on the results of community detection, it is helpful to gauge how robust the community structure is. Could small perturbations produce a different community structure in the network? If the answer is yes, then it would be reasonable to mistrust the divisions made by optimizing modularity. However, if the structures are impervious to random insertions or deletions, this would be clearer evidence of true community structure. To assess robustness, we employ a technique based on statistical bootstrapping that has been modified from Fortunato (2010) for the context of chaining tasks.

For a data set of N student responses, our bootstrapping technique consists of creating a hypothetical data set comprised of $M = N$ responses drawn at random from the N actual student responses. (A specific response in the original data set may be selected more than once for the hypothetical data set; if this weren't the case, the hypothetical data set would be equivalent to the actual data set.) This hypothetical data set is treated as a new data set and a network is constructed from it. The community structure of this new hypothetical network is found, and tests are applied to the hypothetical community structure. The process is then repeated for many iterations, tallying the results of the tests so as to determine how frequent a particular result is. It is suggested to perform as many iterations as possible, but in chaining task analysis, convergence is attained quite easily. Accordingly, in the research described in this manuscript, a standard 1000 iterations were found to be sufficient to obtain reliable information.

Typically, the primary test for a bootstrap iteration is to determine whether or not the community structure in the hypothetical network is the same as the community structure in the actual network. In many cases, one or two elements may not be as tightly bound in a community as the others, and so testing for the exact community structure does not produce enough resolution to determine the strength of a community. Instead, it is helpful to determine, via testing, which elements are most often contained in a given community. This type of test can be applied by selecting an element of interest (such as an answer element) and determining which of the other elements are consistently in the same community as that element. By taking note of the community members in each iteration, a frequency plot can be generated from the results. An example of such a frequency plot is shown in Figure 4-7.

4.4.2.3 Network measures: centrality and clustering

Two network measures, *betweenness centrality and global clustering coefficients*, were utilized in the current work and will be described here. *Betweenness centrality* (Opsahl, Agneessens, & Skvoretz, 2010) is seen as a measure of a node's control over the "flow" in the network. A node's betweenness was originally defined as the number of shortest distance paths through that node divided by the total number of shortest distance paths in the network (Opsahl, Agneessens, & Skvoretz, 2010). This definition applied only to unweighted networks, and so the definition was modified to respect the weights of the various links in the network by defining "shortest distance" as a combination of the traditional "distance" (*i.e.,* number of nodes on a path between two end-nodes) and a "conductance" (*i.e.,* the weighting of the different links on a path between two end-nodes) (Newman, 2001). Opsahl *et. al.* (2010)'s modification of betweenness for weighted networks relies on a similar definition of shortest distance, and is represented as

$$d(i,j) = \min\left(\frac{1}{(w_{ih})^{\alpha}} + \cdots + \frac{1}{(w_{hj})^{\alpha}}\right)$$

where d is the shortest distance between node i and node j, w_{gh} is the weight of the link between nodes g and h, and α is a positive tuning parameter which is set based aspects on the context that the network is representing. When $\alpha < 1$, the number of nodes in a path becomes a greater influence on the distance, whereas for $\alpha > 1$, the weight of the links becomes a greater influence. In chaining networks, the weight of a link represents the number of students who made an association between the two elements and so it should have the most influence over the distance: a path that many students established should be of smaller distance than a short path that only a few students took. However, we don't wish to completely drown out structures created by only a few students. For this reason, we select a value of 1.5 for α. The betweenness is then calculated in the same manner as for unweighted graphs: by finding the ratio of the number of shortest paths through a given node to the number of shortest paths in the network.

Global clustering coefficients were also defined originally for unweighted networks and needed to be updated for weighted networks. The goal of a global clustering coefficient is to quantify how interconnected a network is. The clustering coefficient was originally defined as the number of closed triads (grouping of three nodes all connected to each other) divided by the total number of triads, either open, (*i.e.,* only two links among the three nodes), or closed (*i.e.,* all nodes connected) (Opsahl & Panzarasa, 2009). The direct association network shown in Figure 4-2 would have a clustering coefficient of zero, while the indirect association network shown in that figure would have a clustering coefficient of one. The idea of clustering is extended to weighted networks by assigning a weight, ω, to each triad in the network based on the weights of the links in the triad (Opsahl & Panzarasa, 2009). The weights, ω, are computed from the geometric mean of the weights of the two links stemming from the center node of the triad. The clustering coefficient can then be defined as follows, with τ representing the set of triplets and τ_Δ representing the set of closed triplets:

$$C_\omega = \frac{\text{total value of closed triplets}}{\text{total value of triplets}} = \frac{\sum_{\tau_\Delta} \omega}{\sum_\tau \omega}.$$

Thus, if a network had many closed triads compared to open triads, but the open triads were all of heavier weight, the network may not be considered to be interconnected. Conversely, if a network had few closed triads but these triads weighted most heavily in the network, this network would rightly be considered to be interconnected.

4.5 Research tasks

In this section, we present network analysis of four chaining tasks in a physics context in order to highlight the power of these methods in providing insight into student reasoning. The first task is set in a work and energy context and provides an introduction to the interpretations of the network analysis methods in the context of chaining tasks. The second and third tasks examine reasoning related to friction and reveal the possible utility of network analysis of chaining tasks toward understanding the structure of student knowledge. Finally, in the last section, we detail a set of four isomorphic graph-based tasks that span four content areas: kinematics, potential energy, electric potential, and magnetic flux. Network analysis of these graph-based tasks reveals the development of a more coherent line of reasoning across two semesters of introductory physics instruction.

In summary, this investigation asked and answered the following research questions. To what extent can network analysis methodologies applied to reasoning chain construction task data better characterize the nature of student reasoning on qualitative physics questions? In particular, how can we interpret the results from network sparsification, community detection, and betweenness centralities when applied to networks of reasoning chain elements?

4.5.1 Work-Energy task

Here we focus on a chaining task in the context of work and energy, and we use this task as an example of how the methods of network analysis can be interpreted in the context of chaining tasks. In this section, we describe the task, provide the results of the network analysis techniques described in section 4.4.2, and discuss the insights gained from this approach.

The goal of this task was to answer the following question. How effective are network analysis methodologies at characterizing and differentiating among different lines of reasoning on a physics question that most students can answer correctly?

4.5.1.1 Physics question overview

The work-energy task was adapted from a concept question (Chapter 9, Concept Question 6) appearing in Knight's text (Knight, 2016). In the task, students are told that a point particle moving to the left is slowing down because of a force pushing to the right, and no other forces are acting on the particle. Students are asked if the work done on the particle by the force is positive or negative, or if there is not enough information to tell. The complete prompt as well and the reasoning elements provided to the student are shown in Figure 4-4.

The correct answer is that the work on the particle by the force is negative. There are two viable ways of answering this question. The first involves recognizing that the work done is defined as the dot product between the force and displacement vectors and that a dot product of two vectors pointing in opposite directions is negative in order to establish that the work is similarly negative. This line of reasoning will be referred to as the *work as a dot product argument*. The second line of reasoning, the *work as a change in energy argument*, uses a statement of the work-energy theorem (*i.e.,* $W_{net,ext} = \Delta KE + \Delta PE$) with the observation that the particle is slowing down to argue that since the kinetic energy is decreasing, and a point particle has no change in potential energy, the work done on the particle by the force must be negative. This line of reasoning could be simplified by invoking the work-kinetic energy theorem (*i.e.,* $W_{net,ext} = \Delta KE$), and thus disregarding arguments related to potential energy.

On the basis of student responses to similar questions in other formats, the most common incorrect response involves concluding that the work on the particle by the force is positive because the force is pushing to the right, which is assumed to be the positive direction.

4.5.1.2 Chaining task implementation

The reasoning elements provided to students on the chaining version of the work-energy task were expressly designed to reflect both the *work as a dot product argument* and the *work as change in energy argument*, and are shown in Figure 4-4. While the common incorrect line of reasoning may also be constructed from the elements provided, all of the reasoning elements (with the exception of the incorrect conclusion elements) are true statements.

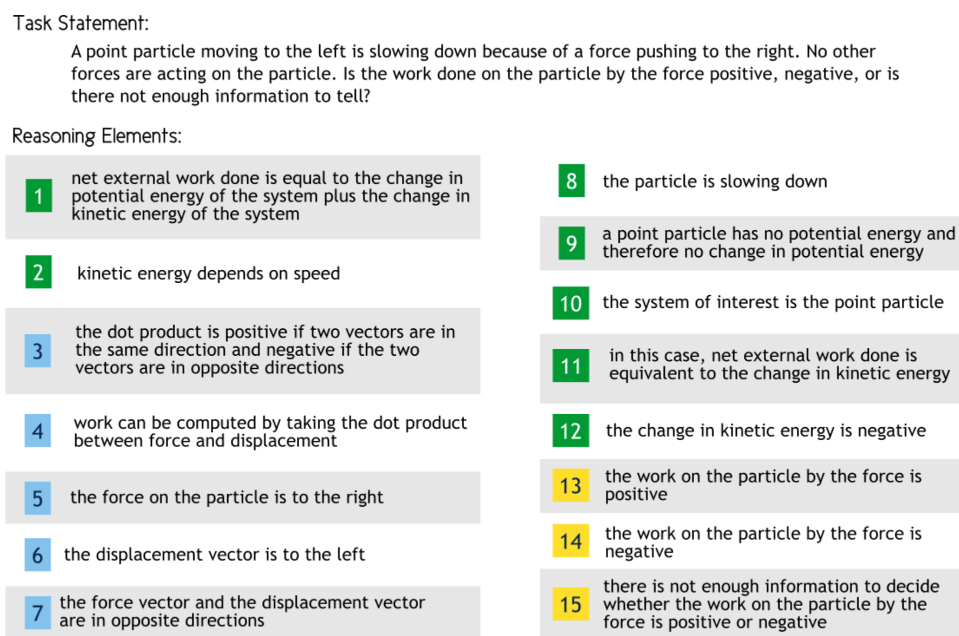


Figure 4-4. Work-energy task. Question prompt and associated reasoning elements provided to students are shown. The elements are numbered for later reference and color coded based on whether they were intended for the *work as a change in energy argument* (green) or for the *work as a dot product argument* (blue) or are conclusion elements (yellow).

4.5.1.3 Performance overview

Of the 119 students who completed the chaining version of the work-energy task, 92% of them answered correctly that the work done by the force on the particle is negative. Of these responses, 69% responded with the *work as a dot product argument*, 12% responded with the *work as a change in energy argument*, and 16% included both arguments. Figure 4-5 shows an example of each type of student response.

We have purposefully chosen to introduce network analysis using the work-energy task due to the unambiguous nature of the collected data set, as this allows us to demonstrate the applicability and power of the network analysis tools before examining more complex, nuanced data sets. Because of the strong overall performance on the work-energy task, it is likely that students had a solid grasp of the reasoning involved in answering the question, and we therefore expected this to be reflected in their reasoning chains. Furthermore, since many students articulated each independent argument (energy and/or dot product), we recognized that these lines of reasoning would be clearly represented in a network constructed from all student responses. As a result, this set of student responses represents an ideal test case for the application of the network analysis methods described above in the context of reasoning chain construction tasks.

Example Work as a Change in Energy Argument
net external work done is equal to the change in potential energy of the system plus the change in kinetic energy of the system
the particle is slowing down
because
the force on the particle is to the right
so
the change in kinetic energy is negative
therefore
the work on the particle by the force is negative

Example Work as a Dot Product Argument
work can be computed by taking the dot product between force and displacement
and
the dot product is positive if two vectors are in the same direction and negative if the two vectors are in opposite directions
because
the force on the particle is to the right
and
the displacement vector is to the left
the work on the particle by the force is negative

Example Response with Both Arguments
net external work done is equal to the change in potential energy of the system plus the change in kinetic energy of the system /and / the system of interest is the point particle / but / a point particle has no potential energy and therefore no change in potential energy / so / in this case, net external work done is equivalent to the change in kinetic energy / and / kinetic energy depends on speed / and / the change in kinetic energy is negative / because / the particle is slowing down / work can be computed by taking the dot product between force and displacement / so / the dot product is positive if two vectors are in the same direction and negative if the two vectors are in opposite directions / because / the force on the particle is to the right / but / the displacement vector is to the left / therefore / the force vector and the displacement vector are in opposite directions / therefore / the work on the particle by the force is negative

Figure 4-5. Examples of each type of response to the work-energy task. The example response with both has been condensed, while the other two examples show what the chain would have looked like to the student.

4.5.1.4 Community detection analysis of correct responses

We constructed both a direct and an indirect association network from the correct responses to the work-energy task and applied the community detection algorithm to each separately. (Recall that, as discussed in Section 4.4.1, a direct association network only links elements that are placed consecutively in a student response, while an indirect association network links each reasoning element in a response with every other reasoning element in that response.) The results from that analysis are shown in Figure 4-6. In the figure, the elements that are important to the *work as a dot product argument* are colored blue and the elements important to the *work as a change in energy argument* are colored green.

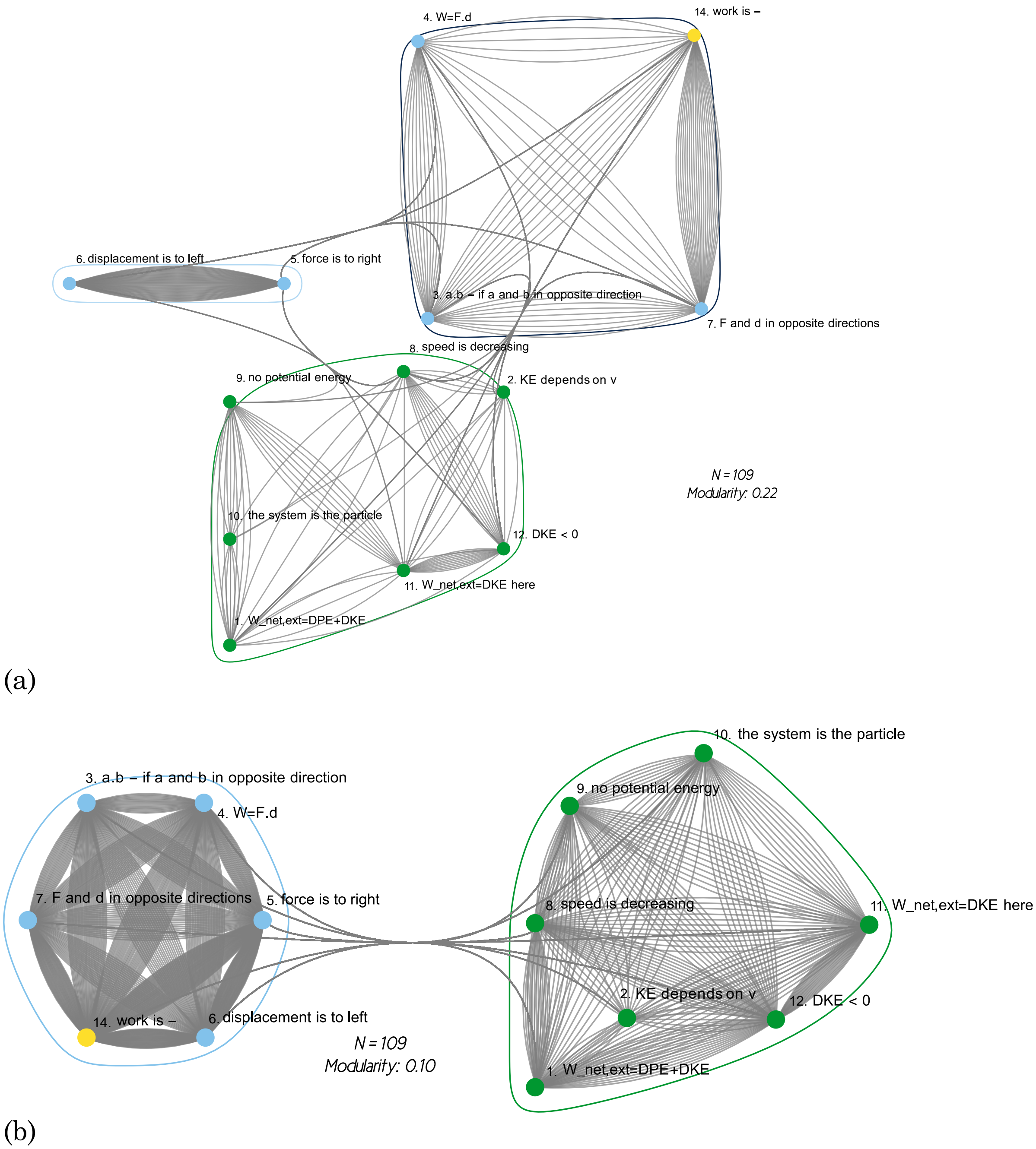


Figure 4-6. A representation of the communities found in (a) a direct association network and (b) an indirect association network built from correct responses to the work-energy task described in Section 4.5.1.2. Elements that are aligned with a *work as dot product argument* are colored blue and the elements aligned with the *work as a change in energy argument* are colored green. The answer element is colored yellow.

In both the direct and indirect association networks, the elements in the *work as a dot product argument* and the elements in the *work as a change in energy argument* are found by the community detection algorithm to be separate from each other. Additionally, the community structure of the direct association network reveals that the work as dot product elements appear to have two groupings: one with the two elements that state that the force vector is to the right and the displacement vector is to the left, and one with the rest of the work as dot product elements.

We wish to note here that these results show that the two types of networks, direct and indirect, yield differing levels of detail and indeed different types of information about the set of student responses represented. Thus, it is valuable to examine both types of networks. More will be said about this in Section 4.5.1.6.

4.5.1.4.1 Bootstrapping Community Detection Results

To assess the stability of the communities found via the optimum modularity community detection algorithm, bootstrap tests were administered by repeatedly testing “hypothetical” networks constructed from resampled correct responses, as explained in Section 4.4.2.2. We first discuss our examination of the communities arising in the direct association network, and then turn our attention to the communities in the indirect association network.

For the direct association network, in every bootstrap test, the elements associated with the work as a change in energy argument and the work as a dot-product argument were well separated from each other. For example, consider the bootstrapping frequency plots shown in Figure 4-7.a and Figure 4-7.b. The plots indicate the percentage of the bootstrapping trials in which each element was included in a specified community. For these tests, we defined membership in the work as a change in energy community as being in the same community as the general statement of the work-energy theorem (*i.e.,* element 1), and membership in the work as dot-product community as being in the same community as the statement of work as a dot-product (*i.e.,* element 4). The frequency plots reveal that the two arguments are well separated in the network since no element associated with the work as a change in energy argument appears in the work as a dot product community, and vice versa, in close to 100% of the trials.

The two-element "force is to the right" and "displacement is to the left" community shown in Figure 4-6.a was only preserved in 35% of bootstrapping runs when testing for the presence of that community on each iteration. On its surface, such a result would seem to call into question the robustness of that structure. However, there is indeed a stronger association between those two elements than any other two elements in the network; there is a link weight of 39 between those two elements, whereas the next strongest link weight is only 18 (not shown). The frequency plot for that community (shown in Figure 4-7.c) shows that the two elements are always coupled together in the same community (1000 times out of 1000) but that between 30% to 40% of the time, the elements concerning the dot product (elements 3 and 4) are also included. Taken together, then, these results indicate that this two-element structure is indeed present in the network and that the frequency plot may be a more reliable method for obtaining information about the robustness of community structure than simply testing for the existence of the community with the initial structure.

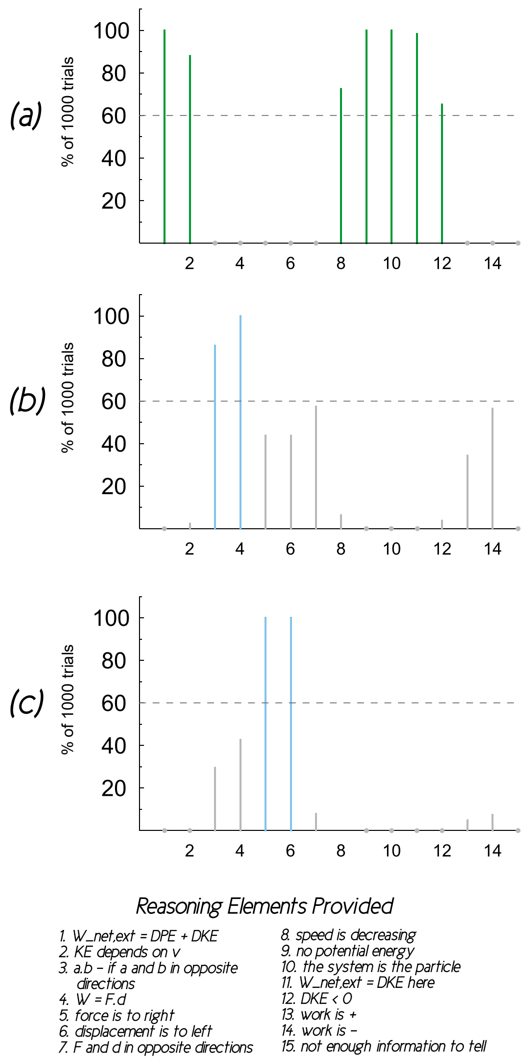


Figure 4-7. Bootstrapping frequency plot for three communities, including (a) the *work as a change in energy* community, (b) the *work as a dot-product* community, and (c) the two-element force and displacement community. The plot indicates the percentage of the trials in which each element was included in the specified community. A dotted line corresponds to the 60% threshold used for ascertaining community membership in the bootstrapping tests.

For the indirect association graph, we administered three bootstrap tests. In the first bootstrap test, we tested the hypothetical network for the exact community structure shown in Figure 4-6.b and found that 88% of networks had that exact same structure. We also conducted bootstrap tests where, on each iteration, we tested which elements were in the same community as (a) the general statement of the work energy theorem (element 1) and (b) the statement of work as a dot product, (element 4), as with the direct association graph. Based on the bootstrapping frequency plots (not shown), all of the *work as a change in energy* argument elements are found 100% of the time in the community with the statement of the general work-energy theorem, and the elements related to the *work as a dot product* argument are likewise found 100% of the time with the statement of work as a dot product. Thus, we felt very confident in the robustness of community structure depicted in Figure 4-6.b.

4.5.1.5 Network Sparsification Method Applied to Work Task Correct Responses

We now explore the usefulness of network sparsification by analyzing a direct association network built from the correct responses to the work task. Figure 4-8 shows a sparsified version of the direct association network at a threshold of $\alpha = 0.2$. The elements in this figure are color coded according to the same color scheme used in Section 4.5.1.2.

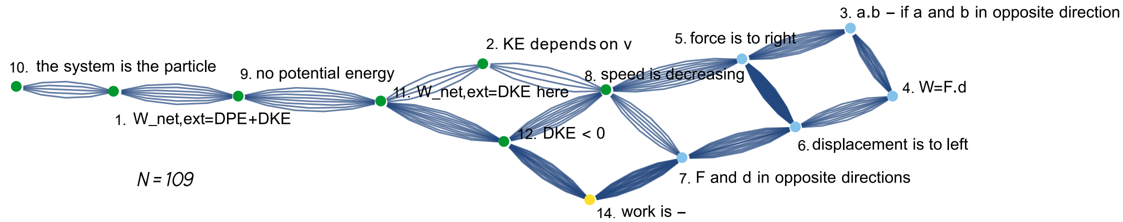


Figure 4-8. A representation of a sparsified ($\alpha = 0.2$) direct association network built from correct responses to the work task. The elements are color coded according to the line of reasoning they are useful for: green elements are useful in the energy argument, and blue elements are useful in the dot product argument.

In Figure 4-8, it can be seen that the two independent arguments are again separated as distinct in the network since the elements associated with the energy argument are separate from the elements associated with the dot product argument. Furthermore, examination of the network reveals the existence of two clear chains of reasoning, each of which appears to include general principles (such as the work-energy theorem or the definition of work as the dot product of the force and displacement vectors) and then to step through the application of the specifics in the problem statement before finally arriving at an answer. By constructing a directed network and calculating the ratio of out-degree to in-degree (as explained in Section 4.4.1), it was shown that the element "the system of interest is the point particle" (element 10) is indeed a starting point for students (out:in is 3.0) as well as "work can be computed …" and "the dot product is…" (elements 3 and 4; out:in is 2.0 and 1.3, respectively). Additionally, the answer element is an end-point with an out:in of 0.1. Thus, based on the sparsified undirected graph and the information about out to in degrees of the directed graph, the students in this case appeared to generally be starting with 1st principles and applying situation specific constraints to arrive at an answer.

4.5.1.5.1 Assessing the fidelity of the sparsified representation

While the features of the sparsified graph are of interest, it is also good to assess, to the extent possible, whether they are true representations of the network structures, or whether they are artifacts of the sparsification process. To assess the fidelity of the sparsified representation, we compare features of the sparsified network to network measures applied to the unsparsified network.

