

ASSESSING CAUSAL MECHANISTIC REASONING:
PROMOTING SYSTEM THINKING

By

Paul J. Weinberg

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Learning, Teaching, and Diversity

August, 2012

Nashville, Tennessee

Approved:

Professor Leona Schauble

Professor Rogers Hall

Professor Richard Lehrer

Professor Sun-Joo Cho

Professor Phillip Crooke

TABLE OF CONTENTS

ACKNOWLEDGMENTS	v
LIST OF TABLES	vi
LIST OF FIGURES	viii
NOTATION	ix
I. INTRODUCTION	1
Children’s Resources for Causal Inferences about Mechanics	1
Mechanistic Reasoning about Simple Machines	5
Assessment of Mechanistic Reasoning about Simple Machines	10
II. CONSTRUCT MAP DEVELOPMENT AND ITEM DESIGN: MECHANISTIC REASONING ABOUT SIMPLE LEVERED MACHINES.....	14
Specifying the Progress Variable	14
Item Design	23
Cognitive Interviews	25
Developing Scoring Exemplars.....	27
III. METHOD	28
Participants	28
Procedure	33
Conduct of the Interview	36

Analysis	36
Analysis of Items	38
Descriptive Item Statistics	39
Item Analysis with Classical Test Theory (CTT) and Item Response Theory (IRT)	40
Reliability and Validity	45
IV. RESULTS	48
Descriptive Item Statistics	49
Item Analysis with Classical Test Theory (CTT) and Item Response Theory (IRT)	53
Reliability and Validity	67
Causal Mechanistic Tracing and Machine Characteristics	79
V. DISCUSSION	82
Children’s Causal Mechanistic Reasoning	82
Assessment Development (Research Question #1)	83
Next Design Iteration	84
Additional Forms of Reasoning to be Assessed	85
The Stability of Mechanistic Reasoning (Research Question #2)	86
System Tracing	87
APPENDIX A: EXEMPLAR EXAMPLES	89
APPENDIX B: TABLES	96

REFERENCES 110

ACKNOWLEDGMENTS

This work would not have been possible without the financial support of the National Science Foundation, who supported the work it is based on with Grant Number 0733209. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

I thank Sarah Weinberg, Leona Schauble, Richard Lehrer, Rogers Hall, Sun-Joo Cho, Phillip Crook, and Rob Rouse.

LIST OF TABLES

Table	Page
2-1: Construct map: Causal mechanistic tracing.....	16
2-2: Item coverage matrix.....	96
3-1: Participants	29
3-2: Demographic information for Centennial Public School District.....	30
3-3: Demographic information for Wordsworth Academy	30
3-4: Demographic information for University Lab Academy	31
3-5: Demographic information for Private Research University.....	31
3-6: Demographic information for Liberal Arts College	32
3-7: Demographic information for Public University.....	32
3-8: The old and new rules of measurement (Embretson, 1996, p. 342).....	41
4-1: Descriptive and classical test theory (CTT) statistics.....	99
4-2: Item difficulty.....	53
4-3: Item discrimination estimates	54
4-4: Item difficulty estimates and standard errors.....	57
4-5: MNSQ fit statistic for each item	61
4-6: Interpretation of parameter-level mean-square fit statistics (Wright & Linacre, 1994)	62
4-7: Percentage of mechanistic elements (scored on items) compared with how they were coded in the cognitive interviews	71
4-8: Item thresholds	73
4-9: Tracing by machine characteristics.....	81

B-1: Item-step estimates and standard errors	100
B-2: Person ability estimates and standard errors	105

LIST OF FIGURES

Figure	Page
2-1: Correct prediction by mechanistic element.....	21
3-1: Camera positioning for cognitive interviews.....	34
4-1a: Score distribution (HFPO).....	51
4-1b: Score distribution (MPA2).....	51
4-1c: Score distribution (STD1).....	51
4-1d: Score distribution (MPD1).....	52
4-1e: Score distribution (STA3').....	52
4-2: Item Wright Map.....	56
4-3: Fit plots for two items.....	63
4-4: Fit plot for item Sequential Tracing-D1'.....	64
4-5: Item-step Wright Map.....	66
4-6: Scatter plot: Person ability estimates v. standard error of measurement (SEM).....	69
4-7a: Item response.....	76
4-7b. Interview gesture.....	76

NOTATION

y_{ij} indicates the response for person j on item i .