The first feature of interest is the observed topology of the network. The topology of the work as a change in energy argument elements, shown in Figure 4-8, is observed to be quite linear, while the topology of the elements associated with the *work as a dot product argument* is more interconnected. These apparent topological differences are reflected in the global clustering coefficients for each argument. Analysis of an unsparsified sub-network composed of solely the elements in the *work as a change in energy* argument yields a clustering coefficient of 0.48. The global clustering coefficient of an unsparsified sub-network consisting of just the elements in the *work as a dot-product argument* is 0.89 -- substantially higher. Thus, the relative interconnectedness of each of these arguments in the original, unsparsified networks (indicated by the clustering coefficients) appears to be preserved even after the sparsification process (indicated by the topology of the sparsified network); this consistency highlights both the fidelity and reliability of the chosen sparsification technique in retaining key characteristics of the network structure.

Another observed feature of the network structure is the element, “the particle is slowing down” (element 8) that bridges the two independent arguments. We sought to ascertain whether or not this element also served as a bridge in the unsparsified network. Bridges tend to have higher betweenness centrality as they are essential to the flow of information through a network (upon which the betweenness centrality is based), which means that betweenness centrality is a good measure to assess whether the feature is a bridge in the unsparsified network. The two elements in the unsparsified network with the highest betweenness are “the change in kinetic energy is negative” (element 12) and “the particle is slowing down” (element 8). These two elements, incidentally, have the same betweenness. Furthermore, in the sparsified network, those two elements also have the highest betweenness centrality. Thus, the sparsified structures appear to be reliable representations of the original network structures on the basis of betweenness centrality as well.

The location of “the particle is slowing down” as a bridge in the network may be attributed to that particular element being used frequently in both the work as a dot-product argument and the work as a change in energy argument. Upon more detailed analysis of student responses, it was found that in the work as a change in energy argument, the element was used to justify why the kinetic energy (and thus the work) is negative, whereas in the work as a dot-product argument, the element was used to describe the consequence of the force and displacement being in opposite directions. This latter use may have stemmed, in part, from students referencing the task prompt, which noted that the particle “is slowing down because of a force pushing to the right”.

4.5.1.6 Discussion of Results

The separation of the elements into two distinct lines of reasoning in both the community detection results and the sparsification results shows that network analysis of data drawn from the reasoning chain construction task can explore, in a meaningful way, the content of the various arguments constructed by students. In particular, the results show the role that each type of network (indirect vs. direct association) can play in examining student reasoning. Based on our analyses, finding communities in the indirect association network seems best suited for determining which lines of reasoning are present among the responses, whereas community detection applied to direct association networks allows for greater resolution of the sub-arguments that make up those lines of reasoning.

Bootstrapping is an indispensable part of community detection. The bootstrapping frequency plot revealed a fairly stable sub-argument structure in the direct association network comprised of the elements “the force on the particle is to the right” and “the displacement vector is to the left”. We would expect those two elements to be more closely associated with each other in the network since they were often placed next to each other in student responses. Indeed, the algorithm is sensitive to that structure. It is important to note that bootstrap testing for an exact community structure is less informative than a bootstrapping frequency plot (recall the two-element sub-structure in Figure 4-6.a, as the latter can determine which elements are more likely to be in a given community.

The sparsified network appears to give information about how students viewed the structure of an argument. The linearity of the work as a change in energy argument and the non-linearity of the work as a dot-product argument suggest a difference in how students approached those two arguments. On the face of it, the linearity or non-linearity of the associations between a group of elements indicate that many students either responded with similar ordering of the elements (creating a linear network) or that there was not a preference for which elements came before others in the reasoning chain (creating a clustered, non-linear network). It could be that this is inherent to the elements provided, or it could be indicative of a particular learned approach to a problem. As noted in the sparsification results section, the students in this case appear to have started with first principles and the application of situation specific constraints in order to arrive at an answer. Perhaps in an energy setting, students recognized that defining a system needed to occur before the application of the general work energy theorem. In contrast, the implication of the dot product on the sign of the work when the vectors were in opposite directions (*i.e.,* element 3) is not necessarily an important next logical step after establishing that "work can be computed from the dot product of force and displacement" (element 4). If it were, the network would have appeared much more linear, with element 4 being linked only to element 3, from which the rest of the network would be linked. Students instead appeared to proceed to information about the force and displacement vectors before discussing the mathematical aspects of the dot product.

Most importantly, the ability to quickly and efficiently determine information about how a large group of students is approaching a line of reasoning can be very useful to instructors and researchers alike, even if the specific interpretation of the structure is not always immediately apparent.

It is important to note, however, that the clear chain of reasoning shown in the sparsified graph does not necessarily represent the chain of reasoning constructed by the majority of individual students. Actually, only 2 students out of 100 responded with chains that included the first four elements of the energy argument (namely, elements 1, 9, 10 and 11) in the order represented in Figure 4-8, and only 8 used all four elements in their chain. Many students only cited parts of the argument, inserted irrelevant elements into their argument, arranged the argument differently, *etc.;* still, these students constructed their arguments in a way that led to the majority of the associations being between those four elements in the ordering shown in Figure 4-8. Thus, the sparsified network represents a "wisdom of the crowd" (Galton, 1907; Surowiecki, 2004) result, a synergistic classroom consensus on how the elements ought to be arranged that transcends the reasoning chains constructed by individual students.

Further evidence of this synergistic consensus or wisdom of the crowd is provided by the results of the betweenness calculations. In the full, unsparsified network of correct student responses, the element "the speed is decreasing" served as a bridge between the two independent arguments and therefore has a high betweenness centrality. However, that particular element was not used by any single student to bridge the two arguments in his or her reasoning chain. Instead, the element's high betweenness centrality offers a glimpse into how the students as a whole viewed that particular element; in the logical landscape of this problem, the information that the speed is decreasing can be seen as relevant to both arguments. An implication of this dual-relevancy is that this single element may serve as a possible pivot point for shifting from one argument to the other during, for example, a classroom discussion of the solution to the task.

A more general implication of the synergistic nature of the reasoning chain network (whether sparsified or not) is that the betweenness centrality of an element is not necessarily related to the position of that element in any given chain, but rather the position of that element in the collection of all chains. The two are coupled, of course, because if a certain element is placed at the beginning of a chain by every student, that element would have a low betweenness score. However, an element that is always placed in the middle of a chain may not necessarily have high betweenness in the resulting reasoning chain network unless that element is shared among many different types of chains or orderings of a particular argument. As an example, consider the element "a point particle has no potential energy and therefore no change in potential energy" (element 9). From the sparsified network, this element was likely consistently placed in the middle of individual student chains, but its betweenness is low (5th from lowest) because it was always placed in the middle of the same student chain (further study of the individual reasoning chains confirmed this to be the case). Thus, betweenness centrality measures centrality to the wisdom of the crowd or classroom consensus reasoning.

This classroom consensus reasoning can be useful in identifying where a class stands with respect to the usage of certain arguments. For instance, the work task was administered to two different calculus-based introductory mechanics courses at the same university, but with different instructors who had different instructional emphases. The sparsified network shown in Figure 4-8 was derived from student responses during one these courses and represents a full *work as a change in energy argument*, whereas the sparsified network of responses from the other class (not included in this paper) gave a truncated work as a change in energy argument that only associates the elements “in this case, the net external work done is equal to the change in kinetic energy” and “the change in kinetic energy is negative” before arriving at an answer. The *work as a dot product argument*, however, appeared to have been articulated in full by students in that same class. Since the arguments associated with the definition of work as a dot product in both classes were similar, the difference in how the work as a change in energy argument was approached by these two classes could be due to factors such as the focus of instruction, the epistemological stance of the instructor and/or students, mastery of work-energy related content, *etc.* Our network data alone cannot isolate the reason for the difference, but they do provide a method of quickly ascertaining the nature of the difference. Thus, we find chaining tasks coupled with network analysis to be a useful diagnostic tool in investigating student reasoning patterns throughout instruction.

While the community detection results and the sparsification results were largely complementary in our analysis of the work-particle task, it isn't necessarily the case that elements found to be tightly associated with each other using community detection will be as tightly associated with each other in the sparsified network. The main reason for this is that each analysis method is answering a different question about the associations made by the students. Community detection answers the question “Which elements are more tightly associated with each other than with the rest of the network?”, whereas the sparsification method answers the question “What is the structure of the associations made between all of the elements?”. As a specific example of how the answers to these questions can differ for the same task, we found in the work-particle task that the element “the particle is slowing down” was more tightly associated with the work as a change in energy argument than with the *work as a dot product argument*; however, sparsification revealed that, structurally, the element was shared between both arguments.

4.5.2 Truck Friction task

In the previous section, the strong student performance on the work-energy task helped us illustrate the power of network analysis methods in characterizing student responses to reasoning chain construction tasks. In this section, we analyze the results of reasoning chain construction task that, like the work-energy task, has two independent pathways for answering correctly, but which is considerably more difficult for students.

The truck friction task examines three main research questions.

1. How effective are network analysis methodologies at characterizing and differentiating among different lines of reasoning on a physics question that is more challenging for students?
2. What are the limitations associated with reasoning chain construction tasks, and can the tasks be modified via adjustments to the list of reasoning elements to address such limitations?
3. To what extent can network analysis be used to identify and document evidence in support of specific theoretical constructs (e.g., dual-process theories of reasoning or resources) in reasoning chain construction task data?

4.5.2.1 Physics question overview

In this task, a box is resting on the back of an accelerating truck, as shown in Figure 4-9. Students are told that "the truck is moving to the right and speeding up (*i.e.,* the truck is accelerating to the right)" and that the box is not moving with respect to the truck. They are asked to determine the direction of the force of static friction from the truck on the box.

There are several approaches that may be used to arrive at the correct answer that the static friction is directed to the right. In a more formal approach, it is recognized that the net force on the box must be in the direction of the acceleration of the box (from Newton's second law), which is to the right. Since the only horizontal force acting on the box is the force of static friction, the net force is equivalent to the static friction force. Thus, the static friction force must be directed to the right. The two main arguments in this approach (net force is in the direction of acceleration and the static friction force is equivalent to the net force) are independent of each other but must both be considered in order to logically deduce that the static friction force must be directed to the right.

A common alternative approach is to construct a hypothetical argument that, in the absence of friction, the box would slide toward the back of the truck (*i.e.,* this is the impending motion). Thus, since the box is not sliding to the left with respect to the truck, the friction force must be opposing the impending motion and is therefore directed to the right.

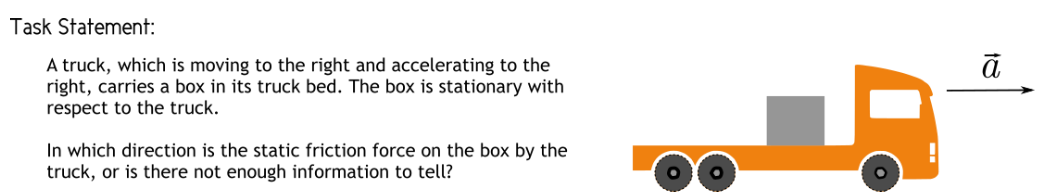


Figure 4-9. Task statement and diagram given to students on the Truck Friction task.

Based on previous research regarding this task and ones similar to it, a common incorrect way of answering this question is to reason that friction opposes the actual motion (as opposed to the relative or impending motion) and that since the box is moving to the right, the friction must point left to be in opposition to that motion. From free response data to this task, we found that students also commonly add that the friction is opposing the force of motion to the right and cite Newton's third law to justify that they are equal in magnitude. (These same students still maintain that the static friction is directed to the right.) The common incorrect line of reasoning is consistent with conceptions of friction noted in literature (*e.g.,* Besson 2007).

From a preliminary study in which this question was asked as a multiple choice plus explanation question, we found that of 115 respondents, 22% of students used the formal Newton's 2nd Law reasoning, 37% of students used the correct hypothetical argument, and 16% of students responded with the common incorrect line of reasoning. The remaining students either gave no explanation (11%) or gave explanations that were either ambiguous or fell into categories too small to be considered separately (< 6% each). On the basis of our data, the hypothetical argument is the predominate lines of reasoning used by those students who gave correct answers.

4.5.2.2 Chaining task implementation

As with the work-energy task, we created reasoning elements (shown in Figure 4-10) that would encapsulate both correct lines of reasoning as well as provide an option for piecing together an incorrect line of reasoning. Again, each reasoning element provided to the students contained a true statement, and students were notified of this fact in the task prompt. Still, some elements could be incorporated into an erroneous line of reasoning if interpreted incorrectly. An example is the element “the force of static friction always opposes the impending motion” (element 8), which could be read incorrectly by some students to mean that friction opposes motion generally and used in the incorrect line of reasoning.

In Figure 4-10, elements that are useful for the formal line of reasoning are color coded based whether they are intended to be part of the sub-argument establishing that the net force is to the right (blue) or part of the sub-argument argument establishing that the net force is equivalent to the static friction (green). The correct hypothetical argument elements, “without friction, the box would move to the left with respect to the truck” (element 13), “the box is not moving with respect to the truck” (element 4) and “the force of static friction always opposes the impending motion” (element 8) are shaded dark blue. Finally, the answer elements are colored yellow and all other elements are colored gray.

1 the truck is accelerating to the right	10 the acceleration of an object is in the same direction as the net force on the object
2 the truck is moving to the right	11 the normal force and the gravitational force sum to zero (i.e., they "balance")
3 the box is moving to the right	12 the static friction force on the box must be in the same direction as the net force on the box
4 the box is not moving with respect to the truck	13 without friction, the box would move to the left with respect to the truck
5 there are three forces acting on the box: the static friction force by the truck, the normal force by the truck, and the gravitational force by the earth	14 the net force is equal to the static friction force
6 the net force on the box is the vector sum of all the forces acting on the box	15 the net force is to the right
7 for every force on object 2 by object 1, there is an equal and opposite force on 1 by 2 via Newton's 3rd law	16 the box is accelerating to the right
8 the force of static friction always opposes the impending motion	17 the static friction force from the truck on the box is to the right
9 the truck and the box have the same acceleration	18 the static friction force from the truck on the box is to the left
	19 the static friction force from the truck on the box is zero
	20 there is not enough information to determine the direction of the force of static friction on the box

Figure 4-10. Reasoning elements provided to the student on the truck friction task. In a modified version of the task (see Section 4.5.2.6), the elements “without friction, the box would move to the left with respect to the truck” (element 13) and “the box is moving to the right” (element 3) are not present. The elements are color coded as explained in the text.

4.5.2.3 Performance overview

On the truck friction task, 50% of students answered correctly on the chaining task by selecting that the static friction was to the right, while 43% of students selected the common incorrect answer (static friction is to the left).

An overview of the categories of reasoning chains constructed by students is given in Table 4-1. The coding of the categories was based on the elements employed. An argument was classified as *formal reasoning* if it included elements from both sub-arguments and also did not include element 13, “without friction, the box would move to the left with respect to the truck”. An argument was classified as *correct hypothetical reasoning* if it included the element “without friction, the box would move to the left with respect to the truck” (element 13) and did not include reference to a net force. Some students appeared to use both the hypothetical and formal arguments in their response, such as the following student response:

“the box is accelerating to the right / but / the box is not moving with respect to the truck / without friction, the box would move to the left with respect to the truck / so / the net force is to the right / because / the acceleration of an object is in the same direction as the net force on the object / and / the static friction force on the box must be in the same direction as the net force on the box / therefore / the static friction force from the truck on the box is to the right”

An argument was classified as “common incorrect reasoning” if the student employed the element “friction opposes motion” and also selected the answer “the static friction from the truck on the box is to the left”, regardless of what other elements the student included in his or her reasoning chain.

Reasoning employed	Percentage of Students (N = 116)
Formal Reasoning	22%
Correct Hypothetical Reasoning	23%
Both Formal and Hypothetical	3%
Common Incorrect Reasoning	42%

Table 4-1. An overview of the categories of reasoning chains constructed by students on the truck friction task.

4.5.2.4 Arguments Found via Community Detection

A representation of the communities found in an indirect association network comprised of all responses to the truck friction task is shown in Figure 4-11. Community detection again reveals meaningful separations among the elements. In the community that includes the common incorrect answer element, “the force of static friction from the truck on the box is to the left”, there are three other elements: “the force of static friction always opposes the impending motion”, “the truck is moving to the right”, and “the box is moving to the right”. These elements are consistent with a common incorrect response.

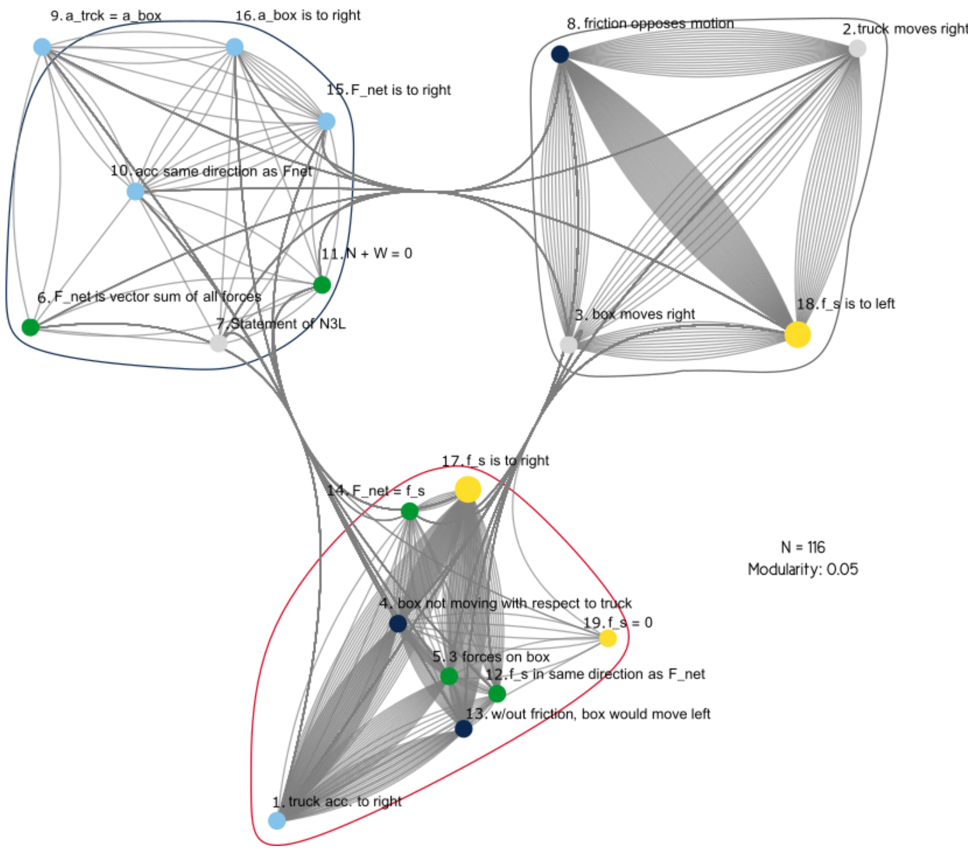


Figure 4-11. A representation of the communities found in an indirect association network comprised of responses to version A of the truck friction task.

The community that includes the correct answer element is more complex, but appears to include elements that we would expect to be associated with the two different lines of correct reasoning -- the hypothetical and the formal. Furthermore, there is a third community, not associated with any answer element in particular, that is comprised mostly of elements regarding the acceleration of the truck and the box. By examining the communities found in direct and indirect association networks comprised of only correct or incorrect answers, we determined that this community appears to be elements that are shared between the two predominant answers ("f_s to the right" and "f_s to the left") and is also conflated with a sub-argument structure for the correct answer (the argument establishing that the net force is directed to the right).

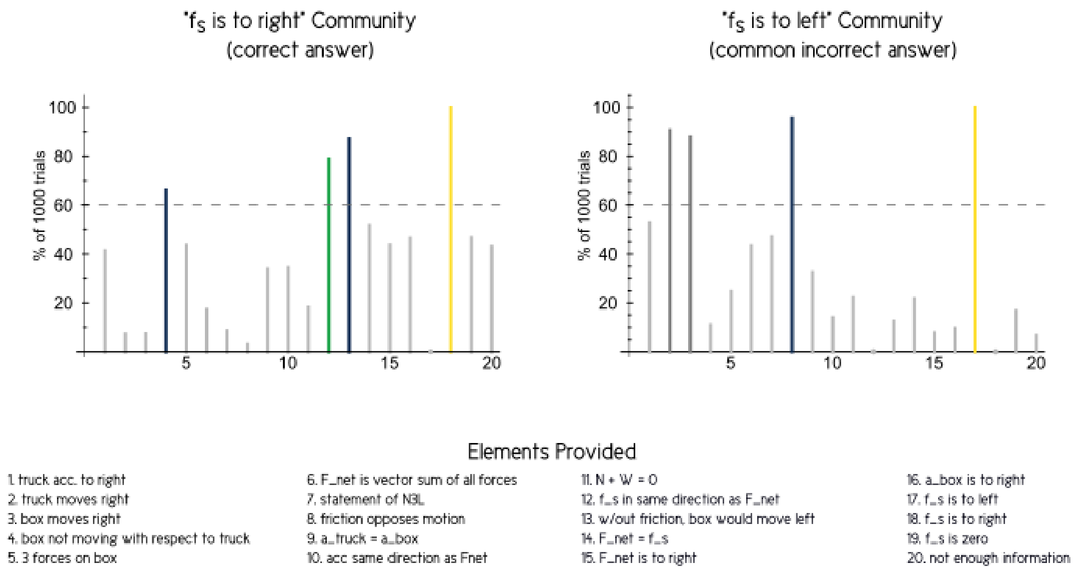


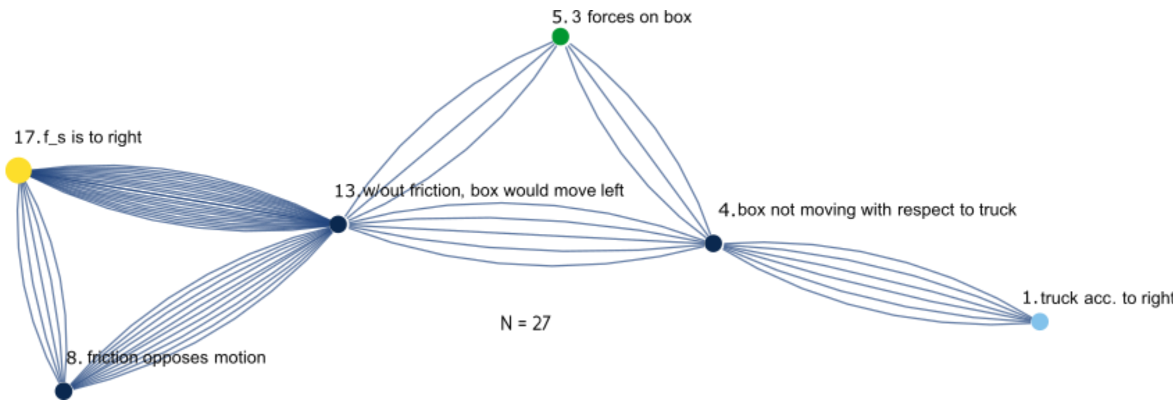
Figure 4-12. Frequency plots of the results for 1000 iterations of a bootstrap that tallied the elements contained in the same community as the indicated answer element. Results are shown for the correct and common incorrect answer. A threshold of 60% is indicated by the horizontal bar. The plots are color coded according to the color of the corresponding elements in the community plot shown in Figure 4-11.

Figure 4-12 shows the results of the element frequency bootstrapping method discussed in Section 4.4.2.2. From the results, it can be seen that the common incorrect community is likely to be comprised of the elements "the box moves to the right", "the truck moves to the right", and "the force of static friction always opposes the motion". For the correct answer element, the community is comprised of "without friction, the box would move to the left with respect to the truck", "the box is not moving with respect to the truck", and "the static friction force must be in the same direction as the net force on the box". Since the elements associated with the hypothetical line of reasoning have a much higher frequency, it indicates that the hypothetical line of reasoning is used more often in support of the correct answer than the formal reasoning. The content of this community (those three elements) are consistent with what we would expect from that line of reasoning.

4.5.2.5 Topology of Argument Structure via Sparsification

We wished to examine the structure of both the correct arguments and the incorrect arguments made by students with the reasoning elements provided to them. In the truck friction task, we separated responses based on which answer element was present in the response and created direct association networks. To better study the topology of the correct hypothetical argument (the predominant line of reasoning employed in the correct responses), we included in the correct response network only responses that included element 13 ("without friction, the box would move to the left…"). We then sparsified these correct response and incorrect response networks to obtain information about the topology of the argument structure. The result is shown in Figure 4-13.a, which is the correct hypothetical argument, and Figure 4-13.b, which is the common incorrect argument. The number of responses in each network is also indicated in the figure.