Indices

j for persons, $j = 1, \dots, J$;

i for items, $i = 1, \dots, I$;

CHAPTER I

INTRODUCTION

Children's Resources for Causal Inferences about Mechanics

The workings of the physical world are governed by many principles that seem to be understood by children from very early ages. For instance, multiple solid objects cannot occupy the same space, or move through one another; in addition, any changes in their movements are the result of internal or external forces (e.g., collisions, gravity). Some features of children's early causal reasoning suggests its potential as a resource for understanding the mechanisms of physical systems.

The research literature on infants' conceptions of the properties of physical objects has significantly increased in the past two decades and demonstrates that even infants have intuitions about cause and effect relationships that help them anticipate how their physical environments work (Baillargeon, 1994). For instance, one class of study addresses young children's intuitions about object permanence and solidity. By their fourth month, infants anticipate that solid objects cannot interpenetrate and continue to exist over space and time, even when out of sight. Baillargeon (1987a; Baillargeon, Spelke, & Wasserman, 1985) presented infants with a flat barrier swinging through 180-degrees of rotation on a surface, such as a table. Infants watched the 180-degree event several times until they became habituated to the stimulus. They were then shown a small, solid object placed behind the barrier in a way that would prevent the barrier from swinging through the full 180-degrees. Infants looked longer at displays in which the barrier appeared to travel through the full 180-

degrees (i.e., impossibly going through the object) than at displays in which the barrier stopped at a place consistent with assumptions of object solidity and permanence. These findings seem to suggest that infants were surprised by the impossible condition. In other research, infants looked longer when a vertically dropped object seemed to end up in a position that implied it must have moved through an intervening, but occluded solid platform, indicating that they had seen an event that violated their expectations (Spelke, 1991).

Although notions like the solidity of objects are in place very early, understanding of the physical world continues to develop over time. For example, young infants are sensitive only to large and obvious conflicts between the barrier and obscured block and do not notice smaller discrepancies, such as when the barrier stops 30 degrees too early (Baillargeon, 1995). Thus, they have not calibrated the geometry of physical events with their consequences. Similarly, Spelke and Kyeong (1992) showed that the development of an appreciation that an unsupported object will fall down also takes time. Irrespective of how quickly infants gain intuitions about the natural and physical world, there is agreement that, by the end of the first year, they have causal expectations consistent with many principles that govern the behaviors of physical objects.

In spite of these results, there are good reasons to doubt that these causal attributions about mechanical systems that infants exhibit are identical to those of adults. Infants' knowledge is probably intuitive and implicit, and is almost certainly not available to reflection in the same way that an older child's knowledge is. Researchers have tried to determine how a one-year-old's mental representations of the world can best be characterized. Leslie (1984) has investigated whether an infant's ability to anticipate the causal behaviors of physical objects can be interpreted as having beliefs like those of an adult. Ongoing research may clarify not

only how older children's physical knowledge becomes more explicit but also how that explicit knowledge interacts with earlier intuitive forms. Leslie (1982) has raised the possibility that young children have predisposed sensitivities to the behavior of physical objects that bear little relation to how adults make causal attributions.

Given infants' rich causal sensitivities about the properties of physical systems, it may seem unusual, but it is not altogether unexpected, that research has concluded that adults have difficulty making causal attributions about the mechanisms of similar systems. In particular adults have difficulty reconciling their intuitions about cause and effect with the forms of mechanistic explanations valued by disciplines. For example, Carmazza, McCloskey, and Green (1981) showed that college students typically did not correctly predict the trajectory of a metal ball suspended by a string, moving in an arc as a pendulum, after the string was cut, even when these students had taken college level physics courses. This gap between findings of early competency and later struggle suggests that our accounts of the development of this form of thinking are incomplete. Carmazza and colleagues argue that adults hold consistent and erroneous beliefs about the physical world, and that many of these beliefs are highly resistant to change by instruction. Much of that literature, especially in mechanics, has focused on high school and college students (Carmazza, McCloskey, & Green, 1981; Clement, 1982; Minstrell, 1983). There have been many fewer studies of younger preschool or elementary schoolchildren (Ioannides & Vosniadou, 2002).

There may, however, be more connection between the early infancy research and the later "misconceptions" research than has been acknowledged. For one thing, viewing adult forms of reasoning as misconceptions may be misleading (Smith, diSessa, & Roschelle, 1993). diSessa (1993), for example, argues that everyday physics is better thought of as both a large

and diverse number of low-level explanatory components that are evoked in different contexts. He further characterizes “misconceptions” as perfectly valid ideas that are used in inappropriate contexts. In addition, there are many areas where students’ causal intuitions accord with formal principles of mechanical reasoning. For example, although students may not perceive the balance of forces when considering a book resting on a table, they do when considering a book resting on an outstretched hand. Thus, neither students nor adults are as devoid of positive intuitions about cause as the misconceptions literature suggests.

In general, little of the literature addresses the forms of reasoning about causal mechanism within STEM disciplines that emerge between the time when individuals enter school (i.e., age five) and the time they exit (i.e., high school and adulthood). Understanding the development of these forms of reasoning requires more than an understanding of mere beginning and endpoints. Moreover, good instruction for elementary and middle school students should capitalize on their naïve causal reasoning about these physical principles as educators engage students in causal explanations of physical mechanisms. To achieve a longer-term portrait of causal reasoning about mechanisms, this dissertation study focuses on the reasoning of elementary school students, middle school students, high school students, lay adults, college undergraduates not enrolled in the hard sciences, and college undergraduates majoring in engineering. All of these populations were asked to make predictions (i.e., causal attributions) about the mechanisms behind the motion of levered machines. Levered machines are ubiquitous in the designed world (and, as well, in states’ science standards). The predictions were used as data for validating an assessment system that characterizes how people think about inspectable machines.

Mechanistic Reasoning about Simple Machines

Mechanistic explanations are clear descriptions of how systems work based on physical laws at specified levels of description (Glennan, 2002). When reasoning about levered systems, mechanistic reasoning indicates an understanding of the “mechanisms” involved in the transmission of force through the system components; the transmission of forces occurs through the push-pull interactions of the connected levers. This research focuses on levered machines because they provide access to an important form of disciplinary reasoning in science and engineering (i.e., mechanistic reasoning) through their “simplicity” (i.e., all machine parts and mechanisms are visible and inspectable). Causal reasoning about mechanism, or “mechanistic reasoning,” is essential to understanding the ways things work in both natural and designed systems; accordingly, mechanistic explanation is an important practice to promote in STEM (Science, Technology, Engineering, and Mathematics) disciplines. Mechanistic explanations are considered complete when they have “bottomed out”; that is, when descriptions of lower-level mechanisms are irrelevant to the explainer’s current goals or interests (Machamer, Darden, & Craver, 2000). For instance, the fields of molecular biology and neurobiology do not typically regress to the quantum level of description to explain chemical bonding. It is important for students (as well as adults) to learn to reason at levels that are appropriate for their particular goals.

Mechanistic reasoning involves not only associating causes with effects within specific systems (in particular domains) but also describing the processes responsible for these associations (Shultz, 1982). By focusing on the processes that produce cause-effect relationships, mechanistic explanations take into account how component entities affect one

another (Machamer, Darnden, & Carver, 2000). Machamer, et al., describe domain general mechanistic causal schemas as “descriptions of mechanisms [that] exhibit productive continuity without gaps from the set up to terminal conditions” (p. 3).

In this study, causal mechanism is addressed within specific domains, systems, and at particular levels of description. The study’s foci are the resources individuals have for mechanistic reasoning, how these resources can be coordinated, and the extent to which this can be assessed within one inspectable system.

Although children have resources for making causal attributions, they often ignore the very features of a mechanism that are critical to its function. This noticing (or lack of noticing) is context specific (diSessa, 1993), and the features of the mechanisms children tend to perceive are critical to their ability to reason mechanistically. Individuals tend to reason about causal mechanism in ways that are local and contextually driven. Thus, Bolger, Kobiela, Weinberg, and Lehrer (2012) have described “elements of mechanistic reasoning” that are specific to the system we have been investigating. In this case, we have been studying learning as students build simple toys that operate with inputs and outputs via combinations of levers. The diagnosis and causal tracing of these mechanistic elements (i.e., of these levered systems) from input to output comprise a complete causal explanation of this system (i.e., Machamer et al., 2000). However, one cannot necessarily perceive and diagnose the motion of a lever in one context and immediately perceive and diagnose all other levers’ motions in all contexts where they appear. Disciplining one’s perception to “see” in such a way as to diagnose lever motion, across lever types and arrangements, no doubt takes time and experience (Stevens & Hall, 1998). With time and experience, these systems (i.e., simple levered machines) can become more generally useful for diagnosing and explaining the many simple and complex levered

machines in the designed world. For example, two levers and a screw are the constituent parts of a pair of scissors; bicycles and eggbeaters are other common examples of compound machines. Developing an understanding of the mechanisms within this system will hopefully aid children in understanding the causal mechanisms of many more systems, including both simple and compound machines.