(a)



(b)

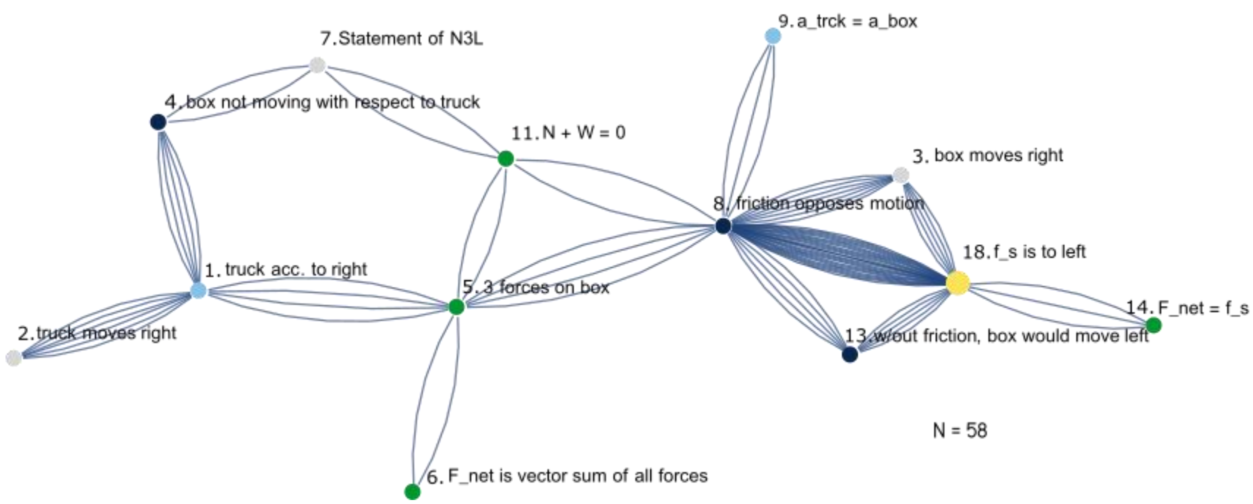


Figure 4-13. Sparsification of direct association networks comprised of (a) correct responses that use the hypothetical line of reasoning (α= 0.2), and (b) responses endorsing the common incorrect answer (α = 0.1).

In the sparsified networks comprised of responses endorsing the common incorrect answer and the correct answer, it is seen that many of the same elements are selected by the student to place in their reasoning chains.

These elements include, “the box is not moving with respect to the truck” (element 4), “the truck accelerates to the right” (element 1), “there are three forces acting on the box […]” (element 5), “the force of static friction always opposes the impending motion” (element 8) and “without friction, the box would move to the left with respect to the truck” (element 13). However, although both networks include the same elements, these elements are arranged in different topologies in the two networks. Additionally, the element “the box moves to the right” (element 3) seems to be uniquely important to the network of incorrect answers. Interestingly, the element “the force of static friction always opposes the impending motion” in the incorrect answer network occupies the same central position as the element “without friction, the box would move to the left with respect to the truck” in the correct answer network. A calculation of betweenness centrality for both networks in their unsparsified form (shown in Table 4-2) reveals that these two elements have high betweenness in their respective networks. Thus, even though both populations of students used the same subset of elements, the structure of the associations made between those elements indicates that emphasis was placed on different elements.

Abbreviated Element Label	Betweenness in Network Comprised of Correct Responses using Hypothetical Reasoning		Abbreviated Element Label	Betweenness in Network Comprised of Common Incorrect Responses
w/out friction, box would move left	93.5 (1.0)		friction opposes motion	136.5 (1.0)
box not moving with respect to truck	44 (0.47)		truck acc. to right	89 (0.70)
friction opposes motion	27 (0.29)		f_s is to left	29 (0.23)
F_net is to right	25 (0.27)		box not moving with respect to truck	22 (0.17)
f_s in same direction as F_net	19 (0.20)		N + W = 0	18 (0.14)

Table 4-2. Weighted betweenness centrality calculations (via (Opsahl, Agneessens, & Skvoretz, 2010)) using unsparsified networks. Only the top five elements are shown in each case. The normalized betweenness is reported in parentheses.

4.5.2.6 Modified version of the truck friction task to study sub argument structure

In order to study the structure of the formal line of reasoning in greater detail, a version of the truck friction task was designed that did not include elements 13 ("without friction, the box would move to the left with respect to the truck") or 3 ("the box is moving to the right"). Removing these elements from the list was intended to preclude the use of the hypothetical argument, thus allowing us to isolate the formal line of reasoning.

This modified version of the task was administered to a different student population: students enrolled in the same course in a different semester. On this modified version of the task 68% of students answered correctly and 28% selected the common incorrect answer. An overview of the categories of reasoning chains constructed by students on the modified version is given in Table 4-3. The coding of the categories was the similar to the coding scheme described in Section 4.5.2.3. In this version, the formal argument was predominant among the correct responses (and indeed among all responses) rather than the hypothetical argument. However, it was observed that on the modified version, in which the hypothetical statement (element 13) was removed, there were a subset of students who were using the "friction opposes impending motion" element in a way that suggested they were attempting to use the hypothetical argument and weren't able to do so fully with the elements provided. An example of this type of response is shown in Figure 4-14. Such students were classified as using the hypothetical correct reasoning in Table 4-3.

"box not moving with respect to truck / and / truck acc. to right /
f_s is to right / because / friction opposes motion"

Figure 4-14. Example student response where the student appeared to be attempting to use the hypothetical argument but were unable to do so because of the constraints of the modified version of the task (i.e., that certain elements were removed from the provided list in that version).

While we suspect that the differences in the percentage of students using the formal line of reasoning is related to our removal of the key element (element 13) essential to the hypothetical argument, we cannot attribute any causality due to the populations having different instructors and being in different courses, *etc.*

Reasoning employed	Percentage of students (N=111)
Formal Reasoning	47%
Correct Hypothetical Reasoning	13%
Both Formal and Hypothetical	4%
Common Incorrect Reasoning	19%

Table 4-3. An overview of the categories of reasoning chains constructed by students on the modified version of the truck friction task.

However, noting that the prevalence of the formal line of reasoning is higher in the modified version of the task compared to base version allows us to study the formal line of reasoning more clearly in that population. Figure 4-15 shows the results of the element frequency bootstrapping method discussed in Section 4.4.2.2. Recall that for the base version, in which the hypothetical argument was accessible, the correct answer element community revealed a strong preference for the hypothetical line of reasoning and only had one element from the formal line of reasoning included. The correct community for the modified version reflects the usage of the formal line of reasoning and shows that all of the elements associated with that line of reasoning are above the threshold for inclusion in the community except for the element describing the three forces acting on the box (element 5). That element was just over the threshold (60% of 1000 iterations) for inclusion in the common incorrect community.

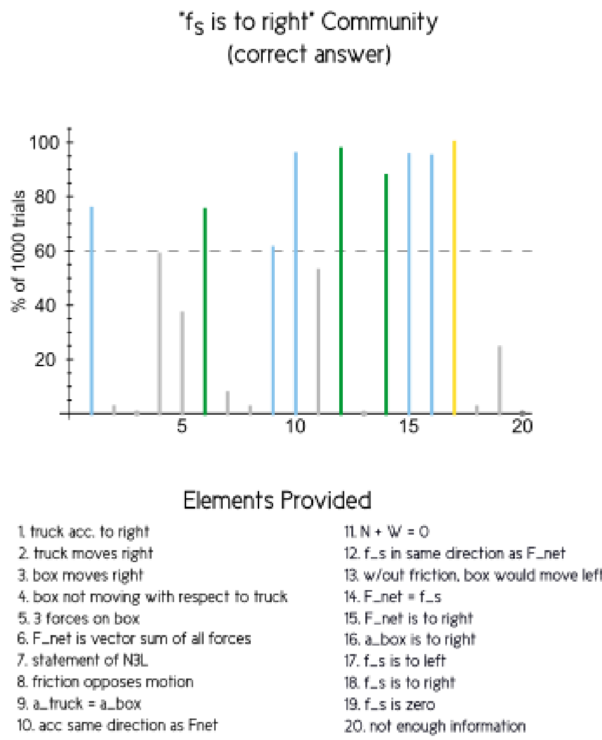


Figure 4-15. Frequency plots of the results for 1000 iterations of a bootstrap which tallied the elements contained in the same community as the indicated answer element. Results are shown for the correct responses to the modified versions of the task. A threshold of 60% is indicated by the horizontal bar. The plots are color coded according to the color of the corresponding elements in the community plot shown in Figure 4-11.

Finally, we constructed a direct association network from the correct responses to the modified version and sparsified that network. The result is shown in Figure 4-16. The sparsified network constructed with correct responses to the modified version of the truck-friction task shows a complex line of reasoning. The cycles (circular structures) in the network imply a multi-path flow in which each path is fairly ordered and linear. Using directed networks to ascertain starting and ending points, it was determined that the primary starting point is the element "the truck accelerates to the right". (This may be due to the fact that this element was listed first in the "items" column, although further work would need to be done to ascertain whether or not that is the reason why this element served as a starting point.) Taking "the truck accelerates to the right" as the starting point, it becomes apparent that students collectively made associations among the elements that would tend to create a flow from one sub-argument (the net force is to the right argument) through the other sub-argument (static friction is equivalent to the net force) to arrive at an answer. It is worth noting that the directed network (not shown) for the formal line of reasoning generally affirms this result. The cycles in this sparsified network show a more complex structure of associations than on the base version of the task (in which correct reasoning primarily relied on the hypothetical argument).

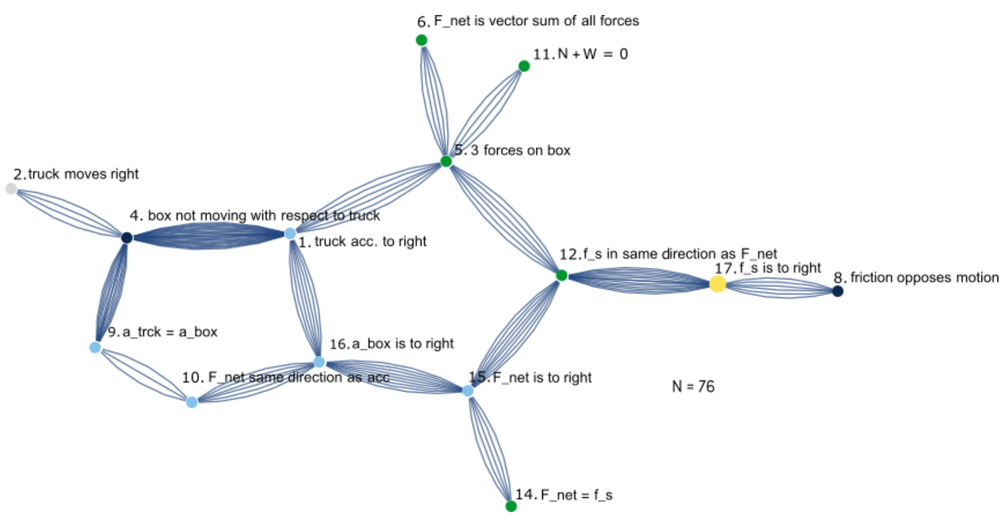


Figure 4-16. Sparsification of direct association networks comprised of all correct responses to the modified version of the truck friction task (α=0.1).

The population of students who answered with the common incorrect answer on the modified version of the task was small enough that the network formed from the responses was too sparse to interpret with any degree of confidence. As a result, the network is not included in this manuscript.

4.5.2.7 Discussion of results

The truck friction task, like the work-energy task, has two independent pathways for answering correctly, but the base physics question for the truck friction task has been shown to be much more difficult for students than the work-energy task. It was thus expected that the task would indicate the extent to which network methods may be applied productively to problems without a strong classroom consensus on the right way to answer a question and, in addition, provide insight into student reasoning surrounding a common student difficulty related to friction.

In general, the results affirmed that network analysis of student responses to chaining tasks may produce meaningful outcomes even in the presence of a common incorrect answer. The community detection algorithm produced distinct communities of elements tightly associated with each of the two predominant answers, namely the correct and the common incorrect answer choices, for both the base and modified version of the task. The results for the correct line of reasoning showed a drastic difference in communities between the base version and the modified version of the task. In the base version, the elements one would associate with the hypothetical line of reasoning were included along with only one element from the formal line of reasoning, reflecting the fact that the hypothetical line of reasoning was indeed predominant among the responses. In the modified version, the community structure included almost all of the elements relevant to the formal line of reasoning. While we cannot attribute this difference to a particular cause due to the two tasks being administered to different populations, it does seem plausible that the difference is due to the lowered accessibility of the hypothetical argument in the modified version (as a result of the absence of element 13). Additionally, even though students who selected the common incorrect answer used a variety of other elements in their response, the algorithm found that the tightest associations were between the three elements that are the foundation of the common incorrect argument. Taken together, we see these results as further evidence of the usefulness of community detection in determining the essential pieces of an argument in favor of a specific answer. We also suggest suspect that community detection is may be most effective when only one line of reasoning per answer is present incompatible with the elements provided.

Furthermore, if the difference in the correct community from the base to the modified version is found to be attributable to the lowered accessibility of the hypothetical argument in the modified version, it would be plausible to use chaining tasks to isolate specific lines of reasoning for detailed study. To support this, consider the sparsified "wisdom of the crowd" structure regarding the two correct arguments. The structure of the hypothetical argument is quite simple, while the structure of the formal argument is complex even if it has hints of linearity in it. Further study of these topological differences in a more controlled experimental design could yield insight into each line of reasoning.

In addition, network sparsification enabled us to examine the possible structure of a common difficulty with friction via network sparsification. Looking at the sparsification results from the base task, we observed that the same sub-set of elements are arranged differently to arrive at correct and incorrect answers. This result is reminiscent of the resource graphs discussed in Section 4.3.4 and hints at another possible avenue of future research using chaining tasks coupled with network analysis. The overlapping subset of elements may represent a shared set of resources among the two populations, with the elements "w/out friction, box would move left" and "friction opposes motion" having a different impact on how resources were coordinated. The high betweenness values of these elements in the correct and common incorrect networks (respectively) is consistent with this speculation. Furthermore, the element "the box moves to the right" (element 3) was tightly associated with the common incorrect answer. It could be that this element represents a resource which, combined with the shared subset of resources with an emphasis on the "friction opposes motion" element, produces the incorrect answer.

If reasoning elements do indeed stand in for the theoretical construct of "student resources" on some level, then it is within the realm of possibility that reasoning chain construction tasks can be utilized to study the structural coordination of student resources by fine tuning the elements to represent a known set of resources. At any rate, the results shown from this task do not represent progress in any theoretical direction but rather represent a phenomenological pattern worthy of further study, whatever theoretical framework one wishes to employ.

4.5.3 Two-Box Friction task

In this section, we present an in-depth network analysis of a chaining version of a task that was originally developed to study the extent to which dual-process theories of reasoning can explain and predict student behavior. This task, the two-box friction task, was originally the focus of an investigation reported in the literature in 2015 (Kryjevskaia, Stetzer, & Le, 2015). A separate paper (presented in Chapter 3 of this dissertation) by the authors of the current manuscript details how reasoning chain construction tasks can be utilized alongside dual-process theories of reasoning to gain greater insight into domain-general reasoning phenomena in physics and to draw upon the findings and theories of cognitive science to increase performance on this particular task. The task is included in this manuscript in order to highlight the findings from network analysis of student responses to this task, which are related to student reasoning more generally. Indeed, the results from this analysis suggest a possible avenue for further investigating cognitive phenomena, including dual-process reasoning, using chaining tasks coupled with network analysis.

The two-box friction task offers another opportunity to revisit two of the research questions related to the truck friction task. How effective are network analysis methodologies at characterizing and differentiating among different lines of reasoning on a physics question that is more challenging for students? To what extent can network analysis be used to identify and document evidence in support of specific theoretical constructs (e.g., dual-process theories of reasoning or resources) in reasoning chain construction task data?

4.5.3.1 Physics question overview

The two-box friction task is drawn from the literature (Kryjevskaia, Stetzer, & Le, 2015) and is part of a question pair expressly designed to study the impact of salient distracting features on student reasoning. In the two-box friction task, students are asked to compare the magnitudes of the friction forces on two identical boxes on different surfaces. Both boxes remain at rest while a 30 N tension force is applied. Coefficients of friction for each scenario are provided to the student in a diagram, shown in Figure 4-17. In order to arrive at a correct comparison, students must realize that the horizontal forces on the box (*i.e.,* the tension and the static friction) are balanced because the box remains at rest, from which they may conclude that the friction force exerted on both boxes is 30 N and the magnitude of friction on box A is therefore equal to the magnitude of the friction on box B. When asked in a multiple choice with explanation format (Kryjevskaia, Stetzer, & Le, 2015), 65% of students answered this way, while 35% of students answered incorrectly that since the coefficient between box A and the surface is less than the coefficient between box B and the surface, the magnitude of friction on A must also be less than the magnitude of friction on B.

Task statement:

Suppose the coefficient of static friction between box A and the floor is 0.4, as show 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?

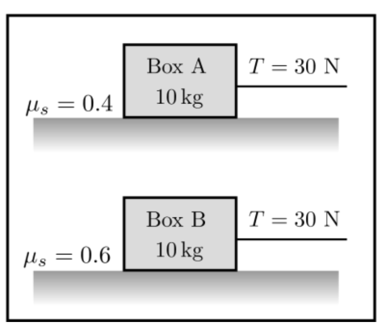


Figure 4-17. Two-box friction task prompt. Diagram given to students on the two-box friction task is replicated from Kryjevskaia, Stetzer, & Le, 2015)

4.5.3.2 Chaining task implementation

The reasoning elements provided to the student are shown in Figure 4-18. Each element included is true, and the students are told this in the prompt for the chaining task implementation of the two-box friction question. It is important to note, however, that some of these true elements are productive in common incorrect lines of reasoning, such as “the coefficient of friction for A is smaller than the coefficient of friction for B”.

The last two elements invite the student to compare the friction force to the applied force on each box, providing them with small attached text boxes in which they can insert a relationship such as “greater than”. The instructions in the prompt explained this option. The prompt also explained the subscript notation used in those elements. (Ultimately, the students did not end up using these customizable elements, so they are not represented in the networks we discuss below.)

The two-box friction task was preceded by a “screening” question in (Kryjevskaia, Stetzer, & Le, 2015), and this screening question was asked here as well in a multiple choice with explanation format. Results from the screening task will not be discussed in this manuscript.

1 $F_{net} = ma$	10 the vertical forces are balanced
2 both boxes have the same mass	11 the net force on both boxes is zero
3 the tension force on box A is equal to the tension force on box B	12 the friction force and the applied force are the only horizontal forces acting on the box
4 both boxes remain at rest	X Ffrct on A is [insert relationship here] Fapp on A
5 the coefficient of friction for A is smaller than the coefficient of friction for B	X Ffrct on B is [insert relationship here] Fapp on B
6 both boxes have the same weight	13 the magnitude of the friction force on box A is greater than the magnitude of the friction force on box B
7 the normal force on box A is equal to the normal force on box B	14 the magnitude of the friction force on box A is less than the magnitude of the friction force on box B
8 neither box is accelerating	15 the magnitude of the friction force on box A is equal to the magnitude of the friction force on box B
9 the horizontal forces are balanced	16 there is not enough information to tell

Figure 4-18. Elements provided to the student on the two-box friction task. The two elements labeled "X" were removed from the analysis as no student used them.

4.5.3.3 Performance overview

Of the 166 students who completed this task, 57% selected the correct answer and 40% selected the common incorrect answer; the performance on the chaining format of this task was generally consistent with previously reported findings (Kryjevskaia, Stetzer, & Le, 2015).

4.5.3.4 Arguments Found via Community Detection

Figure 4-19 shows a representation of the communities identified in an indirect association graph comprised of all responses to the two-box friction task.

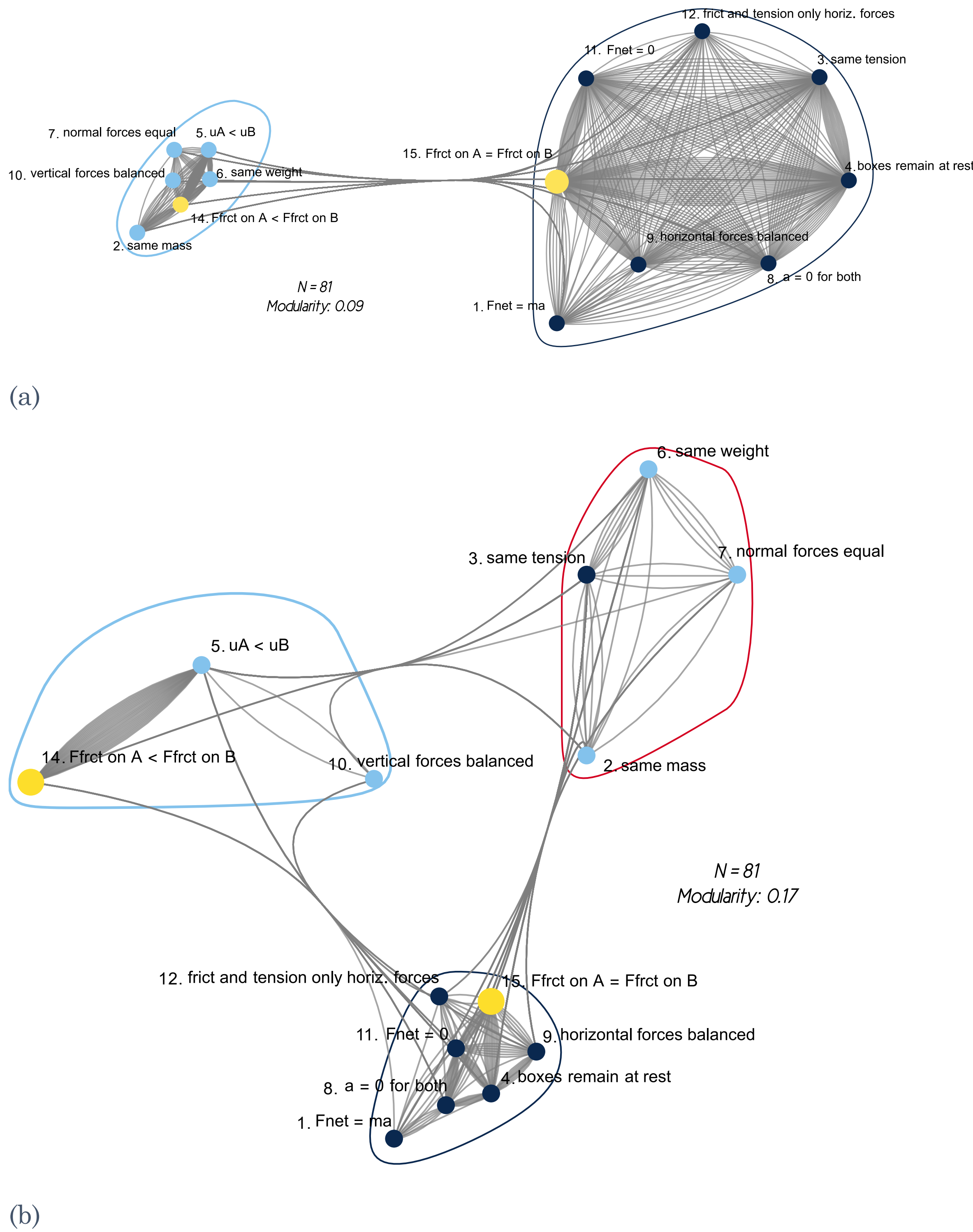


Figure 4-19. A representation of the communities identified in (a) an indirect association network and (b) a direct association network comprised of all responses to the two-box friction task.

Again, the algorithm produces a meaningful separation between the common incorrect and the correct line of reasoning. A frequency plot (shown in Figure 4-20) generated by the method of bootstrapping explained in Section 4.4.2.2 indicates that the community structure is fairly robust. In the plot, the dark blue markers indicate the community that includes the correct answer element, while the light blue markers indicate the community that includes the common incorrect answer element. The elements "the normal force on box A is equal to the normal force on box B" (element 7), "neither box is accelerating" (element 8), and "the friction force and the applied force are the only horizontal forces acting on the box" (element 12) appear to be somewhat shared between the two communities, but all elements in each community structure shown in Figure 4-19 are above a 60% threshold for their respective community, and below 30% for the opposite community.

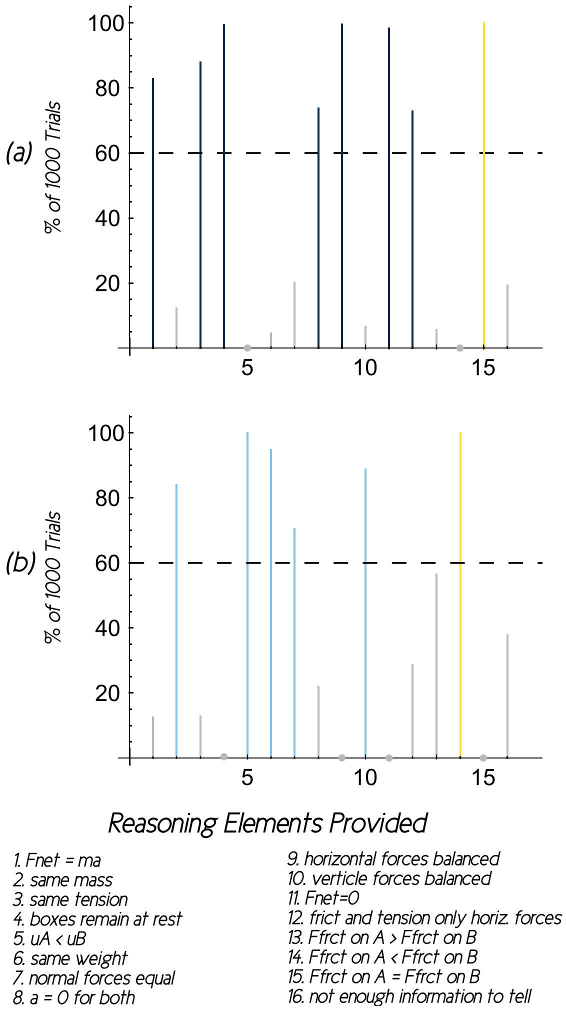


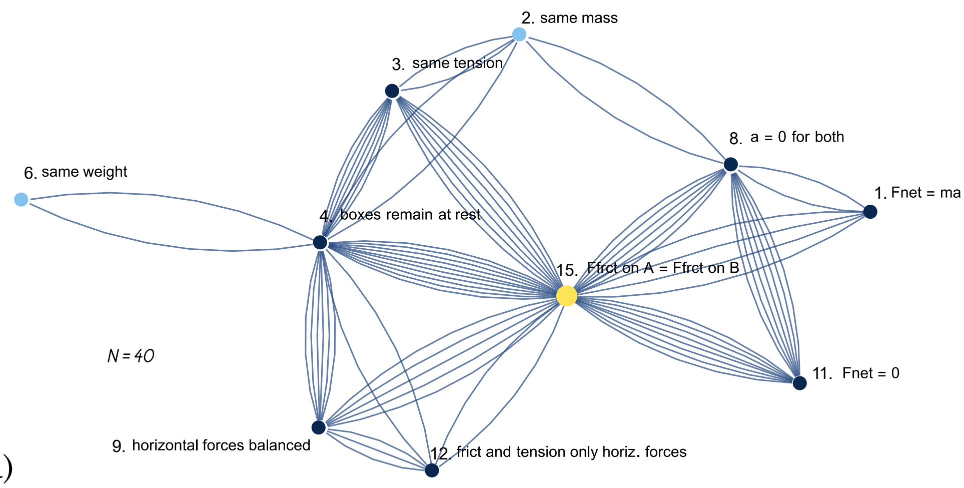
Figure 4-20. A frequency plot of the communities identified in the indirect association network generated by the method of bootstrapping explained in Section 4.4.2.2. The figure shows a test for the community that includes (a) the correct answer element, and (b) the common incorrect answer element. The elements are color coded according to the coloring presented in the community plot shown Figure 4-19.

The community structure of the direct association graph shows a similarly meaningful separation between correct and common incorrect responses, but the graph separated a collection of elements with a similar theme -- namely, the four elements that explicitly state that something is "equal" or "the same". These elements were "both boxes have the same mass" (element 2), "both boxes have the same weight" (element 6), "the normal force on box A is equal to the normal force on box B" (element 7), and "the tension force on box A is equal to the tension force on box B" (element 3).

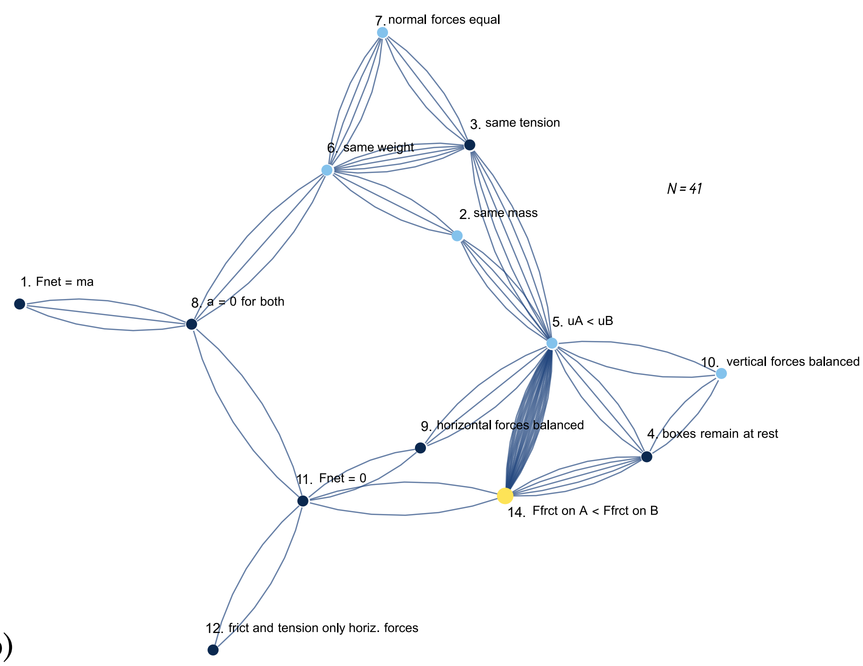
This "sameness" community in the direct association network (Figure 4-19.b) fails a bootstrapping test for the exact community structure shown (success rate of 10%). However, a frequency plot of the elements most often in a community with the "tension is the same" element (not shown) suggests that the elements "same weight" and "normal forces equal" are tightly connected to the tension element and are the only elements above the 60% threshold for robustness (75% and 65%, respectively). Additionally, a bootstrapping test for the presence of those two elements in the community that includes "tension is the same" has a 65% success rate. We conclude that the "sameness" community is moderately robust; it is clearly present but it is fragile to small perturbations in the network structure. The other communities in the direct association network are highly robust with the exception of the element "both boxes have the same mass" (element 2) which is shared among all three communities shown.

4.5.3.5 Topology of Argument Structure via Sparsification

Separating the responses based on answer element used, two direct association networks were sparsified resulting in the networks shown in Figure 4-21.



(a)



(b)

Figure 4-21. Sparsification of direct association networks comprised of (a) correct responses and (b) responses with the common incorrect answer. $\alpha = 0.2$ for both.

The sparsification of the direct association network comprised of responses containing the common incorrect answer element reveals a somewhat linear topology, while the topology of the correct responses mimics a "wheel graph", wherein the central node (in this case the answer element) is connected to every node on a ring of nodes surrounding it. The global clustering coefficients of the unsparsified graphs are 0.48 and 0.6 respectively, which reinforces an interpretation of the incorrect reasoning topology as being more "linear" than the correct reasoning topology.

The elements with the highest betweenness in the unsparsified network of incorrect responses are "the coefficient of friction for A is less than the coefficient of friction for B" (59.3), "the boxes remain at rest" (48.8), and "both boxes have the same weight" (35.5). For the correct responses, the highest betweenness was the element "the boxes remain at rest" (51.2) and the answer tile (40.7).

4.5.3.6 Discussion of results

The network analysis results from the two-box friction task reinforce that community detection, especially of indirect association graphs, gives meaningful information about the elements associated with the lines of reasoning leading to different answers. The elements for the correct and the common incorrect chains were well separated, as indicated by the bootstrapping frequency plot, and were interpretable. Additionally, the direct association graph gave greater resolution on the core of the common incorrect argument: students apparently base their reasoning on the difference between the coefficients of static friction for each scenario. This is, of course, expected as the task was expressly designed to elicit reasoning based on the coefficients; network analysis of chaining tasks has, then, an ability to detect reasoning effects related to salient distracting features.

The sparsification process likewise yielded meaningful differences in the topologies of the two types of reasoning. The "wheel"-like structure of the correct network indicates that there isn't a strong consensus as to the ordering of the specific argument or, more particular to this structure, a strong consensus as to what elements need to be included in an argument supporting the correct answer. Looking at the student chains, it appears that, to the students, there are many ways of saying the same thing. Sparsification of the common-incorrect answer network, on the other hand, showed a strong consensus about the core of the argument, which was comprised of a tight association between the element comparing the coefficients and answer element. The other elements in the common-incorrect network seem somewhat peripheral, but the linear structure also indicates a consensus in how these elements are arranged into arguments.

Based on the known nature of this task as eliciting strong intuitive responses formed around the coefficients (Kryjevskaia, Stetzer, & Le, 2015), it may be that the topology of the common incorrect line of reasoning is indicative of a strong cueing on the coefficients. This view would be consistent with a dual-process theory perspective. From this perspective, students who have a strong intuitive (process 1) answer may attempt to rationalize that answer using formal physics knowledge during a superficial engagement of process 2, but this rationalization will always be post-hoc due to process 1 having already formed a conclusion. Recalling that the sparsified network represents the classroom consensus of the "logical landscape" that the elements create, we interpret the strong association between the element comparing the coefficients and the answer element as mimicking the association formed by process 1 between the cue and the judgement proceeding from that cue. The weaker associations among the elements which would add further justification and detail for the coefficient argument would then be indicative of process 2 having been only superficially engaged, if at all, by the population as a whole. In this perspective, the "wheel"-like nature of the network representing the correct reasoning may be related to a more comprehensive understanding of the elements – each element in the line of reasoning is associated with the answer on some level such that everything is deemed to be relevant to that answer. Further research would be required to bring this speculation into a measurable domain, but this does highlight the possibility that theoretical frameworks, such as dual-process theories, can be explored using chaining tasks coupled with network analysis.

The sparsification results yield the additional insight that the element "both boxes remain at rest" appears to be used in both networks. It could be, then, that the recognition that both boxes remain at rest is not sufficient to cue reasoning related to balanced forces for students who ultimately select the common incorrect answer. Perhaps the difference between students who select the correct answer and students who select the common incorrect answer is a cognitive connection between the cue "both boxes remain at rest" and "the horizontal forces are balanced" (which is prominent in the correct reasoning network but largely absent in the common incorrect network). Attending to that connection during instruction may improve performance on this question. If this hypothesis is eventually confirmed, then reasoning chain construction tasks may be useful in revealing specific portions of arguments in which reasoning chains can be reinforced during instruction.

The betweenness results also help provide insight into what, exactly, betweenness may be measuring in a reasoning chain network. The elements with the highest betweenness for the correct and common incorrect answer networks are cues from the problem statement that we would expect would be indicative of the respective answer ("boxes at rest" for the correct answer, and coefficients for the common incorrect). This result leads to a proposed interpretation of betweenness in reasoning chain networks. Given that betweenness is aimed at measuring the control of the flow of information through a network, the betweenness in a reasoning chain network may be measuring the central idea in students' reasoning; that is, the idea the students "lock on to" in order to frame their reasoning.

Finally, the community detection of the direct association network found that four elements had tighter association with each other than with the rest of the network. These four elements were "both boxes have the same mass" (element 2), "both boxes have the same weight" (element 6), "the normal force on box A is equal to the normal force on box B" (element 7), and "the tension force on box A is equal to the tension force on box B" (element 3). There is an *a priori* reason to believe the first three elements would be associated with each other, namely that there is a direct connection between weight and mass (they are proportional) and because of the direct connection between the normal force and the weight (they are equal in this case and it is common for a student to write $N = mg$ regardless of the situation). But the presence of the "tension" element led us to wonder if there was an underlying reason that these elements would be connected, especially as the tension element is not very useful in an incorrect chain of reasoning. Students may have a desire to express a thought related to the tension in that it does not "overcome" the friction force, but this idea is not represented by this particular element. The element instead simply compares the tension on A to the tension on B.

It could be that this community represents an unconscious tendency to associate things that are the same with one other, similar to a "same is same" p-prim (diSessa, 1993). The mediocre robustness of the community of “same/equal” elements is consistent with an unconscious tendency to associate similar elements because we would expect these unconscious effects to be hard to discern (Gawronski & Payne, 2010). However, the methods and results described here are far from able to assess such an effect, and this proposal is mentioned to illustrate a possible future use of chaining tasks and network analysis for research.

4.5.4 Isomorphic Graph Tasks

In this section, we report on student reasoning on a collection of four similar tasks administered over the course of two subsequent semesters of introductory calculus-based physics. Each of the four tasks is designed to foreground the same line of reasoning in four different contexts. By conducting this experiment, we sought to answer the following research question. To what extent can network analysis methodologies be used in conjunction with reasoning chain construction tasks to track and document the development of a specific line of reasoning over the course of a two-semester introductory physics sequence? Network analysis of these tasks provided evidence for the development of a skill and comfort with this line of reasoning over the course of instruction.

4.5.4.1 Physics question overview

As part of an investigation of the impact of salient distracting features on patterns of student reasoning in the context of introductory physics, we developed four chaining-format graph tasks that are isomorphic in structure and are based upon one task in the literature, which we refer to as the kinematics graph task (Heckler, 2011; McDermott, Rosenquist, & Zee, 1987; Beichner, 1994; Elby, 2000; see also Speirs, Ferm Jr., Stetzer, & Lindsey, 2016).

In the kinematics graph task, shown in Figure 4-22, students are asked to determine when the speeds of two cars are the same by examining a plot of position vs. time with two graphs representing the motion of the two cars. At time A, the slopes of the two graphs are the same, and at time B the two graphs intersect. The correct answer is arrived at by noting that the velocity is the time-derivative of position, which on a graph equates to the slope of the tangent line at a point. Comparing slopes allows students to determine that the speeds (*i.e.,* the magnitudes of the velocities) are the same at time A. However, it is observed in the literature (Heckler, 2011) that many students answer that the speeds are the same at time B, consistent with attending to the intersection point of the two graphs. The phenomenon of incorrect answering on these types of graphs has led to researchers investigating "slope-height confusion" and other difficulties related to interpreting and using graphs in a physics context (McDermott, Rosenquist, and Zee 1987; Beichner, 1994; Christensen & Thompson, 2012), and has also been used to examine the impact of salient distracting features in physics contexts (Heckler, 2011; Speirs, Ferm Jr., Stetzer, & Lindsey, 2016).

The other three tasks are presented, in detail, in Appendix A. All four tasks are structurally parallel and presented in the contexts of kinematics, potential energy, electric potential, and magnetic flux. Each graph task has a correct line of reasoning that relies on an understanding that the desired quantity can be obtained from the derivative of the known quantity, and thus the slopes of the graphs at the point of interest ought to be compared.

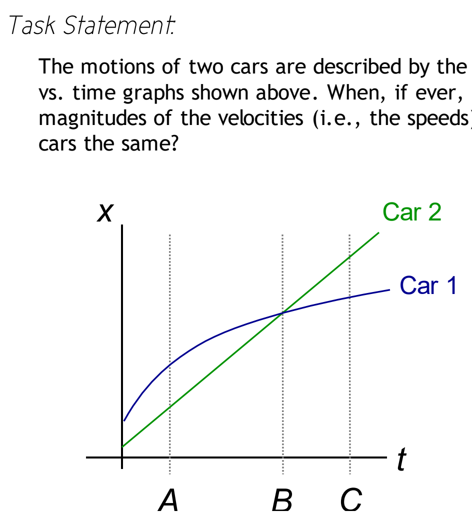


Figure 4-22. The first of four isomorphic graph tasks adapted from (Heckler 2011). The other three graph tasks are shown in detail in Appendix A.

4.5.4.2 Chaining task implementation

The reasoning elements provided to the student in each task have been modified to fit the context but remain isomorphic in their structure. The reasoning elements are shown in Figure 4-23. Unlike the tasks discussed in previous sections, these isomorphic tasks include a large number of elements that are irrelevant to both the correct and common incorrect lines of reasoning; indeed, seven of the twelve elements are not relevant to any common line of reasoning.

	Kinematics Reasoning Elements	Potential Energy Reasoning Elements	Electric Potential Reasoning Elements	Magnetic Flux Reasoning Elements
1	$x(t_f) = x_0 + \int_0^{t_f} v(t)dt$	$U(x_f) = U_0 + \int_0^{x_f} \vec{F}(x) \cdot d\vec{x}$	$V(x_f) = V_0 + \int_0^{x_f} E(x)dx$	$\Phi_B = -\int_0^{t_f} \mathcal{E}dt$
2	$v(t_f) = v_0 + \int_0^{t_f} a(t)dt$	$p(t_f) = p_0 + \int_0^{t_f} F(t)dt$	$U(x_f) = U_0 + \int_0^{x_f} F(x)dx$	$\mathcal{E} = -\int_0^{s_f} \vec{E} \cdot d\vec{s}$
3	$v = \frac{dx}{dt}$	$F = -\frac{dU}{dx}$	$E = -\frac{dV}{dx}$	$\mathcal{E} = -\frac{d\Phi_B}{dt}$
4	$a = \frac{dv}{dt}$	$F = \frac{dp}{dt}$	$F = -\frac{dU}{dx}$	$E = -\frac{d\mathcal{E}}{ds}$
5	the integral,$\int_a^b f(x)dx$, is the area under the graph of *f* vs. *x*	the integral,$\int_a^b f(x)dx$, is the area under the graph of *f* vs. *x*	the integral,$\int_a^b f(x)dx$, is the area under the graph of *f* vs. *x*	the integral,$\int_a^b f(x)dx$, is the area under the graph of *f* vs. *x*
6	the derivative,df/dx, at a specific point is the slope of the tangent line of the *f* vs. *x* graph at that point	the derivative,df/dx, at a specific point is the slope of the tangent line of the *f* vs. *x* graph at that point	the derivative,df/dx, at a specific point is the slope of the tangent line of the *f* vs. *x* graph at that point	the derivative,df/dx, at a specific point is the slope of the tangent line of the *f* vs. *x* graph at that point
7	slope of a position vs. time graph is the velocity	the negative of the slope of a potential energy vs. position graph is the force	the slope of an electric *potential* vs. position graph gives the magnitude of the electric field	slope of a magnetic flux vs. time graph is the magnitude of the induced EMF
8	slope of a velocity vs. time graph is the acceleration	slope of a momentum vs. time graph is the force	the slope of an electric *potential energy* vs. position graph is the force	slope of an induced EMF vs. position graph is the magnitude of the induced electric field
9	area under a velocity vs. time graph is the displacement	area under a force vs. position graph is the work done by the force, which is the negative of the change in potential energy	area under an electric field vs. position graph is the electric potential	area under an induced EMF vs. time graph is the magnitude of the change in magnetic flux
10	area under an acceleration vs. time graph is the change in velocity	area under a force vs. time graph is the change in momentum (or the impulse)	area under a force vs. position graph is the electric potential energy	area under an induced electric field vs. position graph is the magnitude of the change in the induced EMF
11	the lines intersect at time B	the lines intersect at position B	the lines intersect at position B	the lines intersect at time B
12	slopes are the same at time A	slopes are the same at position A	slopes are the same at position A	slopes are the same at time A
13	the speeds are the same at time A	the magnitudes of the forces are the same at position A	the magnitudes of the electric fields are the same at position A	the magnitudes of the induced EMF's are the same at time A
14	the speeds are the same at time B	the magnitudes of the forces are the same at position B	the magnitudes of the electric fields are the same at position B	the magnitudes of the induced EMF's are the same at time B
15	the speeds are the same at time C	the magnitudes of the forces are the same at position C	the magnitudes of the electric fields are the same at position C	the magnitudes of the induced EMF's are the same at time C
16	the speeds are never the same	the magnitudes of the forces are never the same	the magnitudes of the electric fields are never the same	the magnitudes of the induced EMF's are never the same

Figure 4-23. Reasoning elements provided to the student on each of the four isomorphic graph tasks.

There is an inherent logical structure among the productive elements provided to the students (shown in red in Figure 4-23). While, at first glance, it may appear that the elements "$v = dx/dt$", "*the derivative,* $dh(r)/dr$*, at a specific point is the slope of the tangent line of the h(r) vs. r graph at that point*", and "*velocity is given by the value of the slope of a position vs. time graph*" are equivalent and interchangeable statements, they actually constitute a logical argument justifying why the slope is the velocity; namely, the two elements "$v = dx/dt$" and "*the derivative[…] is the slope…*" combine to imply the third element. We refer to the collection of these three elements as the *velocity triad*. We also refer to the element "*velocity is given by the value of the slope of a position vs. time graph*" as a *derived heuristic* because it represents a chunked knowledge piece (National Research Council, 2000) that is derived from two independent principles. While it would be acceptable to many instructors if students were to simply use the "slope is velocity" heuristic, all three elements are needed to provide a logically sound argument. Their inclusion, then, provided an opportunity for additional insight into whether students tend to justify their arguments with first principles or instead rely on derived heuristics learned in class.

4.5.4.3 Performance overview

All tasks were administered after relevant course instruction. Chronologically, the kinematics task was administered first in the year, the potential energy task second, the electric potential task early in the second semester of physics, and the magnetic flux task last. Given the contexts associated with these isomorphic tasks, data were collected in both semesters (fall and spring) of the on-sequence calculus-based introductory physics course. Because the four graph tasks were administered across a single academic year, most students who completed the introductory calculus-based sequence would have seen and completed multiple, and likely all four, tasks.

Student performance for these tasks is shown in Table 4-4. The percentage of responses answering correctly increases very slightly over the two-course sequence, but it can be seen that salient distracting feature (the intersection point) remains a strong distractor, with more than a quarter of students answering consistent with attending to the intersection point.

Response	Kinematics (N = 149)	Potential Energy (N = 76)	Electric Potential (N = 97)	Magnetic Flux (N = 88)
Time A	**57%**	**43%**	**73%**	**66%**
Time B	29%	51%	21%	28%
Time C	0%	1%	1%	5%
Never	14%	4%	5%	1%

Table 4-4. Overview of student performance on the four isomorphic graph tasks.

4.5.4.4 Arguments Identified via Community Detection

Each indirect association network (not shown) built from responses to the graph tasks generally breaks into three communities: the correct answer community, which includes the elements isomorphic to "$v = dx/dt$", "velocity is given by the value of the slope of a position vs. time graph", and "the slopes are the same at time A"; the common incorrect answer community, which includes the element isomorphic to "the lines intersect at time B"; and a third community including all of the other elements in a loosely connected network. These elements were not relevant to any common line of reasoning. Interestingly, the element "the derivative, $dh(r)/dr$, at a specific point is the slope of the tangent line of the of the *h(r) vs. r* graph at that point" (element 6), which is very relevant to the correct line of reasoning, was found in the common incorrect answer community for the kinematics and potential energy graph task, but was found in the correct answer community in the electric potential and magnetic flux task. We would have expected this element to always be associated with the correct answer. To investigate this phenomenon more fully, we examined community structure in indirect association networks comprised of just the correct responses to each task. The resulting networks are shown in Figure 4-24. The elements that make up the full, detailed correct line of reasoning are colored red in the figure, while all other elements are colored dark blue except the answer element, which is colored yellow. One can notice that the *derivative is slope* element is not in the main answer community for the first two graph tasks but becomes more tightly associated with the correct answer in the final two graph tasks.

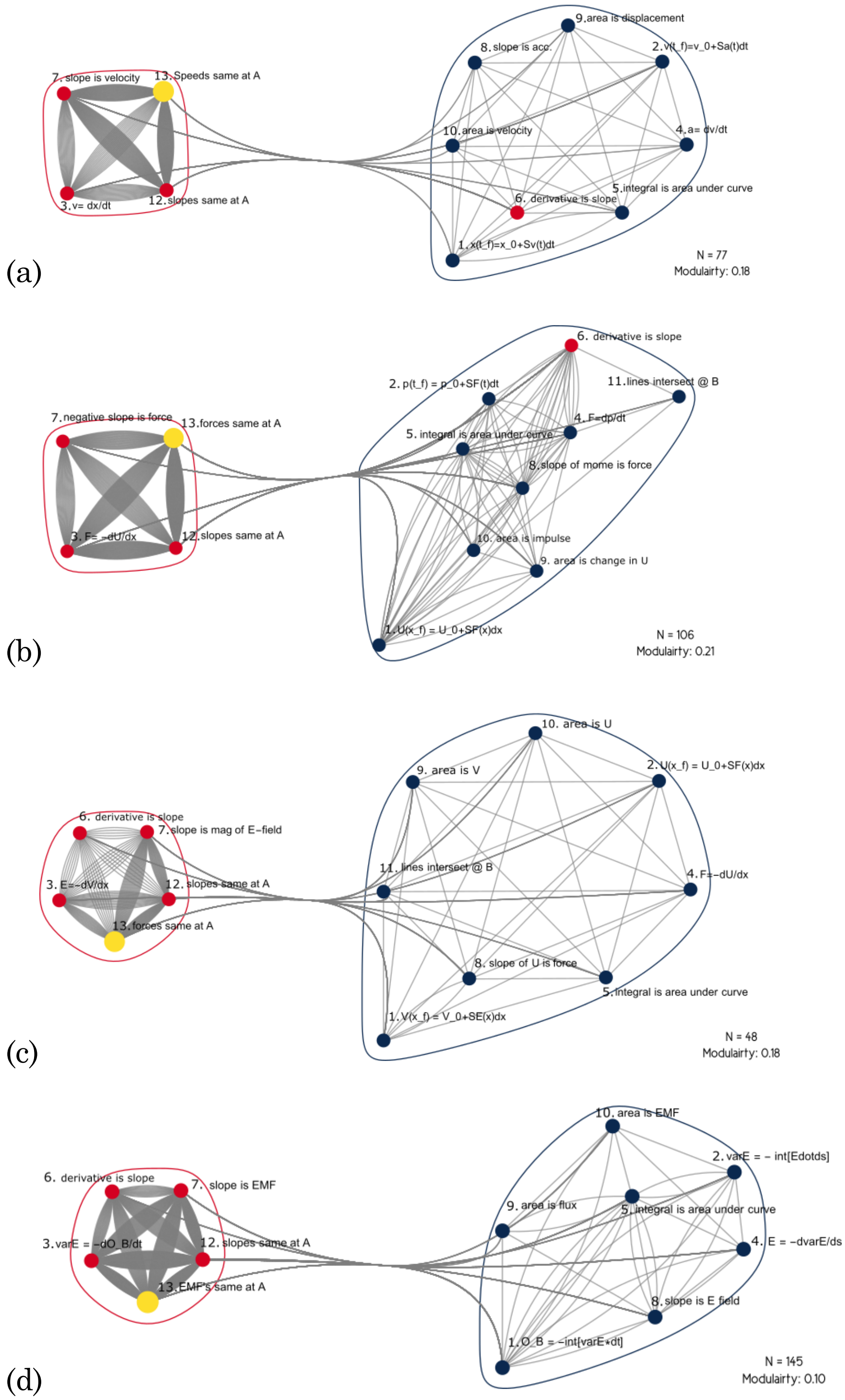


Figure 4-24. Community structure detected in indirect association networks comprised of correct responses to the graph task as posed in the context of (a) kinematics, (b) potential energy, (c) electric potential, and (d) magnetic flux.