A predisposition for seeking out causal mechanism in the natural and designed world is valuable for inquiry in the STEM disciplines. There is evidence that focusing on mechanism is central to children's development of capacities to engage in scientific explanation and argumentation (Bolger et al. 2012; Russ, Sherr, & Hammer, 2009).

Regardless of the literature about early competencies, our previous studies (Bolger et al., 2012; Bolger, Weinberg, Kobiela, Rouse, Lehrer, 2011; Kobiela, Bolger, Weinberg, Rouse, Lehrer, 2011) indicate that even for simple systems, such as those we have investigated, constructing coherent mechanistic explanations is not trivial. Even though children have resources for these forms of reasoning under specific conditions (e.g., within developmental psychology studies where the materials are constructed to operate in obvious ways), they may not necessarily deploy these forms of reasoning spontaneously in unfamiliar or untutored contexts; this is the case because these forms of reasoning are heavily dependent on the cues that are provided (and attended to) when participants are considering which components are relevant to systems functioning. Many characteristics of simple machines, including aspects of their appearance, embeddedness within other components, and possibly other attributes as well, will likely influence the difficulty with which individuals can diagnose and causally trace their mechanisms.

In spite of the multitude of simple machines in our culture, most individuals continue to find explaining their causal mechanism challenging (Bolger et al., 2012; Bolger et al., 2011; Lehrer & Schauble, 1998; Metz, 1985; 1991). In their studies of reasoning about gear trains, Lehrer and Schauble as well as Metz explored children's tendencies to use causal reasoning to explain visible mechanisms (i.e., in these cases, gears). Lehrer and Schauble compared explanations of gear trains by children in grades 2 and 5 and found that both age groups knew that contact between gears determined output direction and speed of the final gear in a train. In addition, some children also mentioned the push and pull interaction among teeth of the gears to explain the gears' coordinated motion. These explanations included causal mechanistic tracings of the trains from input to output, involving gears, gear size, and the connection of gear teeth. However, most children did not mention the gear teeth when explaining why gears must be connected to transmit motion. In fact, even when asked, a significant number of the younger children provided no mechanism at all to explain the motion of the gear trains. For example, when children were asked to explain the motion of machines like eggbeaters (Lehrer & Schauble, 1998) they tended not to examine the machines and ignored the intermediate components (e.g., the gears). Children rarely traced the transmission of motion from the input through each of these intermediate parts to the output. In addition, students more frequently noticed structural features of the machines (e.g., how relevant parts of the system of levers were organized without attention to motion) rather than how motion was transmitted within the system. Even when presented with opportunities to manipulate mechanical components, children did not tend to spontaneously search for mechanistic explanations. They infrequently explained the mechanism governing the causal relations that they discerned. Instead, children

most frequently noticed the perceptual features of the linkages (e.g., overall appearance, individual parts or motions).

In our previous work on toys with levers (Bolger et al., 2012; Bolger et al., 2011) most students had difficulty explaining the mechanism of the system with which they worked. In order to both develop and support children's mechanistic reasoning we focused on mechanical linkages consisting of inputs and outputs (i.e., systems of levers and connecting pivots/fulcrums) with no components hidden from view. The intent was to make the linkages accessible to children, and, presumably, to provide optimal candidates for eliciting mechanistic reasoning and explanation. We developed and verified mechanistic elements from both our own sense of their workings and those of professionals in engineering and physics. The elements, were (a) *linked direction* (i.e., attention to the coordinated direction of the input and output of a linkage, "When you push the input up, the output goes down"), (b) *rotation* (i.e., attention to the rotary motion of the levers, "The output goes around"), (c) *lever arms* (i.e., attention to the coordinated opposite motion of the two lever arms, "When this side [of the lever] goes up, this side goes down"), and (d) *constraint via the fixed pivot* (i.e., attention to the causal relation between the pivot being fixed to the board and the resultant motion, "Because the brad is stuck to the board, the link is going to go that way"). These mechanistic elements accorded with the children's predictions and explanations of the machines' motions. We characterized mechanistic reasoning as comprised of these mechanistic elements for these simple levered systems (Bolger et al., 2012). Although all children's explanations demonstrated some of these elements of mechanistic reasoning, few were able to orchestrate these elements to constitute "causal mechanistic tracing." Causal mechanistic tracing was determined according to analysis that determined whether children sequentially animated the

components of linkages. First, it was determined whether they referred sequentially, in talk or gesture, to each of the links in linkages. Second, it was determined whether a correct determination, in talk or gesture, was expressed for the direction of motion for each component in the sequence. Within a tracing episode, children must have diagnosed all mechanistic elements.