A bootstrapping frequency plot for the correct community, shown in tabular form in

Table 4-5 for ease of reading, revealed that the derivative is slope element is indeed increasing in use across the four tasks (administered in the sequence shown), and thus increasing over the course of the two-semester introductory calculus-based physics sequence.

	Kinematics	Potential Energy	Electric Potential	Magnetic Flux
derivative is slope	46%	29%	74%	100%
"v=dx/dt"	85%	100%	95%	100%
slope is "velocity"	100%	100%	100%	100%
slopes same at A	100%	100%	100%	100%

Table 4-5. The results of a bootstrapping frequency plot in tabular form for the correct answer community. Results are shown in table form rather than a plot for ease of reading. Elements referencing velocity are in quotes to remind that in the non-kinematics graph tasks, this element was cast into the appropriate context.

The community structures of the direct association networks for the four graph tasks (not shown) also reveal a shift in how the derivative is slope element is used by students. In the responses to the kinematics and potential energy tasks the element is not a member of the correct answer community or in the same community as the other productive elements, whereas in the responses to the electric potential and magnetic flux tasks the element is more closely associated with the productive elements. A particularly compelling community structure is found in the direct association network built from correct responses to the magnetic flux task and is therefore shown in Figure 4-26. The community structure shows a subcommunity made up of the “velocity triad” elements. Recalling that, in direct association networks, a connection is formed between two elements when they are placed consecutively, the sub-community of the “velocity triad” elements means that those three elements were consistently placed next to each other in student responses.

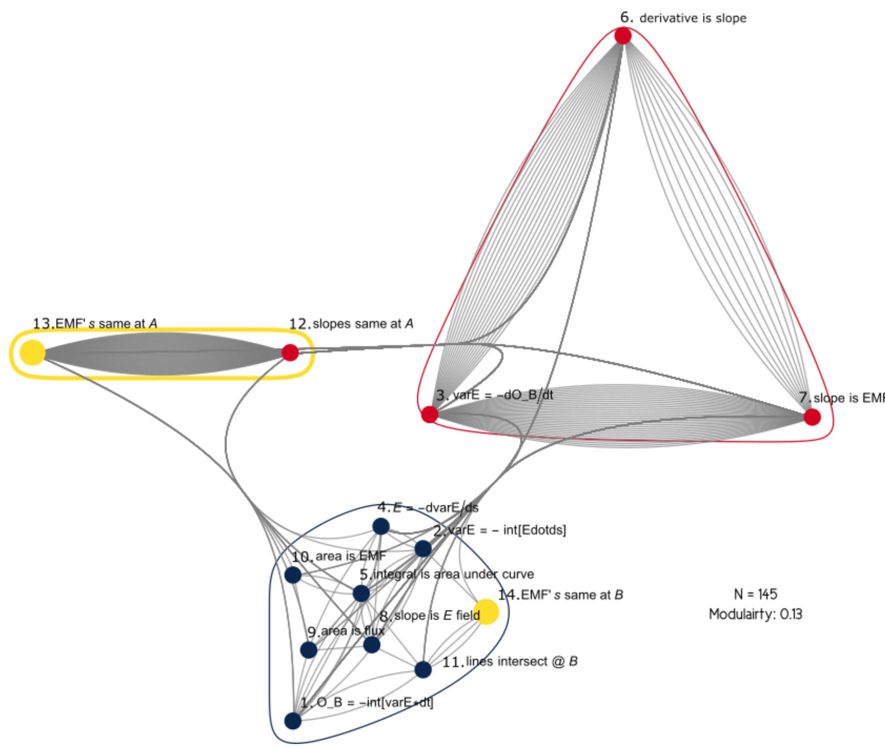
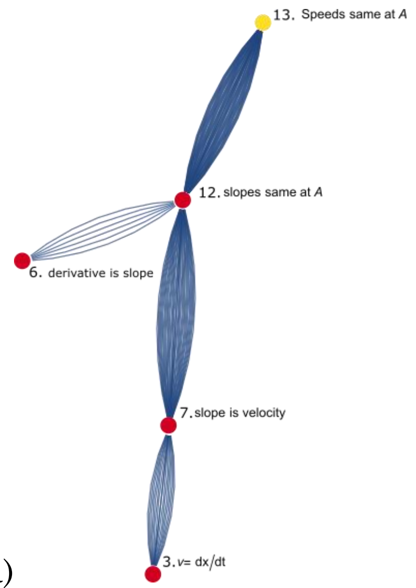


Figure 4-25. Community structure found in the direct association network comprised of correct responses to the magnetic flux graph task. The sub-community of the “velocity triad” elements means that those three elements were consistently placed next to each other in student responses.

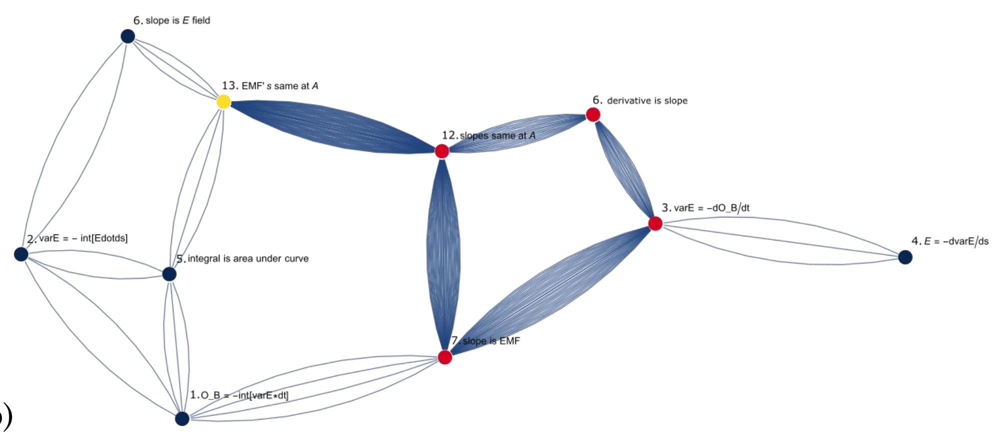
4.5.4.5 Topology of Argument Structure via Sparsification

The basic result that the derivative is slope element becomes more integrated into the correct line of reasoning is also revealed in the sparsification of the direct association networks. For space, we only show the sparsified correct answer networks for the kinematics and magnetic flux tasks (see Figure 4-26).

The sparsified network of correct responses to the kinematics task, shown in Fig. 4-26a appears to be a linear path from v = dx/dt through the derived heuristic “slope is velocity” (element 7) to the answer. The “derivative is slope” element constitutes an extension of this, another independent piece of information that must be brought in to secure the logic of the argument. The sparsified network of correct responses to the potential energy task (not shown), however, reveals that the *derivative is slope* element is heavily connected to the unproductive element "slope of momentum is force" and, as in the kinematics task, is somewhat connected to “the slopes are the same at position A”. In the sparsified correct answer network for the electric potential task (also not shown), the *derivative is slope* element is placed into the main line of reasoning, which consists of “derivative is slope”, “$E = -\frac{dV}{dx}$”, “slopes same at position A”, and then the correct answer element; it is no longer peripherially attached to the main line of reasoning as in the previous two tasks. However, it is still only somewhat connected to that chain of elements. Finally, in the sparsified network of correct responses to the magnetic flux task shown in Fig. 26b, the *derivative is slope* element serves as a bridge between the elements "$\varepsilon = -\frac{d\Phi_B}{dt}$" (element 3) and "the slopes are the same at time A" (element 12) and is heavily connected to both of those elements. An examination of directed networks showed that element 3 ("$v = dx/dt$" for the kinematics task) is a common starting point for the correct responses to all tasks.



(a)



(b)

Figure 4-26. Sparsified direct association networks comprised of correct responses to the (a) kinematics graph task ($\alpha = 0.1$), and (b) magnetic flux task ($\alpha = 0.1$).

While the sparsified networks for the correct responses appear to become less linear over the four tasks, the global clustering coefficients for the correct response networks for the four tasks range from 0.61 to 0.75, indicating fairly clustered networks throughout. Even so, the magnetic flux task has a higher clustering coefficient (0.75) than the kinematics task (0.61), which does suggest some increase in clustering.

The betweenness centrality of the elements in the correct response network for each of the four tasks is of interest. Normalized betweenness centrality calculations for the three elements that comprise the "velocity" triad are shown in Table 4-6. As can be seen, one of the elements (the derivative is slope element) had a betweenness of zero, whereas in the later tasks, all elements had non-zero betweenness. Additionally, the average betweenness of the velocity triad elements increases across the four tasks.

Isomorph Element Abbreviation	Kinematics	Potential Energy	Electric Potential	Magnetic Flux
"v = dx/dt"	0.82	1.00	0.87	1.00
derivative is slope	0.00	0.22	0.77	0.28
slope is "velocity"	1.00	0.50	0.38	0.99
Average	0.61	0.57	0.68	0.75

Table 4-6. Normalized weighted betweenness centrality (Opsahl, Agneessens, & Skvoretz, 2010) calculations for the unsparsified network comprised of correct responses for each graph task. The element label is shown in the kinematics context but is meant to be general to all contexts.

As might be expected, sparsification of the common incorrect answer networks show strong associations between “the lines intersect at (time/position) B” and the corresponding incorrect answer element. In the kinematics context, the sparsified incorrect answer network also revealed a tendency to associate the elements “velocity is slope” and “$v = dx/dt$” with the intersection element and the answer element, but this was the only context to do so. Because of there being very few elements (an average of 2.5 elements per chain) used in the incorrect responses, the sparsified networks appeared linear, and the clustering coefficients for the unsparsified incorrect response networks indicated linear structure with coefficients ranging from 0.17 to 0.48. The typically low number of elements per chain combined with the lower number of students selecting the incorrect answer rendered the betweenness centrality measure uninterpretable for the later tasks, so betweenness centrality is not reported for the incorrect answer networks.

4.5.4.6 Discussion of results

The results of network analysis of the four isomorphic graph tasks again demonstrate that community detection can meaningfully separate lines of reasoning in the responses according to the answer choice. Additionally, the wisdom of the crowd results from sparsifying the direct association networks reveal meaningful structures in the responses such as the relative (compared to the other tasks) linearity of the correct line of reasoning in the kinematics task or the tight association between the cue (lines intersect at B) and the common incorrect answer. Thus, the key result that network analysis of chaining task data provides useful and interpretable information is replicated in this task.

Perhaps the most important result from the isomorphic graph tasks is the observed development of a cohesive line of reasoning regarding the “velocity triad” of elements, seen in both the community detections and network sparsifications. The identified communities in both the direct and indirect association networks indicate that the *derivative is slope* element was not tightly associated with the other productive elements (including the correct answer element) for the mechanics tasks but was tightly associated with those elements for the electromagnetics tasks.

For an indirect association network, membership in the correct answer community alongside the other productive elements implies that the *derivative is slope* element either increases in frequency of use in correct responses overall or compared to responses that also include unproductive elements. The proportion of correct responses that include the "derivative is slope" element is 14% for the kinematics task, 24% for potential energy task, 24% for electric potential task, and 27% for magnetic flux task, indicating that the frequency of use overall is not increasing much over the last three tasks. Instead, the element must have been more frequently placed in responses that include only the other productive elements, rather than being placed in responses that include unproductive elements as well – that is, the element is being used “more productively”.

The fact that the derivative is slope element joins the community of the other productive elements in the direct association networks is also indicative of associating that element with productive rather than unproductive elements as the introductory physics course sequence progressed. This is because the connections in a direct association network are formed based on an element’s proximity to other elements. Thus, if two elements are tightly tied together, they are more often used in proximity to each other and are thus more associated. The observation that the derivative is slope element gains membership in communities with the other productive elements then implies that the element was used in closer proximity to the other productive elements, which supports the interpretation that the element was used more productively over time.

This interpretation is further bolstered with the results from sparsification. There, the element starts out as peripheral to the “classroom consensus” on the correct line of reasoning but progresses to become more central to that line. Thus, the coherence between the derivative is slope element and the other productive elements increases. If betweenness centrality indeed stands as a proxy for core ideas in a reasoning chain, as suggested by the results from other tasks, the increasing betweenness centrality of the “triad” elements would be further evidence in support of the development of a coherent chain of reasoning.

We propose that, as the sequence progresses, the students in these tasks either better understand the connection between that element and the other elements or are more comfortable with the use of that element alongside the other elements.

Why would this shift occur?

One explanation for the relative non-use of the element among correct respondents on the kinematics graph task is that the phrase "the velocity is the slope" is often a "chunked" cognitive element or heuristic, even among experts[3]. We presume that the students who answer correctly on this task in the context of kinematics employ the learned heuristic that the slope of a position versus time graph is the velocity and ignore the first principles from which that heuristic is derived. When asked the question in a context in which they haven't formed such a heuristic, they may then resort to a wider examination of the separate elements.

The heuristic may have been formed to varying degrees in the other contexts. For instance, in the magnetic flux task, it may be that student were less familiar with the application of Faraday's law to a graph of magnetic flux

[3]We have administered the chaining version of the kinematics graph task to physics and other STEM educators and a frequent comment we hear is that the three elements "v = dx/dt", "derivative is slope", and "velocity is slope" are functionally equivalent. Only when it is pointed out that the former two are independent statements that combine to justify the latter is it agreed upon that the three elements are actually logically different.

than they were with, say, how to get an electric field from a graph of electric potential. Because of a lack of familiarity, students may have relied more on the calculus to make a connection between Faraday's law and the graph, as opposed to simply knowing from the features of the graph how to obtain an answer. This is supported by a brief review of the curriculum. In the course textbook (Knight, 2016), there are many examples of switching between field and potential graphically, but most examples concerning Faraday's law were centered in non-graphical considerations. Thus, the heuristic was probably more familiar in the electric potential task than it was in the magnetic flux task, with both being less familiar than the kinematics task.

Another possibility is that the students, over the course of the two semesters, became more comfortable and/or more proficient with the language and concepts of calculus, such that they felt comfortable endorsing elements that explicitly included those concepts. Some of the think-aloud interviews conducted with students seemed to support this interpretation as well, at least in the aspect of student’s not feeling comfortable with the language on the kinematics task. Further work would need to be done to determine the extent to which comfort with calculus impacts the use of the derivative is slope element, but this is a very real possibility to consider, as a significant percentage of students were concurrently taking the first calculus course as a co-requisite at the time the kinematics task was administered and derivatives were covered later in the semester in calculus.

While the cause of the shift can't be ascertained from our data alone, the evidence of a shift points to the usefulness of the network analysis of chaining tasks to examining student formation of specific reasoning chains. We see a shift across multiple metrics, including community detection on indirect association networks, community detection and sparsification of direct association networks, and betweenness calculations for direct association networks. Thus, network analysis techniques are sensitive to shifts in reasoning chains over time and, as such, could be used to gauge how students are building reasoning skills over time.

Finally, the results regarding the incorrect answer networks revealed a tight association between the element "the lines intersect at (time/position) B" and the incorrect answer element. This tight association is reminiscent of the association between the element comparing the coefficients and the common incorrect answer in the two-box friction task and may be due to a similar phenomenon. The intersection point in each graph task is a salient distracting feature and commands attention. Thus, the tight association between the intersection element and the answer element could be related to a tight link between a perceptual cue and a process 1 judgement based on that cue. Recalling that the correct line of reasoning was less linear than the common incorrect line of reasoning in both the two-box friction task and the graph tasks, there could also be a relationship between the cue-judgement phenomenon and the linearity of the networks. However, it may also be that those wishing to respond with the "time B" answer simply had no other elements they could use to describe their reasoning, which would create both the tight association and the linearity of the network. Further investigations would be necessary to examine the extent to which the observed phenomenon documented through network analysis primarily stems from underlying cognitive mechanisms or features of the particular task discussed here.

4.6 Conclusions and future work

The overarching goal of this manuscript was to illustrate how a new methodology, network analysis of student responses to *reasoning chain construction tasks,* can generate valuable knowledge surrounding how students reason on physics questions, specifically those questions that require stepping through a series of qualitative inferences. As we have shown, network analysis of responses to chaining tasks generates novel data sources related to both the content and structure of student arguments. Here, we discuss general affordances seen across tasks, and then highlight how these affordances, and other patterns observed in the data, can be used to bolster existing analysis methods or generate entirely new research questions.

Across all tasks, we have demonstrated that network analysis of chaining task data has the ability to separate lines of reasoning associated with a particular answer. Via community detection, we were consistently able to find elements that were more tightly associated with a given answer than the other elements in the set; these tight associations were interpretable as typical reasoning seen from students in free response / interview settings. One affordance of the network methodology is that the categorization of the elements into lines of reasoning associated with a particular answer is automatic through the use of the community detection algorithm, so large data sets can be analyzed quickly. Furthermore, by studying the community structure in both direct and indirect association networks, one can determine a set of elements that are core to an argument, and which are associated but somewhat peripheral to arriving at a particular answer. As an example, recall that in the box friction task, which included a salient distracting feature in the form of the given coefficients of friction, the indirect network showed associations between many of the element expected to be tied to a common incorrect answer, but the direct association graph showed that the main core of the argument was the comparison of the two coefficients. Clear distinctions between correct and incorrect arguments were also seen in the sparsification results across tasks, indicating once again that the lines of reasoning associated with particular answers can be meaningfully separated in chaining task data.

Network sparsification yields further insight into another aspect of student reasoning with the provided elements: on each task shown, sparsification was meaningfully interpreted as the "wisdom of the crowd" consensus about the structure (or logical landscape) of the identified arguments. In most of the tasks reported on, the structure of the associations among the elements revealed information that would not have been available from an examination of the responses individually. For instance, in the work-energy task, the linear structure of the *work as a change in energy* argument compared to the clustered structure of the *work as a dot product* argument would have been hard to ascertain from simply studying the individual responses alone. One affordance of knowing the structure of an argument is to ascertain how students are responding to specific lines of reasoning. For instance, in the two-box friction task, it was seen that the students who responded with the common incorrect answer had a strong consensus to the core argument elements, whereas those students who responded with the correct answer choice did not have a strong consensus in the ordering or arrangement of the reasoning elements. Likewise, the structure of the correct, formal reasoning in the truck friction task indicated a more complex view of that specific line of reasoning compared with the relatively straight forward hypothetical reasoning.

One outcome of the ability to separate and structurally study different lines of reasoning is that specific instructional implications can develop. For instance, the element “both boxes remain at rest” in the two-box friction task is used by students in both the correct and common incorrect lines of reasoning. However, in the correct line of reasoning, that particular element is associated closely with the element indicating that the "horizontal forces are balanced", whereas in the common incorrect line of reasoning, that particular element is not heavily associated with anything else. Attending to developing a connection between the boxes remaining at rest and the idea of balancing horizontal forces during instruction may improve performance on these types of questions.

A further, perhaps more powerful use of network analysis of chaining task data is to isolate and observe specific lines of reasoning before, during and after instruction. The truck friction task demonstrated that it may be possible to isolate specific lines of reasoning (such as the formal line of reasoning) by not including elements from other lines of reasoning. The isomorphic graph tasks revealed that over the course of two semesters, a specific reasoning element regarding the relationship between the slope of a graph and the derivative of the function represented on the graph was more productively incorporated into a line of reasoning. These two results suggest that network analysis of reasoning chain construction task data can be used to isolate and study the development of specific reasoning skills. This could be helpful in assessing the impact of instructional materials on student reasoning with specific arguments. For instance, many instructional materials (especially scaffolded tutorials) step students through qualitative inferential arguments while forming physics conceptual knowledge or teaching problem solving strategies. These same qualitative inferential arguments are then expected to be used on new but similar questions such as those found on exams, for instance. Chaining tasks could be used to study student use of these arguments before, during and after instruction. We likewise feel that chaining tasks, coupled with network analysis techniques, can be utilized to study many types of arguments, and specifically arguments related to the reasoning difficulties identified in physics education research literature.

Results from reasoning chain construction tasks can support analyses drawn from other theoretical and experimental methodologies. In the second task, the truck friction task, we gave evidence from community detections and network sparsification that suggested that students who answer correctly and incorrectly on a friction task are drawing largely upon the same reasoning elements. The difference in the populations was the topology of their argument and the elements on which emphasis was most placed. This finding is reminiscent of studies using the resources framework that posit that different reasoning outcomes may share a subset of similar resources, with only one or two resources not in common with each other. We therefore have hopes that reasoning chain construction tasks coupled with the network analysis techniques described here can be used to support research regarding the resources framework, specifically where resource graphs have been helpful in the past.

Similarly, in the box friction task, we gave evidence suggesting that network analysis could possibly be able to detect unconscious phenomena such as being cued towards a specific answer based on task features. The high betweenness of the observational elements core to the correct and incorrect line of reasoning suggest that students are highly influenced by these features. Additionally, the observed linear topology of the common incorrect line of reasoning and the non-linearity of the correct reasoning suggested a dual-process interpretation wherein the common incorrect line of reasoning was the result of an intuitive process 1 judgement without much consideration of other models. This same trend in topology was seen in the graph tasks reasoning patterns as well, with the correct answer network having more interconnections than the intuitive answer pattern.

On the basis of the demonstrated affordances of facilitating the investigation of specific reasoning chains through novel data generation, assisting in theory building, and informing instruction in new ways, we believe that the network analysis of reasoning chain construction tasks has the potential to become a valuable tool for researchers in physics education. Perhaps most importantly, we are confident that it will be a distinct asset to ongoing efforts to investigate and strengthen student reasoning in physics, particularly those that attend to domain-general reasoning phenomena.

5 EXAMINING STUDENT TENDENCY TO EXPLORE ALTERNATE POSSIBILITIES

5.1 Abstract

A broad goal of physics education is to provide students with a strong repertoire of problem-solving strategies, a familiarity with mathematizing real-world situations, and a strong set of critical thinking skills related to qualitative inferential reasoning. A growing body of research has demonstrated that some patterns in student responses to qualitative physics questions may be attributable to processes general to all human reasoning, and not necessarily related to physics content. Theories from the psychology of reasoning posit that the ability to consider and explore alternate possibilities is a hallmark of strong reasoning skills. Furthermore, recent findings suggest that there may be a link between student ability to consider alternative possibilities and student performance on physics problems — particularly problems in which salient distracting features appear to prevent students from accessing relevant knowledge. We have piloted new tasks designed to measure student ability to consider multiple possibilities when answering a physics problem. These tasks measure the relative accessibility of a mental model (or possibility) as well as student ability to recognize whether or not this model is consistent with given problem constraints. Using these tasks across three physics content areas, we find that a model in which two objects are in opposition (such as two fans pushing in opposite directions)

is less accessible than models in which the objects are not in opposition. This result suggests that a domain-general mechanism may control model accessibility. We expect that this underlying mechanism is a tendency to avoid expending cognitive resources on multiple, complicated models and instead reason from a single, easy-to-represent model.

5.2 Introduction

A typical physics course is full of new vocabulary, procedures for problem solving, and strategies for applying concepts to real-world situations. In addition to learning definitions, procedures, and strategies related to each physics concept, physics students are also often expected to apply their knowledge to reason their way through new and difficult physics problems. Research-based instruction has shown a marked improvement in student performance on questions assessing conceptual understanding and other related abilities (Finkelstein & Pollock, 2005; Saul & Redish, 1997; Sokoloff & Thornton, 1997; Beichner R. , 2007; Crouch & Mazur, 2001). However, despite research-based instruction, some physics questions continue to prove difficult for students, even when students demonstrate that they can generate correct lines of reasoning on questions targeting the same concepts (Heckler, 2011; Kryjevskaia, Stetzer, & Grosz, 2014).

A growing body of research suggests that processes general to all human reasoning and not necessarily associated with physics content may be primarily responsible for the observed discrepancies. As such, it is important to investigate the interplay between domain-general reasoning processes and reasoning in a physics context to understand more clearly how to best prepare students for applying their knowledge to new situations. One such reasoning process is a tendency to search for alternate possibilities. Searching for alternate possibilities is associated with more productive reasoning (Johnson-Laird, 2009; Evans, 2007; Lawson, 2004; Tishman, Jay, & Perkins, 1993) and in some cases may be foundational to productive reasoning. For instance, in Johnson-Laird's mental model's theory of reasoning (Johnson-Laird, 2009), the failure to fully flesh out possibilities is the fundamental mechanism for all reasoning errors.

A student's tendency to explore alternate possibilities can be impacted by the cognitive accessibility of an idea. Cognitive accessibility is a measure of how easily a concept or model is retrieved from memory (Higgins, 1996), and so a search for alternate possibilities can be truncated if the accessibility of an initial idea is much higher than the accessibility of the other possibilities. Heckler and Bogdan recently investigated the effects of accessibility on physics questions (Heckler & Bogdan, 2018). They first measured the relative cognitive accessibility of causal factors in different physics contexts, such as length and mass in the context of determining the period of a pendulum. They then found that when a highly accessible factor was offered in a problem statement, students tended not to explore alternate factors - even when the factor offered was causally irrelevant to the physics scenario (*e.g.,* the mass of a pendulum). Furthermore, when the less accessible factor was offered students did explore alternate factors, namely the highly accessible factor. They surmised from this that accessibility could represent a "soft contour" (*i.e.,* a control mechanism) that influences the trajectory of a reasoning process.

The notion of accessibility is generally applied to the ease of recall of information stored in memory. In this paper, we extend the notion of accessibility to the ease of generating novel possibilities. We aim to examine the relative accessibility of various generated mental models within the context of three tasks, one in a non-physics domain and two in physics domains. In doing so, we aim to provide additional insight about how accessibility might impact reasoning in a physics domain and to shed light on factors that contribute to the relative accessibility of a model in a given context. This work thus serves to deepen researchers' understanding of the interplay between domain-general and domain-specific reasoning in physics.

5.3 Background and Theoretical Framework

When examining the concept of cognitive accessibility in physics, it is critical to have a solid understanding of the relevant frameworks for understanding human reasoning. We will discuss, in detail, two related theoretical frameworks: a class of theories collectively referred to as *dual-process theories of reasoning and decision-making* (Evans & Stanovich, 2013) as well as the *mental models theory of reasoning* developed by Philip Johnson-Laird (Johnson-Laird, 2009). Once this theoretical background has been established, we discuss the notion of accessibility in greater detail along with a focused discussion about accessibility in the context of physics reasoning.

5.3.1 Theoretical frameworks for student reasoning

Dual-process theories posit two processes for reasoning: an automatic, subconscious process 1; and an effortful, slower process 2 (Evans & Stanovich, 2013). Process 1 is responsible for constructing the most plausible model based on contextual clues and prior knowledge. When there is a reason to expend effort, process 2 comes into play by recruiting working memory to run simulations, test hypotheses, or execute an algorithm. This process is helpful with problems such as long division, deducing a result from first principles, or deciding which tax cut to take. In most dual-process theories, the searching for alternate models occurs only if process 2 finds the default model supplied by process 1 to be insufficient in some way (*e.g.,* (Evans, 2006)) or if the reasoner has an innate disposition to execute that search

(Tishman, Jay, & Perkins, 1993; Thompson, 2009). Otherwise, the default model tends to be the only model considered.

Another theoretical framework for reasoning is Johnson-Lairds' theory of mental models (Johnson-Laird, 2009). For Johnson-Laird, a mental model is a mental representation of the relationships between objects; reasoning is then a process of simulation based on that representation. Reasoning errors are due primarily to the failure to represent all possible models of a given situation. Khemlani and Johnson-Laird (2016) extended the theory of mental models to include a dual-process perspective and posit that process 1 puts forward a single mental model from which an intuitive judgment is made. This mental model is limited by a human tendency to use reduce the load on working memory and other cognitive resources (*i.e.,* cognitive miserliness). In the theory, if more cognitive resources are available and there is a need to recruit such resources, counterexamples to the original judgment are then sought after by representing more possibilities until all possibilities are fully fleshed out.

To describe Johnson-Laird's theory in greater detail, we provide an example to show each step in the theorized mental model reasoning process. Consider the following logical statement[4]: *If there is a circle, then there is a*

[4] For a mathematician, this statement and the following discussion may appear odd. This statement was originally phrased by Johnson-Laird and could be modified to read "If there is X, then there is Y", which is logically equivalent. Johnson-Laird's representation of mental models is also not supposed to

triangle. Johnson-Laird represents mental models on paper via a spatial arrangement of icons (*i.e.,* words) that reflect theory's stance that real mental models are also spatial arrangements of mental icons. Using Johnson-Laird's representation, a single, fully fleshed out mental model of the information in this statement would consist of three distinct possibilities:

Circle	Triangle
No Circle	Triangle
No Circle	No Triangle

Due to a tendency to preserve resources, Johnson-Laird's framework predicts that the typical mental model produced by process 1 would not be fully fleshed out, but instead be abridged:

Circle	Triangle
…	

Rather than represent all three possibilities explicitly in the model, the mind keeps a mental "footnote" (Johnson-Laird's representation of this footnote is an ellipses) as a reminder that other possibilities exist and that the model would need to be fleshed out to include representations of these possibilities if the task requires it so to be.

Which models get "footnoted" and which get explicitly represented in the model produced by process 1 is partially governed by the "principle of truth",

represent a logical truth table or logically equivalent statements. They simply simulate possible configurations that are consistent with the premise.

which says that models represent what is true in a possibility rather than what is false. Thus, the situation in which "Circle" is false (*i.e.,* "No Circle") is not explicitly represented but rather is left to be explored if the situation demands it. In contrast, if the logical statement had been "If there is not a circle, there is a triangle," then the intuitive mental model would be

No Circle	Triangle
	…

with the ellipsis denoting the two non-represented possibilities where there is a circle. In this case, the semantic content of the initial phrase implies the object to consider is "No Circle".

In the *mental models theory of reasoning,* the tendency of the human mind to "footnote" certain possibilities is the source of all observed systematic reasoning errors. For instance, consider the following logical problem: "If there is a circle, then there is a triangle. There is not a circle. What follows?" A common answer to this problem is "there is not a triangle", but this answer is incorrect. As indicated by the fully fleshed out mental model shown above, there are two possible outcomes associated with the absence of a circle: either there is a triangle or there is not a triangle. Thus, nothing follows deductively from the two statements given in the problem. Reasoners make an error on this problem because they are reading a judgement directly from the intuitive, abridged mental model produced by process 1 (*i.e.,* the second one depicted) rather than fully fleshing out the model to include all possibilities and reading a judgement from that more thorough simulation.

5.3.2 Cognitive accessibility and availability

We now examine the notion of cognitive accessibility. Accessibility is best understood in contrast to availability. Knowledge (concepts, mental models, procedures, *etc.*) is available if it (or some of its constituent parts) is stored in memory, whereas the accessibility of knowledge is a measure of how readily this knowledge can be activated or brought into working memory. In other words, accessibility is an "activation potential of available knowledge" (Higgins, 1996). The accessibility of specific knowledge structures is posited to be primarily dependent on the strength of associations between it and other relevant structures. For instance, "fleas" is a highly accessible explanation for a scratching dog because fleas and dogs are strongly associated with each other (Quinn & Markovits, 1998). These strong associations are mostly formed through repeated exposure to the association during the course of everyday experiences. However, the accessibility of a mental construct can also be temporarily increased through priming effects. If a particular concept (*e.g.,* a stereotype, see Wheeler & Petty, 2001) is unconsciously primed (*e.g,* by subliminal exposure to words related to the stereotype), ideas associated with that concept become temporarily more accessible and it is possible to study the time-decay of that accessibility

(Higgins, 1996). As such, the accessibility of a given knowledge structure is both context-dependent and time-dependent.

The accessibility of various knowledge structures impacts which of those structures process 1 draws upon during the act of reasoning. For instance, when two or more models are in competition, the model with the higher accessibility tends to be constructed or selected for use in reasoning. According to most dual-process theories, reasoners tend to utilize a single model while reasoning; thus, a highly accessible model can hinder a student's exploration of possible alternate models. This was shown clearly in Heckler and Bogdan's study of accessibility in a physics context (Heckler & Bogdan, 2018). When highly accessible explanations for physics phenomena were offered in the question prompt, students did not tend to reason via alternate explanations, whereas when less accessible explanations were offered, they did.

In that study, and in line with other studies regarding accessibility (*e.g.,* Quinn and Markovits), relative cognitive accessibility was operationally defined as the relative number of times that an explanation is listed in a free-recall task. As an example, one such free-recall task told students that "Pendulum A swings with a longer period (time) than pendulum B" and were prompted to "list the possible reason(s) why pendulum A has a longer period". Heckler and Bogdan report data regarding which explanations were listed, which were listed first, which were listed singly, and the number of times that all explanations were offered. Considering all of these measures together, they determined which explanations were highly accessible and subsequently manipulated the presentation of physics questions to control for the explicit mention of these explanations. For instance, they presented the observation that "Pendulum A swings with a longer period (time) than pendulum B" and asked students if the statement "Pendulum A has a longer string than pendulum B" was a valid explanation for the observation (length being a highly accessible factor). By varying the offered explanation, they determined that the accessibility of the offered explanation impacted whether students would explore alternate explanations for the stated observed phenomena.

5.3.3 Applying accessibility and availability to mental models

The language of accessibility and availability as employed by Heckler and colleagues has generally referred to the recall of knowledge structures (such as the relevancy of length to the period of a pendulum). Since we were interested in exploring Johnson-Laird's *mental models framework* as a means of studying the extent to which particular possibilities could be generated or identified in a given research task, we have applied the same notions of accessibility and availability to reasoner *generated* possibilities.

To illustrate the difference, consider the syllogism "All artists are bakers, some bakers are chemists. What follows?" A typical response is to say that "Some artists are chemists" because, according to Johnson-Laird, the initial, abridged mental model would be

Artist	Baker	Chemist
Artist	Baker	Chemist
Artist	Baker	
	…	

Indicating a model in which there is the possibility of an artist-baker not being a chemist. Because reasoners don't initially represent what is not true, the possibility of there being bakers who aren't artists does not readily occur to reasoners and reading off of the model above they conclude that some artists, but not all, are chemists. A more fleshed out model may look something like

Artist	Baker	Chemist
Artist	Baker	Chemist
Artist	Baker	
	Baker	
	Baker	Chemist
	…	

in which case the reasoner would readily read off that there was no definite relationship between the state of being an artist and being a chemist. (That is, it could be that none of the artists are chemists.)

The point of this discussion is to illustrate that a student would likely have never considered specific arrangements of artist-status and baker-status prior to the task. Thus, they can't reasonably be said to be recalling information about these arrangements. Rather, they are generating novel models in response to the task. However, the concept of accessibility still applies. In the context of the task, where artist is listed first, possibilities in which bakers are also artists are more accessible than possibilities in which bakers are not artists. This difference in accessibility has ramifications for the model that is constructed for use in reasoning: it is unlikely that an initial model would appear as

	Baker	Chemist
	Baker	Chemist
Artist	Baker	
	…	

In this study, we extended the concept of accessibility to reasoner-*generated* models in a physics context and used this to pursue a greater understanding of the tendency to explore alternate possibilities. In particular, the investigation was designed to answer the following research questions. Can investigating the relative accessibility of mental models in both physics and non-physics contexts provide a better understanding of the mechanisms that control reasoning in a physics context? Can such an investigation also yield insight into domain-general reasoning phenomena occurring while students answer physics questions?

5.4 Methods

To deepen an understanding of the interplay between domain-general and domain-specific reasoning, we aimed to study how accessibility might impact reasoning in a physics domain as well as to investigate factors that contribute to the relative accessibility of a possibility in a given context. Accordingly, we created three isomorphic tasks that probe student tendency to explore possibilities. These tasks span three content areas: a purely numerical context, a force and motion context, and a circuits context. In all tasks, students are asked to identify all possible arrangement or configurations consistent with the given premise. The tasks are intentionally isomorphic in construction. By this, we mean that each task has a similar underlying reasoning pathway that requires students to determine a set of discrete values that sums to a given positive number.

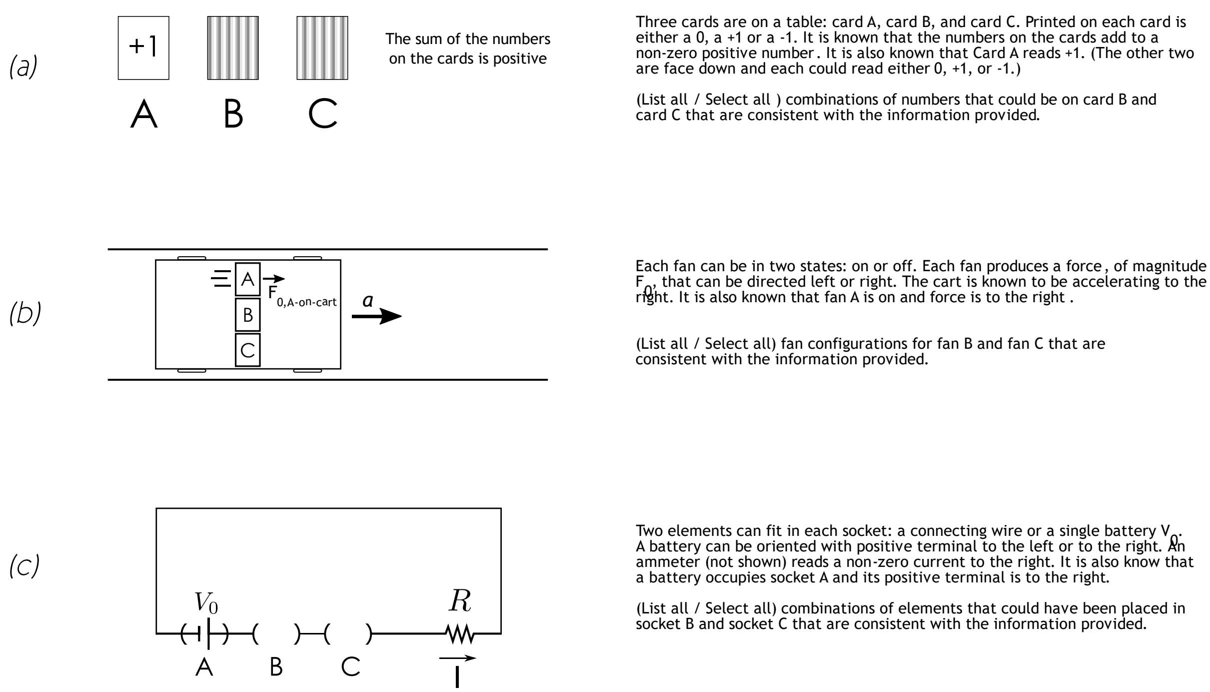


Figure 5-1. Three isomorphic tasks designed to investigate relative cognitive accessibility in (a) a numerical “physics-less” context, (b) a forces context, and (c) a circuits context.

In the first task (see Figure 5-1.a), the question states that three cards are laid out on a table, and that printed on each card is either a 0, a +1, or a -1. The student is told that the first card reads +1 and that the other two cards each could read either 0, +1, or -1. The students' task is to determine the combinations of numbers that could be printed on cards B and C such that the cards sum to a non-zero positive number. There are 6 combinations of numbers that could satisfy the premises: (i) both cards could read 0, (ii) card B could read +1 and card C could read 0, (iii) card B could read 0 and card C could read +1, (iv) both cards could read +1, (v) card B could read -1 and card C could read +1, and (vi) card B could read +1 and card C could read -1.

The other two tasks set up the same basic situation in the context of forces and circuits. In the forces task (shown in Figure 5-1.b), a fan cart has three fans that can direct a force on the cart to the right or to the left. The fans can also be turned off. The students are told that the fan cart is accelerating to the right and also that fan A is on and the force on the cart by the fan is to the right. In the circuits task (shown in Figure 5-1), a circuit has a battery of voltage V_0 oriented with its positive terminal to the right, two sockets (each of which could hold a battery, with its positive terminal oriented either to the right or to the left, or a straight connecting wire), and a resistor. It is indicated that an ammeter measures a non-zero current to the right (as shown in the diagram). Thus, in all three cases, the students must consider how two object states, each of which could be represented as a value that could be either zero or signed positive or negative, combine with a given third object state to produce an effect that is signed positive.

The accepted way of combining the objects is context specific. In the numerical context, the rule is given to the students: the numbers on the cards must be added. In the forces context, the forces from the fans on the cart can be represented as vectors, with the sign of the vector indicating the direction of the force on the cart. Newton's 2^{nd} Law provides the underlying rule governing how these forces are combined to produce an acceleration: the vectors are added and the direction of the sum of the force vectors is the direction of the acceleration. In the circuits context, Ohm's law governs the relationship between the cause (a potential difference ΔV) and the effect (a current). The potential difference is additive and signed based on the orientation of each battery (*i.e.,* where its positive terminal is with respect to the rest of the circuit).

In all three of these tasks, there are only nine combinations of states among the two objects. If one represents the state of each of the remaining objects (B and C) using the symbols +, -, and 0, one can represent all nine combinations in a concise fashion, as shown in Figure 5-2. Of these nine, only six are consistent with the information provided in the task statement.

These six are indicated in Figure 5-2 via the absence of shading. It should be noted that some of these combinations (e.g., 0+ and +0) would be equivalent if the two objects (B and C) were indistinguishable. Our treatment of such combinations during data analysis will be described in greater detail in the results section. As a clarifying note, the term possibility will be used interchangeable with configuration and combination of states in the following sections and does not necessarily denote a possibility in the Johnson-Laird sense, except where explicitly referenced as such.

		Object C		
		0	+	−
Object B	0	00	0+	0−
	+	+0	++	+−
	−	−0	−+	−−

Figure 5-2. A table of the nine possible combinations of states. Only six of these are consistent with the premises given in the three iso-morphic possibilities tasks. The other three have been shaded in the table to indicate that they are inconsistent with the task premises.

Each task was administered online using the Qualtrics survey software. The tasks were administered on homework assignments or exam reviews for students enrolled in an introductory calculus-based physics sequence, along with other questions relevant to the course but not relevant to the content found in the research task. These assignments counted for participation credit in most cases, although extra credit was awarded in some cases. In all cases, the tasks were administered after relevant lecture, laboratory, and small-group recitation instruction at a research-intensive university in New England. Research-based materials from *Tutorials in Introductory Physics* (McDermott & Shaffer, 2001) were used in the recitation section. Given that the tasks were all administered across a single academic year, most students who completed the year-long introductory calculus-based sequence would have seen and completed all three tasks.

To examine the relative accessibility of the different possibilities inherent in these questions, we used a between-student methodology and randomly split students into two conditions: (1) a *select condition* in which students were asked to select possibilities from a list of configurations, and (2) a *generate condition* in which students were asked to generate their own list of possibilities. Together, the data from the two conditions enable us to gather information about which possibilities are available to students (*i.e.,* possibilities that students are able to recognize as consistent with the given information and the rule for adding the quantities, captured by data from the select condition) and which are relatively accessible (*i.e.,* possibilities that are easily generated when students are left to come up with their own, captured by data from the generate condition).

In our analysis of these data, we operationalized relative accessibility three different ways, in the tradition of Heckler and Bogdan, 2018. In the first approach, we simply determined how many students cited a particular possibility in the generate condition. The second way we operationalized relative accessibility was to determine how many students in the generate condition listed a particular possibility *first* in their response. This approach implicitly assumes that students did not edit their responses but simply listed their models in the order in which they came to mind (or at least quickly listed the first model that came to mind). We recognize, of course, that this may not always be the case. Finally, since accessibility is proposed to inhibit the exploration of alternate possibilities, the third operationalization of relative accessibility was to examine the relative prevalence of models listed by students who only generated one possibility in their responses. Given that each approach had associated limitations, we used all three approaches to estimate the relative accessibilities of the various possibilities, thereby ensuring that our results were reliable. It is worth noting that a similar multi-operationalization approach was employed by Quinn and Markovits (1998) as well as Heckler and Bogdan (2018). As we discuss later, all three operational definitions yielded similar results and served to strengthen our results.

5.5 Results

The results are grouped into three sections. In the first, we introduce a data-driven coding scheme used in the subsequent results sections. In the second, we provide general results, and in the third section we provide results from the three different ways of operationalizing accessibility. The three tasks are considered together in each section.

5.5.1 Coding Scheme Development

As discussed in Section 5.4, possibilities such as +0 and 0+ could be considered to be equivalent if objects B and C were effectively indistinguishable. For this reason, in our initial analysis of the data we probed the ways in which students treated those possibilities that would be equivalent for indistinguishable objects. The two quotes presented below are illustrative of the kinds of responses students gave when ask to generate possible configurations in the forces task.

> *"B and C can be off, both be applying force to the right, or one applying force to the right and the other either off or applying force to the left"*

> *"- B and C may both be turned off*
>
> *- B and C may both be turned on and to the right*
>
> *- B may be on and to the right, C may be on and to the left*
>
> *- B may be on and to the left, C may be on and to the right"*

In the first response, the student treated components B and C as indistinguishable, noting that one could be to the right while the other could be off to the left and suggesting that it doesn't matter which is which. In the second response, the student treated the components as distinguishable but explicitly mentioned both the +- possibility and the -+ possibility. It was observed that in over 98% of student responses, the components (B and C) were either treated as indistinguishable or both possibilities in a given set of distinguishable possibilities were mentioned.

Therefore, for coding purposes, we collapsed the six consistent configurations down to four, and the three inconsistent configurations down to two. In particular, we established a single *opposition configuration* code, denoted +-, in which the two components are competing. Configurations in which one component is "on" and the other is "off" (*i.e.,* +0, 0+, -0, and 0-) were denoted either + (when "pushing" to the right) or - (when "pushing" to the left). Finally, the possibility that neither component B nor C is "pushing" was denoted 0.

5.5.2 General Results

In the numerical context, the *select* and *generate* conditions were similar in the number of possibilities chosen (see Figure 5-3.a). Almost 70% of students recognized all of the possibilities that were consistent with the premises given (select condition), and close to 60% of students were able generate and endorse these possibilities on their own (generate condition). Another 40% generated and endorsed three of the four consistent possibilities.

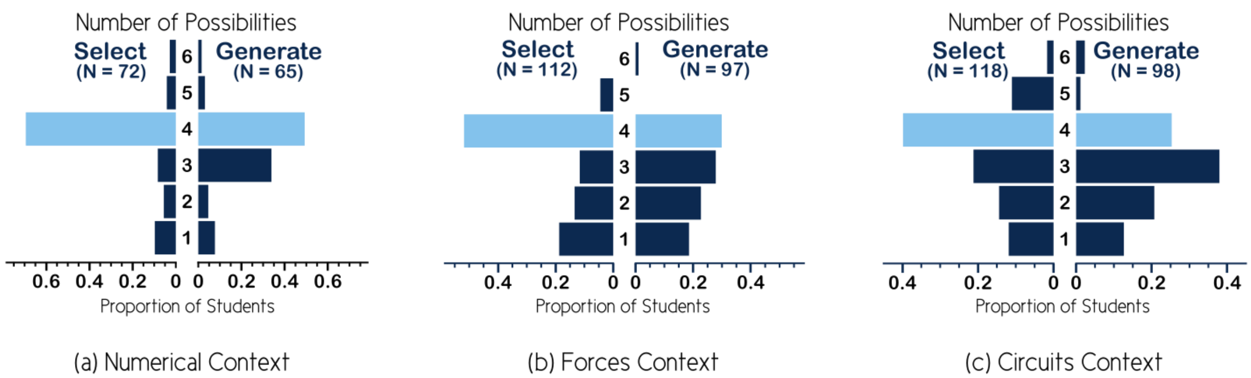


Figure 5-3. Plots showing the proportion of students selecting a given number of possibilities for (a) the numerical context, (b) the forces context, and (c) the circuits context.

In the forces context, there was a stronger performance difference between the Select and Generate conditions. As shown in

Figure 5-3.b, 50% of students recognized all four possibilities as consistent with the premises in the select condition, while only 30% were able to generate all four possibilities. The circuits context yielded very similar results (

Figure 5-3.c).

5.5.3 Accessibility measures

In this section, we examine the results of measuring accessibility via three different operationalizations of the concept, namely accessibility as measured by the percentage of students who generated the possibility, by the percentage of students who listed the possibility first in their response, and by the percentage of students who listed the possibility as the only possible configuration.

5.5.3.1 Accessibility as measured by the percentage of students who generated the possibility

Figure 5-4 shows a comparison of the percentage of students who endorsed a configuration in their response in the select and generate conditions for each task. Each row constitutes a distribution of endorsed possibilities. Typically, the configurations consistent with the premises were endorsed by the majority of students, while the two inconsistent configurations were not highly endorsed. The tables of percentages are shaded according to the percentages in a linear function with zero shading (white) corresponding to the maximum percentage in the table, and max shading (black) corresponding to the minimum percentage for the table.

		0	++	+	+−	−	− −	Chi-square results (First Four Possibilities)
Numerical	Generate (N = 66)	77%	76%	80%	80%	17%	5%	$p = 0.9,\ V = 0.02$
	Select (N = 72)	81%	76%	85%	90%	15%	14%	$p = 0.2,\ V = 0.08$
Forces	Generate (N = 97)	62%	77%	71%	47%	2%	12%	$p < 0.01,\ V = 0.14$
	Select (N = 112)	70%	69%	68%	71%	18%	15%	$p = 0.96,\ V = 0.01$
Circuits	Generate (N = 98)	66%	64%	69%	41%	7%	7%	$p < 0.01,\ V = 0.14$
	Select (N = 118)	76%	67%	75%	61%	31%	19%	$p = 0.04,\ V = 0.08$

Figure 5-4. A comparison of the percentage of students that endorsed a configuration in their response in the select and generate conditions for each task. For each distribution, a chi-square test was used on the four consistent configurations to gauge the extent to which observed variations were statistically significant. The tables are shaded according to the percentages in a linear function with zero shading (white) corresponding to the maximum percentage in the table, and max shading (black) corresponding to the minimum percentage for the table.