One student consistently supported the construction of such a complete mechanistic explanation through: (a) latching mechanistic elements together (e.g., how the *constraint via the fixed pivot* caused the *rotation* of the output link), (b) using linking words (e.g., *because*, *and so*, and *then so*) to coordinate entities and properties (e.g., “Because there’s two fixed brads this time (*constraint via the fixed pivot*), and so when you push it, these [the fixed pivots] will stay and and then so they’ll [the two horizontal links] kind of twist” (*rotation*), and (c) the gesturally tracing the motion of the system (Bolger et al., 2012). This performance, although rare in our sample, suggests that tracing is not out of reach of elementary-aged students.

Assessment of Mechanistic Reasoning about Simple Machines

The purpose of the current study was to develop an assessment instrument that is capable of characterizing mechanistic reasoning about basic mechanical systems (i.e., simple levered machines). Ideally, this assessment will play a role in the characterization and investigation of children’s mechanistic reasoning without a complete reliance on one-on-one interviews, which require a significant amount of time, personnel, and materials to conduct.

There are presently no assessments that (1) leverage children’s early capacities to reason causally about properties of mechanical objects and (2) promote a highly profitable

disciplinary form of reasoning. At this time the most widely used assessment of ideas about force and motion is the Force Concept Inventory (FCI) (Hestenes, Wells, & Swackhamer, 1992). This instrument assesses how well high school and college students are prepared for introductory physics courses. It qualitatively discriminates between students who hold Newtonian compared with more naïve conceptions of mechanical force. The FCI takes a top-down perspective on physics instruction. That is, it measures how closely students' conceptions accord with those of Newtonian principles by asking students to reason about those principles in the context of real world situations. For example, the FCI assesses individuals' understandings of Newton's Third Law in the context of a collision of two marbles in terms of a "conflict metaphor" in action. In contrast, the assessment proposed here takes a different approach to understanding ideas about the properties of mechanical objects. That is, it tracks individuals' abilities to mechanistically parse systems of simple machines, characterizing their forms of reasoning as they are observed without trying to fit them into a Newtonian framework. This assessment leverages children's early capacities to make sense of forces such as pushes and pulls, force vectors, and geometry as an opportunity to develop their mechanical knowledge.

From this perspective, introducing students to general mechanical principles through the causal mechanistic tracing of these simple systems may provide a foundation for the building of important knowledge about mechanical objects; for example, using vectors to consider force and motion problems.

The assessment instrument being developed in this study assesses an individual's use of the following mechanistic elements (i.e., machine mechanisms) common to many simple and compound machines: (1) *linked direction*, (2) *rotation*, (3) *lever arms*, and (4) *constraint via*

the fixed pivot. The instrument assesses children’s ability to diagnose these mechanistic elements as they attempt to predict and explain the operation of these levered systems from input to output.

This assessment is being designed using IRT modeling that was outlined by Mark Wilson and the Berkeley Education Assessment Research (BEAR) Center. In psychometrics, item response theory (IRT), also known as latent trait theory, strong true score theory, or modern mental test theory, is a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables. It is based on the application of related mathematical models to testing data.

Design of the assessment system was guided by Wilson’s (2005) “building blocks” for developing assessments. Wilson proposes that an assessment instrument should be precipitated by a theory of the structure and progression through the knowledge, ideas, or reasoning to be measured. This theoretical structure/progression is referred to as a construct or a progress variable. The construct is the first building block of the system. This study investigates one progress variable (i.e., causal mechanistic tracing) and defines learning performances that represent benchmarks of knowledge and skill.

The remaining three building blocks include item design, the development of “exemplars” (i.e., scoring guides that are specific to different construct benchmarks), and the modeling of participant responses.

Research Questions

I hope to answer the following research questions through the development and administration of this assessment: (1) Can mechanistic reasoning be assessed via a standard

assessment instrument? And, (2) Can this assessment provide insight into the features of machines that are most likely to disrupt an individual's capacity to reason mechanistically?