For each distribution, a chi-square test was used on the four consistent configurations to gauge the extent to which observed variations were statistically significant. These tests revealed a general trend in which all consistent configurations were equally likely to be selected from a list, but also indicated statistically significant differences in each of the two physics contexts for the *generate* condition.

5.5.3.2 Accessibility as measured by how many times the possibility was listed first

The percentage of students in the generate condition who listed a specific possibility first in their response is given in Figure 5-5. A consistent pattern is shown across all three tasks: the 0 configuration appears most prevalent, and the +- configuration is the least likely (of the four consistent possibilities) to be listed first.

		0	++	+	+−	−	− −
Listed First	Numerical (N = 65)	31%	29%	18%	17%	5%	0%
	Forces (N = 97)	41%	34%	16%	5%	0%	3%
	Circuits (N = 87)	48%	26%	16%	7%	1%	1%

Figure 5-5. Percentage of students in the generate condition that listed a given possibility first. These values also include students who only listed one possibility.

5.5.3.3 Accessibility as measure via the number of students who listed the possibility alone

By looking at which configurations were the only ones listed by a particular student (see Figure 5-6), we found that in all contexts, students who only listed one consistent possibility tended to list the 0 or the + configurations (>60%), while the +- configuration was only rarely listed alone, if at all (< 9%). The sample size was fairly small, however, so firm conclusions are hard to make about which possibility was most accessible according to this measure alone. Nevertheless, it seems certain from this measure that the opposition configuration (+-) was among the lowest accessibility models taking the three contexts together.

		0	++	+	+−	−	− −
Listed Alone	Numerical (N = 5)	1	0	2	0	2	0
	Forces (N = 18)	8	5	1	1	0	3
	Circuits (N = 11)	3	1	5	1	1	0

Figure 5-6. Absolute number of students who listed only one possibility broken down by what possibility they listed. The absolute number of students is shown rather than a percentage because the number of students in this condition was so small.

5.6 Discussion

On these tasks, students treated the two objects as indistinguishable when generating possible configurations – even when they distinguished between the two objects in their response. We view this result as suggesting that a student who generates the configuration 0+ and the configuration +0 are generating both configurations from the same underlying mental simulation which treats the components indistinguishably. Thus, we are inclined to believe our coding scheme represents the 6 different underlying mental simulations used by students when reasoning through these tasks, four of which produce results which are consistent with the premises of the problem.

In the following discussion, we refer to these simulations as *models.* Because there is a difference in the way that the term *mental model* is used in Johnson-Laird's framework and in the more general dual-process theories, we wish to introduce a notation that will aid the reader in distinguishing what is meant. Johnson-Laird refers to a collection of possibilities as a single mental model of the premises, whereas in dual-process theories generally a single possibility is considered the mental model. Therefore, when we refer to a collection of possibilities, we will use the term *JL mental model.* Otherwise, when using the terms *mental model* or *model*, we are referring to the single underlying model that corresponds to a single possible configuration.

To summarize the basic results, students were able to select more configurations from a list than they were able to generate on their own. Also, in all contexts, each consistent model was generally equally available to more than half of the students, as evidenced by the results of the chi-square test on the distributions in the select condition. Additionally, the inconsistent models were avoided by most students.

While our results indicated that the models were generally available to most students, we discovered a difference in the relative accessibility of each model. Taking the three methods of measuring accessibility together, it appears that the opposition model, +-, was less accessible for students than the other models. Additionally, the 0 model was consistently among the top for accessibility.

These results can be interpreted in a few different ways. In the first interpretation, we look to the proposed mechanism for accessibility. Since the primary mechanism driving accessibility is posited to be the strength of associations between the knowledge structure and the context in which the knowledge is being activated, and given that the specific configurations of objects in these tasks is largely novel to the student, one could ask "what is being associated with the context of these tasks when generating a model to determine possible configurations?"

One might propose that each configuration is constructed from a pairing of an abstract knowledge element -- like the fine-grained phenomenological primitives described in diSessa (1993) -- with the conceptual content of the task (*e.g.,* the vector nature of Newton's second law), and that it is these abstract structures that are more or less associated with the context of the task. For instance, a "status quo" structure (such as the "WYSIWYG" knowledge element from Elby, 2001) might seek to take the context "as-is" without hypothetical simulations. This structure could combine with the specific task features to create the 0 model. Likewise, a "conformity" structure (which seeks homogeneity) would combine with the state object A is in to create the + and ++ models; or an "opposition" structure (like diSessa's "canceling" p-prim (diSessa, 1993)) produces the +- model in the context of these tasks. Note that our purpose isn't to define these structures, but simply to propose their existence and effect. One could say, then, that these abstract structures are invoked with certain relative accessibilities in each context due to the level of their association with the particular context.

If this were true, one could argue that the associations with the abstract structures were based more on the underlying structure of the problem (*i.e.,* entities combining through vector addition) rather than the specific task features or content area (*i.e.,* whether the entity was a battery or a card) because it appears that the least accessible and most accessible models for each context are the same (+- and 0, respectively). This would represent a domain-general effect related more to problem structure than physics content.

An alternate interpretation uses the perspective of cognitive miserliness – that is, a reasoner's tendency to avoid large expenditures of cognitive resources such as working memory when reasoning through a task. It could be that the observed lack of accessibility of the +- model is due to the cognitive effort involved in mentally representing that model. Recall that the 0 model seemed most accessible across the three tasks, with the + and ++ models next, and finally the +- as least accessible. We would expect that a model in which nothing changes (the 0 model) would be the easiest to maintain in working memory since it only requires that one object be represented (object A). Likewise, models in which only one extra component needs to be represented (+ and –) or where the components are in the same state (++ and --) would be more easily maintained in memory than the opposition model, in which it is necessary to represent two extra components in different states.

Of note is that while two out of the three accessibility measures showed the opposition model as less accessible in the numerical context, the opposition model was just as accessible as the other consistent models in the numerical context when accessibility was operationalized as the number of times a model was cited in the generate condition. We suspect that the construct of cognitive miserliness may contribute the apparent inconsistencies in accessibility of the opposition model in the numerical context. Given that two out of three accessibility measures indicate that the opposition model is less accessible in the numerical context and given that the opposition model was found to be less accessible in the other two contexts, we are inclined to believe that the opposition model is, in fact, less accessible in the numerical context as well. However, because fewer cognitive resources were devoted to interpreting the content and representing the cards (the physical objects) as abstract mathematical objects in the numerical context, students had more resources available for exploring alternate possibilities. Thus, from the perspective of cognitive miserliness, the opposition model is more likely to be generated in the numerical context than in the two physics contexts, as the latter two contexts require more resources for navigating the specific conceptual content of the tasks. Findings from the "Listed First" measure of accessibility (see Figure 5-5. Percentage of students in the generate condition that listed a given possibility first. These values also include students who only listed one possibility.) are consistent with this interpretation. The results shown in Figure 5-5 indicate that the 0 model (the simplest one) increased in accessibility while the other models (including the opposition model) decreased in accessibility while going from the numerical context to the circuits context, with the latter being the most complex of the contexts.

Indeed, this interpretation through the lens of cognitive miserliness is also consistent with Johnson-Laird's *mental models theory of reasoning* since, in his framework, factors such as working memory capacity are tightly linked to how many possibilities are explicitly listed in a JL mental model and which are "footnoted". As discussed in Section 5.3, the "principle of truth" partially governs which possibilities are explicitly represented in the intuitive model. This principle states that the intuitive process typically yields a JL mental model that only includes those things that are "true" and does not represent models that include things that are "not true". The problem statement in all three contexts alludes to object A having a positive-signed quantity and the net effect being positively signed. It could be that the problem statement in all three contexts sets up the reasoning such that the positive-signed quantities are the "true" quantities to represent. Thus, according to Johnson-Laird's "principle of truth", the mind is biased against representing things that run counter to positive-signed quantities; that is, the reasoner is predisposed to not put any negative-signed quantities in the possibilities, unless those are structural consequences of the premise.

Of the two interpretations of our data, we favor the cognitive miserliness interpretation because it accounts for the finer details of the results and it combines multiple theoretical perspectives into a single coherent model of student reasoning in physics. However, further research would need to be done to establish for certain whether cognitive miserliness was indeed the controlling factor for the observed relative accessibility. If further research supports this interpretation, it would constitute a control mechanism for accessibility beyond simple associations and might serve to enhance the breadth and applicability of the concept of accessibility. Even in the absence of a coherent, robust understanding of the phenomenon observed in this work, the empirical results alone still have the potential to inform instruction, as discussed in the next section.

5.7 Conclusions, implications for instruction, and future work

In this study, we examined the relative cognitive accessibility for reasoner-generated mental models inside and outside of physics contexts. Three isomorphic tasks were developed to probe student tendencies to explore alternate possibilities consistent with a given premise. These tasks all had the same underlying structure (the addition of three signed/vector quantities) but were in different contexts and they probed student tendency to explore alternate possibilities consistent with a premise. We analyzed results to these tasks to determine the availability and accessibility of the different possible configurations. We found a consistent pattern across three content areas suggesting that a model in which two objects are in opposition (such as two fans pushing in opposite directions) is less accessible than models in which the objects are not in opposition. Because this pattern spanned two different physics contexts, we are inclined to believe that a domain-general mechanism may control model accessibility. In particular, we speculate that this underlying mechanism is cognitive miserliness, or the tendency to avoid expending cognitive resources on multiple, complicated models and instead reason from a single, easy-to-represent model.

Regardless of the mechanism responsible for the phenomenon, our findings have specific implications for instruction and further research. First, the finding that students can recognize that a model is consistent with premises but have difficulty generating the model on their own suggests that physics questions in which possibility generation is used as a measure of availability of conceptual knowledge may in fact be testing the accessibility of that knowledge instead. Such questions, therefore, if used alone, may not be appropriate for assessing student knowledge of a particular concept, as they will tend to underestimate the corresponding level of knowledge.

Moreover, our findings suggest that accessibility-related phenomena could impact reasoning on questions leading to a competition between an opposition model and a more accessible model such as a "null" or 0 model. Based on this work, it would be expected that students would exhibit a preference for lines of reasoning based on the latter, more accessible models. Future work should be directed toward verifying this claim.

Finally, we anticipate that future research will explore the extent to which these findings might help uncover a mechanism behind some of the documented conceptual difficulties in certain areas of physics, particularly when involving vector quantities. For instance, there is a tendency for students to treat momentum as a scalar when totaling the momentum of a system. It may be that underlying this difficulty is a bias toward not explicitly representing mental models that require opposing quantities. Further work is needed to link these two largely independent lines of existing research.

Research-based materials have focused primarily on conceptual understanding and scientific reasoning skills. The overall results of this paper (and related studies) point to a need for a better understanding of the interplay between domain-general reasoning processes and content-specific reasoning with physics concepts. With an improved understanding of this interplay, a next generation of research-based materials can be developed that help students navigate these domain-general reasoning processes in the context of physics while also preparing them for more effective reasoning outside of a physics context (*e.g.,* in a future career).

6 CONCLUSIONS AND FUTURE WORK

The goal of the work presented in this dissertation was to provide new methodologies to examine qualitative inferential reasoning that separate reasoning skills from understanding of a particular physics concept. This dissertation presented two new methodologies, the reasoning chain construction task and the possibilities task, and demonstrated their utility in exploring mechanistic processes related to the generation of qualitative inferential reasoning chains and in revealing insight into the nature of student reasoning generally. In this section, we review the results of each investigation, discuss broad implications, and then discuss future directions for research and implications for instruction.

6.1 Review of Results from Chapter 3

In Chapter 3, we presented the results of a study in which the reasoning chain construction task was utilized to probe the extent to which dual-process theories could account for and predict student behavior on tasks with salient distracting features. From Evans' heuristic-analytic theory (Evans, 2006), we developed a working hypothesis stating that students would be unlikely to shift away from an incorrect default model cued by process 1 unless they were provided with information that explicitly refuted the satisfactoriness of that model. Two sets of experiments built on the chaining task methodology were devised to test this hypothesis. In the first, students were given graph tasks with a known salient distracting feature (the intersection point, see Figure 3-1) which had been cast into a chaining format; the reasoning elements in the chaining task version of the graph task served to give students access to relevant conceptual information, thus testing whether or not this improved access would be sufficient to increase performance. In the second set of experiments, we gave students access to information (via the analytic intervention element, or AIE) that refuted a common incorrect default model about static friction in order to determine whether the presence of this information improved performance, as suggested by our working hypothesis.

Several important lessons emerged from these experiments. The first set of experiments showed that providing increased access to relevant, correct information was not enough to produce a large shift in performance on a kinematics question with a salient distracting feature. Instead, the correct information was used by many students to justify an incorrect (and therefore inconsistent) answer. Additionally, the salient distracting feature had a recognizable effect in three other content domains as well, and correct reasoning elements provided in each domain were not enough to negate the effects of the salient distracting feature on the reasoning process.

The second set of experiments showed that a large increase in performance could in fact be realized by providing access to information (via the analytic intervention element, or AIE) that refuted a common incorrect default model cued by the salient distracting feature on the two-box friction task. This set of experiments revealed that the AIE had a greater impact on students who had previously demonstrated relevant mindware (*i.e.,* answered a screening question correctly with correct reasoning) and that there was no statistically discernible change in performance for those students who had not demonstrated that they possessed the relevant mindware. Together, these results provide further support for the use of dual-process theories as a mechanistic framework for making and testing predictions about student performance and behavior, particularly about which models are selected and why, in turn, some are abandoned.

6.2 Review of Results from Chapter 4

The overarching goal of the investigation detailed in chapter 4 was to show the usefulness of network analysis of data stemming from this methodology towards the goal of gaining insight into the composition and structure of student reasoning chains. In addition, we illustrated the many ways in which the novel data resulting from network analysis of reasoning chain construction tasks could be leveraged for future research regarding reasoning difficulties, reasoning resources, and reasoning mechanisms.

We provided four tasks that highlighted various aspects of the usefulness of network analysis on chaining task data. The first task established the uses of various network analysis methods and measures. The second task provided evidence that students drew upon the same set of reasoning elements when arriving at both correct and incorrect conclusions, but placed emphasis on different elements, consistent with studies using the resources framework to study the topology of student causal nets via resource graphs. The third task hinted at the possible use of network analysis techniques on chaining task data to provide insight into dual-process effects by revealing a sub-community of “same is same” elements and showing that salient distracting features had a “short-cutting” effect on student reasoning chains. Finally, the fourth set of tasks showed evidence for the development, over the course of a two-semester physics course, of a justification argument for a graph-reading heuristic, showing the usefulness of reasoning chain construction tasks for assessing the development of specific reasoning skills before, during, and after scaffolded instruction.

The results of this investigation point to the usefulness of chaining tasks, coupled with network analysis techniques, to study many types of arguments, particularly those arguments related to the reasoning difficulties identified in physics education research literature. Additionally, reasoning chain construction tasks may also be leveraged to investigate how students coordinate reasoning resources while solving through a physics problem. Finally, it may be that the associations students make while assembling a reasoning chain on a chaining task are reflective of subconscious associations and reasoning processes. Thus, network analysis of such tasks may be useful in studying the effects of domain-general mechanisms.

6.3 Review of Results from Chapter 5

Chapter 5 introduced the possibilities tasks, which were designed to examine the relative cognitive accessibility for generating various mental models inside and outside of a physics context. The tasks revealed a consistent pattern across three content areas, which suggested that a model in which two objects are in opposition (such as two fans pushing in opposite directions) is less accessible to reasoners than models where the objects are not in opposition. Because this pattern spanned two physics content areas, it was proposed that a domain-general mechanism controls which model is accessible. This mechanism was proposed to be cognitive miserliness, or the tendency to avoid expending cognitive resources and instead reason from single, easy to represent models.

Thus, chapter 5 illustrates the process of gaining insight about a domain-general mechanism in the context of physics by implementing a new methodology (*i.e.,* by studying the relative accessibility of each mental model in three different contexts using possibility generation tasks).

6.4 Implications from all three studies

Each of the studies described in this dissertation explored the interplay between domain-general reasoning and content-specific reasoning. The first showed that introducing a cognitive scaffold based on a domain-general reasoning mechanism produced an increase in performance for the students, thus giving more definition to the cognitive mechanism interacting with the specific context. The second showed that it is possible to generate new forms of data that are useful to studying content-specific reasoning, and possibly to uncover insights into domain-general reasoning mechanisms as well. The third highlights how another reasoning mechanism - cognitive miserliness – can impact the tendency to search for alternate models. In each case, the results spanned multiple contexts, thereby allows us to more thoroughly characterize the interplay between context and domain-general reasoning.

6.5 Future directions

The results from network analysis of reasoning chains appear to be robust as the interpretations of the network measures were consistently applied across many contexts. However, if one were to continue exploring network analysis of chaining task data, a productive route would be to observe the behavior of the analysis methods in a wider variety of contexts to further verify that the interpretations remain consistent across contexts. Secondly, an exploration of other network analysis measures could be productive. In particular, it may be that a stochastic block modeling community detection algorithm (the favored algorithm of Fortunato (2010)) produces more reliable communities, and this may provide greater consistency across tasks or reveal inconsistencies across tasks leading to further insight. Furthermore, working with directed graphs more extensively and analyzing the shortest and most probable paths could also help us better understand student reasoning patterns and phenomena.

Perhaps more exciting are the possibilities for utilizing chaining tasks to study various reasoning phenomenon already identified in the literature. For instance, scaffolded materials such as *Tutorials in Introductory Physics* and *Open Source Tutorials in Physics Sensemaking* often step students through a series of qualitative inferences. These series of inferences constitute a chain of reasoning that could be built into a chaining task, and differences in the quality of students' chains could be studied *before*, *during,* and *after* scaffolded instruction. That is, one could more formally explore the research question about the extent to which scaffolded materials aid students in developing context-specific reasoning skills.

Additionally, these scaffolded materials could be scrutinized for domain-general skills addressed in a context-specific way, such as the compensation reasoning difficulty addressed in the contexts of work and energy, bouyancy, and the ideal gas law. Then, a chaining task could be devised that isolated a compensation reasoning argument in a novel (and unfamiliar) context and student reasoning chains could be studied before and after relevant scaffolded instruction. Such an investigation to explore the extent to which addressing a domain-general reasoning skill on a context-specific basis leads to proficiency at that skill.

Another exciting avenue for future research is to craft reasoning elements that reveal information about students' coordination of resources while reasoning. Procedural resources have been identified for separation of variables in a physics context (Black & Wittmann, 2009), and other resources have been proposed in the contexts of Newton's 3rd law (Smith & Wittmann, 2008) and waves (Wittmann, 2006). There could be a way for these resources to be directly incorporated into a chaining task. Because chaining tasks can be implemented online, a large amount of data can be accumulated and analyzed fairly efficiently. Furthermore, with the capability of gathering time-dependent data on student construction of reasoning chains, it may be that patterns can emerge that corroborate the ideas put forward in literature regarding resource graphs. Other insights about the coordination of resources as students work through physics problems may also emerge.

The possibilities task has considerable potential for future development in a number of ways also. Along one dimension, the ability to represent many different mental models is linked with good reasoning skills (Johnson-Laird, 2006; Tishman, Jay, & Perkins, 1993; Lawson, 2004). It would be productive, therefore, to use the results of possibilities tasks, perhaps coupled with scores on the cognitive reflection test (Frederick, 2005; Wood, Galloway, & Hardy, 2016), to correlate the skill of generating hard-to-access models with success on physics problems (including those in existing concept inventories) that elicit a strong intuitive response. If there were positive correlations between these three factors, one might propose a possible direction for increasing student performance on such problems, by developing and implementing interventions expressly focused on increasing the tendency to explore (alternate) possibilities.

Another dimension for possibilities tasks to investigate is the link between documented conceptual difficulties and the cognitive structures used in reasoning. If, as proposed in Chapter 5, the tendency to treat momentum as a vector quantity is related to cognitive miserliness, the possibilities tasks can play a key role in better understanding that connection. Furthermore, there may be other contexts in which similar phenomena occur. Possibility tasks written for these contexts may be helpful in exploring the interplay between cognitive miserliness and the construction of particular cognitive constructs related to scalar vs. vector quantities.

Finally, the work presented in this dissertation is directly applicable to the development of a new generation of research-based materials that attend to domain-general reasoning mechanisms and bolster domain-general reasoning skills. The tasks presented here could be used to assess the productivity of such materials, but they could also play a key role in the instruction itself. For instance, new tutorials could target the exploration of alternate possibilities in a variety of contexts, thus giving students more practice with this skill. Chaining tasks could be used as a vehicle to discuss claim-evidence based reasoning (McNeill & Krajcik, 2008) or to examine the effects of salient distracting features on the use of specific lines of reasoning.

This latter goal of raising awareness of and addressing the impact of high-salience features on productive reasoning could perhaps be accomplished with chaining task modifications that ask a student to construct a line of reasoning leading to each of two answers and then asking them to reflect on which of the two answers seems more accurate based on (1) gut feeling and (2) quality of formal reasoning (similar to the Elby pairs (Elby, 2001) discussed in Chapter 2). Another way to address this may be to use "analytic intervention elements" followed by a series of follow-up questions that address the use or non-use of specific reasoning elements. Whatever the specific tactic employed, it seems that eliciting reflection on reasoning phenomena related to intuitive answers seems a promising avenue for future instructional materials that attend to student reasoning in physics more comprehensively (for instance, Elby, 2001 and Smith & Wittmann, 2007; see also Le, 2017).

REFERENCES

Beichner, R. (2007). The Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) Project. In Research-Based Reform of University Physics (Vol. 1). PER Central.

Beichner, R. (1994). Testing student interpretation of kinematics graphs. American Journal of Physics, 62, 750-762. doi:10.1119/1.17449

Black, K., & Wittmann, M. C. (2009). Procedural Resource Creation in Intermediate Mechanics. 2009 Physics Education Research Conference Proceedings. AIP. doi:10.1063/1.3290980

Bodin, M. (2012). Mapping university students' epistemic framing of computational physics using network analysis. Physical Review Special Topics - Physics Education Research, 8. doi:10.1103/physrevstper.8.010115

Braine, M. D., & O'Brien, D. P. (Eds.). (1998). Mental Logic. Psychology Press.

Brewe, E., et al. (2018). Toward a Neurobiological Basis for Understanding Learning in University Modeling Instruction Physics Courses. Frontiers in ICT, 5. doi:10.3389/fict.2018.00010

Christensen, W. M., & Thompson, J. R. (2012). Investigating graphical representations of slope and derivative without a physics context. Physical Review Special Topics - Physics Education Research, 8. doi:10.1103/physrevstper.8.023101

Close, H. G., & Heron, P. R. (2010). Research as a guide for improving student learning: An example from momentum conservation. American Journal of Physics, 78, 961-969. doi:10.1119/1.3421391

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. American Journal of Physics, 69, 970-977. doi:10.1119/1.1374249

diSessa, A. A. (1993). Toward an Epistemology of Physics. Cognition and Instruction, 10, 105-225. doi:10.1080/07370008.1985.9649008

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? International Journal of Science Education, 20, 1155-1191. doi:10.1080/0950069980201002

Docktor, J. L., et al. (2016). Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. Physical Review Physics Education Research, 12. doi:10.1103/physrevphyseducres.12.010130

Elby, A. (2000). What students learning of representations tells us about constructivism. The Journal of Mathematical Behavior, 19, 481-502. doi:10.1016/s0732-3123(01)00054-2

Elby, A. (2001). Helping physics students learn how to learn. American Journal of Physics, 69, S54--S64. doi:10.1119/1.1377283

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

Evans, J. S. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. Psychonomic Bulletin & Review, 13, 378-395. doi:10.3758/BF03193858

Evans, J. S. (2007). Hypothetical Thinking: Dual Processes in Reasoning and Judgement (Essays in Cognitive Psychology). Psychology Press.

Evans, J. S., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition. Perspectives on Psychological Science, 8, 223-241. doi:10.1177/1745691612460685

Finkelstein, N. D., & Pollock, S. J. (2005). Replicating and understanding successful innovations: Implementing tutorials in introductory physics. Physical Review Special Topics - Physics Education Research, 1. doi:10.1103/physrevstper.1.010101

Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486, 75-174. doi:10.1016/j.physrep.2009.11.002

Foti, N. J., Hughes, J. M., & Rockmore, D. N. (2011). Nonparametric Sparsification of Complex Multiscale Networks. (F. Rapallo, Ed.) PLoS ONE, 6, e16431. doi:10.1371/journal.pone.0016431

Frederick, S. (2005). Cognitive Reflection and Decision Making. Journal of Economic Perspectives, 19, 25-42. doi:10.1257/089533005775196732

Freeman, S., et al (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111, 8410-8415. doi:10.1073/pnas.1319030111

Galton, F. (1907). Vox Populi. Nature, 75, 450-451. doi:10.1038/075450a0

Gawronski, B., & Payne, B. K. (Eds.). (2010). Handbook of Implicit Social Cognition: Measurement, Theory, and Applications. GUILFORD PUBN.

Gentner, D., & Stevens, A. (1983). Mental Models (Cognitive Science Series). Lawrence Erlbaum Associates.

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. doi:10.1103/physrevphyseducres.14.010113

Gigerenzer, G. (2008). Why Heuristics Work. Perspectives on Psychological Science, 3, 20-29. doi:10.1111/j.1745-6916.2008.00058.x

Graham, T., & Berry, J. (1996). A hierarchical model of the development of student understanding of momentum. International Journal of Science Education, 18, 75-89. doi:10.1080/0950069960180107

Hammer, D. (2000). Student resources for learning introductory physics. American Journal of Physics, 68, S52--S59. doi:10.1119/1.19520

Hammer, D., Elby, A., Scherr, R., & Redish, E. F. (2005). Transfer of Learning: Research and Perspectives. In J. Mestre (Ed.). Information Age Publishing Inc.

Heckler, A. F. (2011). The Role of Automatic, Bottom-Up Processes: In the Ubiquitous Patterns of Incorrect Answers to Science Questions. In Psychology of Learning and Motivation (pp. 227-267). Elsevier. doi:10.1016/b978-0-12-387691-1.00008-9

Heckler, A. F., & Bogdan, A. M. (2018). Reasoning with alternative explanations in physics: The cognitive accessibility rule. Physical Review Physics Education Research, 14. doi:10.1103/physrevphyseducres.14.010120

Heckler, A. F., & Scaife, T. M. (2014). Patterns of Response Times and Response Choices to Science Questions: The Influence of Relative Processing Time. Cognitive Science, 39, 496-537. doi:10.1111/cogs.12166

Heron, P. R. (2004). Empirical investigations of learning and teaching, part I: Examining and interpreting student thinking. Proceedings of the International School of Physics Enrico Fermi, 156, 341-350. doi:10.3254/978-1-61499-012-3-341

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. doi:10.1103/physrevphyseducres.13.010131

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, 1191-1206. doi:10.1037/a0013025

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Kruglanski (Ed.), Social psychology: Handbook of basic principles (pp. 133-168). Guilford Press.

Hsu, L., Brewe, E., Foster, T. M., & Harper, K. A. (2004). Resource Letter RPS-1: Research in problem solving. American Journal of Physics, 72, 1147-1156. doi:10.1119/1.1763175

Ibrahim, B., Ding, L., Heckler, A. F., White, D. R., & Badeau, R. (2017). Students' conceptual performance on synthesis physics problems with varying mathematical complexity. Physical Review Physics Education Research, 13. doi:10.1103/physrevphyseducres.13.010133

Johnson, J. G., & Raab, M. (2003). Take The First: Option-generation and resulting choices. Organizational Behavior and Human Decision Processes, 91, 215-229. doi:10.1016/s0749-5978(03)00027-x

Johnson-Laird, P. (1983). Mental Models (Cognitive science series). Harvard University Press.

Johnson-Laird, P. (2009). How We Reason. Oxford University Press.

Kahneman, D. (2013). Thinking, Fast and Slow. Macmillan USA.

Kautz, C. H., Heron, P. R., Shaffer, P. S., & McDermott, L. C. (2005). Student understanding of the ideal gas law, Part II: A microscopic perspective. American Journal of Physics, 73, 1064-1071. doi:10.1119/1.2060715

Khemlani, S., & Johnson-Laird, P. N. (2016). How people differ in syllogistic reasoning. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Ed.), Proceedings of the 38th Annual Conference of the Cognitive Science Society.

Knight, R. D. (2016). Physics for Scientists and Engineers: A Strategic Approach with Modern Physics (4th Edition). Pearson.

Koponen, I. T. (2013). Systemic view of learning scientific concepts: A description in terms of directed graph model. Complexity, 19, 27-37. doi:10.1002/cplx.21474

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. doi:10.1103/physrevstper.10.020109

Kryjevskaia, M., Stetzer, M. R., & Heron, P. R. (2012). Student understanding of wave behavior at a boundary: The relationships among wavelength, propagation speed, and frequency. American Journal of Physics, 80, 339-347. doi:10.1119/1.3688220

Kryjevskaia, M., Stetzer, M. R., & Le, T. K. (2015). Failure to Engage: Examining the Impact of Metacognitive Interventions on Persistent Intuitive Reasoning Approaches. 2014 Physics Education Research Conference Proceedings. American Association of Physics Teachers. doi:10.1119/perc.2014.pr.032

Lawson, R. A., & McDermott, L. C. (1987). Student understanding of the work-energy and impulse-momentum theorems, *American Journal of Physics*, **55**, 811

Lawson, A. E. (2004). Reasoning and Brain Function. In The Nature of Reasoning. Cambridge University Press.

Lê, T. K. (2017). Using Contrasting Cases to Build Metacognitive Knowledge About the Impact of Salient Distracting Features in Physics Problems. Ph.D. dissertation, University of Maine.

Lindsey, B. A., Heron, P. R., & Shaffer, P. S. (2009). Student ability to apply the concepts of work and energy to extended systems. American Journal of Physics, 77, 999-1009. doi:10.1119/1.3183889

Loverude, M. E., Kautz, C. H., & Heron, P. R. (2002). Student understanding of the first law of thermodynamics: Relating work to the adiabatic compression of an ideal gas. American Journal of Physics, 70, 137-148. doi:10.1119/1.1417532

Loverude, M. E., Kautz, C. H., & Heron, P. R. (2003). Helping students develop an understanding of Archimedes' principle. I. Research on student understanding. American Journal of Physics, 71, 1178-1187. doi:10.1119/1.1607335

Madsen, A., Rouinfar, A., Larson, A. M., Loschky, L. C., & Rebello, N. S. (2013). Can short duration visual cues influence students' reasoning and eye movements in physics problems? Physical Review Special Topics - Physics Education Research, 9. doi:10.1103/physrevstper.9.020104

McCloskey, M. (1983). Mental models. In &. A. D. Gentner (Ed.). Erlbaum.

McDermott, L. C. (1991). Millikan Lecture 1990: What we teach and what is learned—Closing the gap. American Journal of Physics, 59, 301-315. doi:10.1119/1.16539

McDermott, L. C., (1995). Physics by Inquiry: An Introduction to Physics and the Physical Sciences, Volume 2. JOHN WILEY & SONS INC.

McDermott, L. C. (2001). Oersted Medal Lecture 2001: Physics Education Research—The Key to Student Learning. American Journal of Physics, 69, 1127-1137. doi:10.1119/1.1389280

McDermott, L. C., & Shaffer, P. S. (2001). Tutorials in Introductory Physics. Pearson College Div.

McDermott, L. C., Rosenquist, M. L., & van Zee, E. H. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. American Journal of Physics, 55, 503-513. doi:10.1119/1.15104

McLeod, P., Reed, N., & Dienes, Z. (2003). How fielders arrive in time to catch the ball. Nature, 426, 244-245. doi:10.1038/426244a

McNeill, K. L., & Krajcik, J. (2008). Inquiry and scientific explanations: Helping students use evidence and reasoning. Science as inquiry in the secondary setting, 121-134.

McPadden, D. (2018). Examining Students' Representation Choices in University Modeling Instruction. Ph.D. dissertation, Florida International University.

N.G.S.S. Lead States. (2013). Next Generation Science Standards. National Academies Press. doi:10.17226/18290

National Research Council. (2013). Adapting to a Changing World - Challenges and Opportunities in Undergraduate Physics Education. National Academies Press. doi:10.17226/18312

National Research Council, D. o., Sciences, S., Education, C., & Sensory Sciences Board on Behavioral, C. o. (2000). How People Learn. National Academies Press.

Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103, 8577-8582. doi:10.1073/pnas.0601602103

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2, 175-220. doi:10.1037/1089-2680.2.2.175

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education, 66, 211-227. doi:10.1002/sce.3730660207

Quinn, S., & Markovits, H. (1998). Conditional reasoning, causality, and the structure of semantic memory: strength of association as a predictive factor for content effects. Cognition, 68, B93--B101. doi:10.1016/s0010-0277(98)00053-5

Redish, E. F. (2004). A theoretical framework for physics education research: Modeling student thinking. Proceedings of the International School of Physics Enrico Fermi, 156, 1-63. doi:10.3254/978-1-61499-012-3-1

Redish, E. F., & Hammer, D. (2009). Reinventing college physics for biologists: Explicating an epistemological curriculum. American Journal of Physics, 77, 629-642. doi:10.1119/1.3119150

Rosiek, R., & Sajka, M. (2016). Eyetracking in Research on Physics Education. In Springer Proceedings in Physics (pp. 67-77). Springer International Publishing. doi:10.1007/978-3-319-44887-9_6

Saul, J. M., & Redish, E. F. (1997). Final Evaluation Report for FIPSE Grant #P116P50026: Evaluation of the Workshop Physics Dissemination Project .

Scherr, R. E. (2008). Gesture analysis for physics education researchers. Physical Review Special Topics - Physics Education Research, 4. doi:10.1103/physrevstper.4.010101

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. Psychological Review, 112, 610-628. doi:10.1037/0033-295x.112.3.610

Selden, A., & Selden, J. (2008). Overcoming students' difficulties in learning to understand and construct proofs. In Making the connection: Research and teaching in undergraduate mathematics. Mathematical Association of America.

Shaffer, D. M., Krauchunas, S. M., Eddy, M., & McBeath, M. K. (2004). How Dogs Navigate to Catch Frisbees. Psychological Science, 15, 437-441. doi:10.1111/j.0956-7976.2004.00698.x

Smith, T. I., & Wittmann, M. C. (2007). Comparing three methods for teaching Newton's third law. Physical Review Special Topics - Physics Education Research, 3. doi:10.1103/physrevstper.3.020105

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. doi:10.1103/physrevstper.4.020101

Sokoloff, D. R., & Thornton, R. K. (1997). Using interactive lecture demonstrations to create an active learning environment. The Physics Teacher, 35, 340-347. doi:10.1119/1.2344715

Speirs, J. C., Ferm Jr., W. N., Stetzer, M. R., & Lindsey, B. A. (2016). Probing Student Ability to Construct Reasoning Chains: A New Methodology. 2016 Physics Education Research Conference Proceedings. American Association of Physics Teachers. doi:10.1119/perc.2016.pr.077

Stanovich, K. E. (2010). What Intelligence Tests Miss. Yale University Press.

Sternberg, R. J. (2004). What do We Know about Reasoning? In The Nature of Reasoning. Cambridge University Press.

Surowiecki, J. (2004). The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Doubleday.

Susac, A., Bubic, A., Martinjak, P., Planinic, M., & Palmovic, M. (2017). Graphical representations of data improve student understanding of measurement and uncertainty: An eye-tracking study. Physical Review Physics Education Research, 13. doi:10.1103/physrevphyseducres.13.020125

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In In two minds: Dual processes and beyond (pp. 171-196). Oxford University Press. doi:10.1093/acprof:oso/9780199230167.003.0008

Tishman, S., Jay, E., & Perkins, D. N. (1993). Teaching thinking dispositions: From transmission to enculturation. Theory Into Practice, 32, 147-153. doi:10.1080/00405849309543590

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5, 207-232. doi:10.1016/0010-0285(73)90033-9

Wason, P. C. (1968). Reasoning about a rule. Quarterly Journal of Experimental Psychology, 20, 273-281. doi:10.1080/14640746808400161

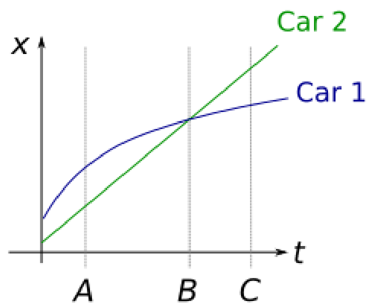
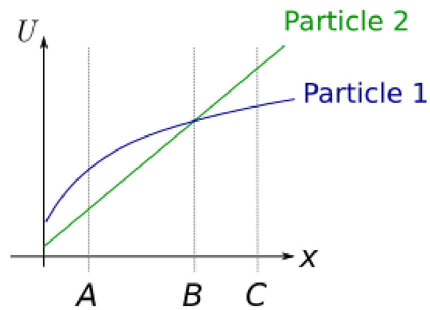
Wittmann, M. C. (2006). Using resource graphs to represent conceptual change. Physical Review Special Topics - Physics Education Research, 2. doi:10.1103/physrevstper.2.020105

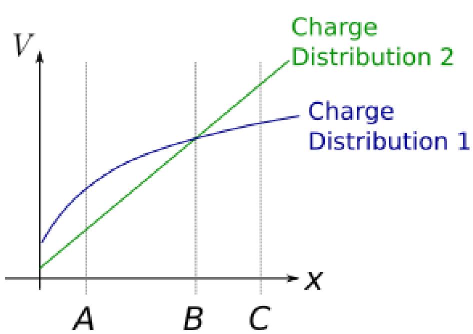
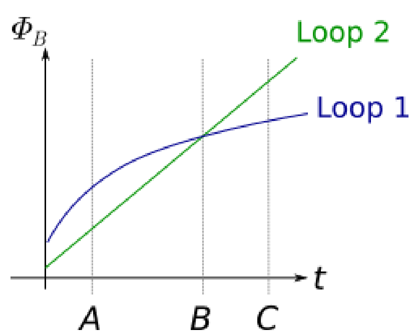
Wittmann, M. C., Steinberg, R. N., & Redish, E. F. (2004). Activity-Based Tutorials: Introductory Physics, The Physics Suite. Wiley.

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. doi:10.1103/physrevphyseducres.12.023101

APPENDIX A: ISOMORPHIC GRAPH TASKS

A.1: Task statements

Task	Kinematics Graph Task	Potential Energy Graph Task
Figure		
Task Statement	The motions of two cars are described by the position vs. time graphs shown above. When, if ever, are the magnitudes of the velocities (i.e., the speeds) of the cars the same?	The potential energy of system 1, in which only particle 1 can move, is described by the potential energy vs. position graph shown. Likewise, the potential energy of system 2, in which only particle 2 can move, is shown. The two systems don't interact. Where, if anywhere, are the magnitudes of the forces on the particles the same?

Task	Electric Potential Graph Task	Magnetic Flux Task
Figure		
Task Statement	The electric potentials set up by two charge distributions located far away from each other are described by the electric potential vs. position graphs shown above. Where, if anywhere, are the magnitudes of the electric fields due to the charge distributions the same?	The magnetic fluxes through two different conducting loops in different magnetic fields are described by the magnetic flux vs. time graphs shown above. When, if ever, are the absolute values of the induced EMF's ($\mathcal{E}_1$ and $\mathcal{E}_2$) the same?

A.2: Reasoning Elements Provided

In consultation with the members of the advisory committee and external collaborators, the elements were refined as the project continued. The network analysis described in Chapter 4 was conducted on an earlier data set based on a longer list of elements. In accordance with feedback from members of the advisory committee and other external collaborators, the element list was subsequently refined and shortened. This refined list was used for the investigation of phenomena related to dual-process theories of reasoning and decision-making documented in Chapter 3.

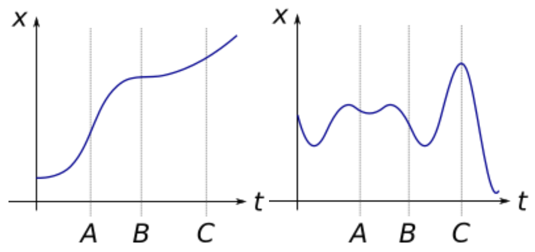
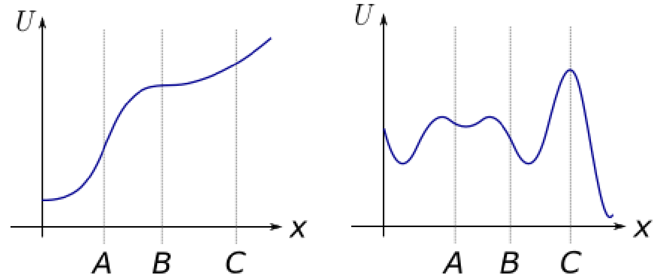
Elements used for network analysis task (Chapter 4):

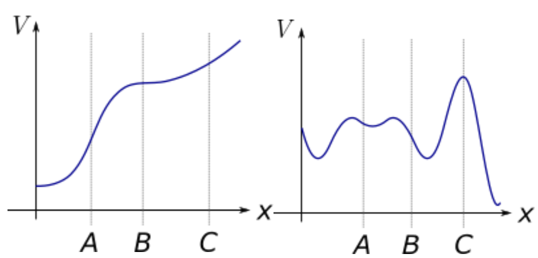
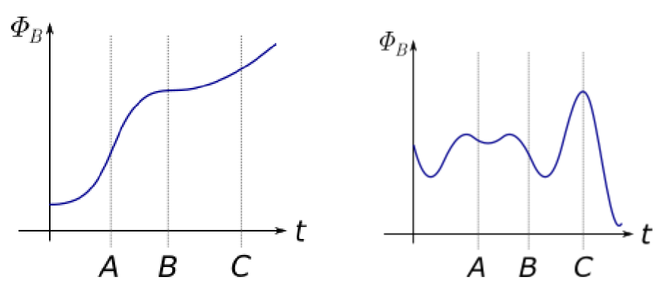
	Kinematics Reasoning Elements	Potential Energy Reasoning Elements	Electric Potential Reasoning Elements	Magnetic Flux Reasoning Elements
1	$x(t_f) = x_0 + \int_0^{t_f} v(t)dt$	$U(x_f) = U_0 + \int_0^{x_f} \vec{F}(x) \cdot d\vec{x}$	$V(x_f) = V_0 + \int_0^{x_f} E(x)dx$	$\Phi_B = -\int_0^{t_f} \mathcal{E}dt$
2	$v(t_f) = v_0 + \int_0^{t_f} a(t)dt$	$p(t_f) = p_0 + \int_0^{t_f} F(t)dt$	$U(x_f) = U_0 + \int_0^{x_f} F(x)dx$	$\mathcal{E} = -\int_0^{s_f} \vec{E} \cdot d\vec{s}$
3	$v = \frac{dx}{dt}$	$F = -\frac{dU}{dx}$	$E = -\frac{dV}{dx}$	$\mathcal{E} = -\frac{d\Phi_B}{dt}$
4	$a = \frac{dv}{dt}$	$F = \frac{dp}{dt}$	$F = -\frac{dU}{dx}$	$E = -\frac{d\mathcal{E}}{ds}$
5	the integral, $\int_a^b f(x)dx$, is the area under the graph of *f* vs. *x*	the integral, $\int_a^b f(x)dx$, is the area under the graph of *f* vs. *x*	the integral, $\int_a^b f(x)dx$, is the area under the graph of *f* vs. *x*	the integral, $\int_a^b f(x)dx$, is the area under the graph of *f* vs. *x*
6	the derivative, df/dx, at a specific point is the slope of the tangent line of the *f* vs. *x* graph at that point	the derivative, df/dx, at a specific point is the slope of the tangent line of the *f* vs. *x* graph at that point	the derivative, df/dx, at a specific point is the slope of the tangent line of the *f* vs. *x* graph at that point	the derivative, df/dx, at a specific point is the slope of the tangent line of the *f* vs. *x* graph at that point
7	slope of a position vs. time graph is the velocity	the negative of the slope of a potential energy vs. position graph is the force	the slope of an electric *potential* vs. position graph gives the magnitude of the electric field	slope of a magnetic flux vs. time graph is the magnitude of the induced EMF
8	slope of a velocity vs. time graph is the acceleration	slope of a momentum vs. time graph is the force	the slope of an electric *potential energy* vs. position graph is the force	slope of an induced EMF vs. position graph is the magnitude of the induced electric field
9	area under a velocity vs. time graph is the displacement	area under a force vs. position graph is the work done by the force, which is the negative of the change in potential energy	area under an electric field vs. position graph is the electric potential	area under an induced EMF vs. time graph is the magnitude of the change in magnetic flux
10	area under an acceleration vs. time graph is the change in velocity	area under a force vs. time graph is the change in momentum (or the impulse)	area under a force vs. position graph is the electric potential energy	area under an induced electric field vs. position graph is the magnitude of the change in the induced EMF
11	the lines intersect at time B	the lines intersect at position B	the lines intersect at position B	the lines intersect at time B
12	slopes are the same at time A	slopes are the same at position A	slopes are the same at position A	slopes are the same at time A
13	the speeds are the same at time A	the magnitudes of the forces are the same at position A	the magnitudes of the electric fields are the same at position A	the magnitudes of the induced EMF's are the same at time A
14	the speeds are the same at time B	the magnitudes of the forces are the same at position B	the magnitudes of the electric fields are the same at position B	the magnitudes of the induced EMF's are the same at time B
15	the speeds are the same at time C	the magnitudes of the forces are the same at position C	the magnitudes of the electric fields are the same at position C	the magnitudes of the induced EMF's are the same at time C
16	the speeds are never the same	the magnitudes of the forces are never the same	the magnitudes of the electric fields are never the same	the magnitudes of the induced EMF's are never the same

Elements used in investigation of phenomena related to dual-process theories of reasoning (Chapter 3):

Kinematics Reasoning Elements	Potential Energy Reasoning Elements	Electric Potential Reasoning Elements	Magnetic Flux Reasoning Elements
$\Delta x_{t_1 \to t_2} = \int_{t_1}^{t_2} v\, dt$	$\Delta U_{a \to b} = \int_a^b \vec{F}(x) \cdot d\vec{x}$	$\Delta V_{a \to b} = -\int_a^b \vec{E}(x) \cdot d\vec{x}$	$\Delta \Phi_{B, t_1 \to t_2} = -\int_{t_1}^{t_2} \mathcal{E}(t) dt$
$v = \frac{dx}{dt}$	$F = -\frac{dU}{dx}$	$E = -\frac{dV}{dx}$	$\mathcal{E} = -\frac{d\Phi_B}{dt}$
the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*	the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*	the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*	the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*
the derivative, $\frac{dh(r)}{dr}$, at a specific point is the slope of the tangent line of the *h(r) vs. r* graph at that point	the derivative, $\frac{dh(r)}{dr}$, at a specific point is the slope of the tangent line of the *h(r) vs. r* graph at that point	the derivative, $\frac{dh(r)}{dr}$, at a specific point is the slope of the tangent line of the *h(r) vs. r* graph at that point	the derivative, $\frac{dh(r)}{dr}$, at a specific point is the slope of the tangent line of the *h(r) vs. r* graph at that point
velocity is given by the value of the slope of a position vs. time graph	force is given by the negative of the value of the slope of a potential energy vs. position graph	electric field is given by the negative of the value of the slope of an electric potential vs. position graph	induced EMF is given by the negative of the value of the slope of a magnetic flux vs. time graph
displacement is given by the area under a velocity vs. time graph	change in potential energy is given by the negative of the area under an electric field vs. position graph	change in electric potential is given by the negative of the area under an electric field vs. position graph	change in magnetic flux is given by the negative of the area under an induced EMF vs. time graph.
the lines intersect at time B	the lines intersect at time B	the lines intersect at time B	the lines intersect at time B
slopes are the same at time A	slopes are the same at time A	slopes are the same at time A	slopes are the same at time A

A.3: Screening Question Task Statements

Task	Kinematics Screening Questions	Potential Energy Screening Questions
Figures		
Task Statement	The motion of a car is described by the position vs. time graph shown above. At which of the three labeled times is the magnitude of the velocity (i.e., the speed) of the car the greatest?	The potential energy of a system in which only one particle can move is described by the potential energy vs. position graph shown. At which of the three labeled positions is the magnitude of the force on the particle the greatest?

Task	Electric Potential Screening Questions	Magnetic Flux Screening Questions
Figure		
Task Statement	The electric potential set up by a charge distribution is described by the electric potential vs. position graph shown above. At which of the three labeled positions is the magnitude of the electric field due to the charge distribution the greatest?	The magnetic flux through a conducting loop is described by the magnetic flux vs. time graph shown above. At which of the three labeled positions is the absolute value of the induced EMF the greatest?

BIOGRAPHY OF THE AUTHOR

J. Caleb Speirs is a Research Fellow and Doctoral Candidate in the Physics and Astronomy Department at the University of Maine. He has a M.S. in Applied Physics and a B.S. in Engineering Physics from the Colorado School of Mines, having done work in the field of scanning laser microscopy. He has taught physics at various community colleges in the greater Denver area, at College of the Atlantic in Bar Harbor, Maine, and is currently employed as an Assistant Lecturer at the University of New England in Biddeford, Maine.

He and his wonderful wife, Ellen, have three daughters, $\sum_{\text{girls}} age < 8\ \text{yrs}$, with $\text{MEAN}(age) = 2.4\ \text{yrs}$, $\text{STDEV}(age) = 2\ \text{yrs}$, and $\text{MAX}(age) = 4\ \text{yrs}$. They are interested in dance parties, cutting vegetables, and pretending colored pencils have personalities, and are learning to take breaks when angry, apologize when hurtful, and forgive when hurt. They are his life.

J. Caleb Speirs is a candidate for the Doctor of Philosophy degree in Physics from the University of Maine in May 2019.