Considering the resources even young children have to reason about causal mechanism, this assessment diagnoses which elements (Bolger et al., 2012) students are able to employ, how frequently they employ them, and what kinds of machine characteristics make mechanistic reasoning difficult. In addition, it assesses the extent to which participants can causally trace the mechanistic elements from a machine's input to output (i.e., participant uses and causally connects all four elements). If the assessment system can measure children's abilities to diagnose the mechanistic elements of these simple inspectable machines, it may be profitable to next attempt to extend the system to a wider variety of machines (e.g., compound machines), because of the ubiquity of both simple and compound machines in our culture.

An important feature of this assessment is its capacity to make qualitative distinctions between individuals who do and do not causally connect mechanistic elements (e.g., "Because [causally connecting the mechanistic elements of constraint via the fixed pivot and rotation] the pivot is fixed to the board [*constraint via the fixed pivot*], this link will rotate [*rotation*]) as they explain the motion of these simple systems. This distinction has not been fully explored in our previous research because of the small number of individuals who causally connected the mechanistic elements. The larger sample explored here may provide further information about what makes this kind of thinking challenging.

CHAPTER II

CONSTRUCT MAP DEVELOPMENT AND ITEM DESIGN: MECHANISTIC REASONING ABOUT SIMPLE LEVERED MACHINES

The development of an assessment begins with three steps (i.e., “building blocks”; Wilson, 2005). First, the progress variable or construct is specified. The most important features of a construct are: (1) a coherent and substantive definition of the content of the construct; and (2) an ordering of ability or skill levels from lowest to highest. Research and theoretical considerations are presented to support these levels. Next, the item design is described. Finally, item “exemplars” are introduced. Item exemplars are scoring guides that map item responses to the construct map.

Specifying the Progress Variable

An assessment design begins with the specification of the construct or progress variable (Wilson, 2005). The specification of this progress variable is based on an analysis of the literature on reasoning about the workings of simple mechanical systems. This research focused on what was more and less difficult for children to understand when diagnosing these systems in one-on-one interviews (Metz, 1985; 1991; Lehrer & Schauble, 1998; Bolger et al., 2012; Bolger et al., 2011). The analysis resulted in a distinct dimension of mechanistic reasoning about levered systems (in both instructional and everyday contexts), which is called “causal mechanistic tracing.” This construct refers specifically to an individual’s mechanistic

parsing of simple levered systems. It includes the four mechanistic elements described in Bolger et al. (2012) (*linked direction, rotation, lever arms, and constraint via the fixed pivot*) in addition to one additional element, *tracing*. These elements are ordered as levels in the construct. The most sophisticated level, *tracing*, requires that a participant correctly diagnose and causally connect all mechanistic elements within a machine, from input to output, without gaps. The five construct levels, descriptions, and examples are listed in Table 2-1.

Table 2-1

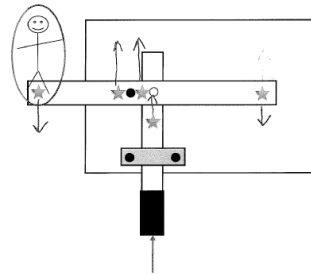
Construct map: Causal mechanistic tracing.

Level	Mechanistic Element	Mechanistic Element Descriptions	Mechanistic Element Example
5	Tracing	Participant predicts all the mechanistic elements correctly.	
4	Constraint via the Fixed Pivot	Participant correctly draws the opposite and/or rotary motion of the two closest points on opposite sides of the fixed pivot.	
3	Lever Arms	Participant draws arrows with opposite directions from stars on opposite sides of a lever's arms.	
2	Rotation	Participant draws arced paths (they may show the incorrect direction). However, the location of these paths must reasonably approximate	

fractions of circles either centered around the fixed or floating pivot(s).

1 **Linked Directions**

Participant draws the correct motion of input and output(s).



The structure of this construct is motivated by previous research conducted with interview methodologies (Bolger et al., 2012). All the construct levels (i.e., *linked direction*, *rotation*, *lever arms*, *constraint via the fixed pivot*, and *tracing*) were designed to capture, within this paper-and-pencil assessment, how students had previously reasoned about the physical systems in the interview contexts. In the previous research, students found it more difficult to recognize some of these mechanistic elements, in comparison to others. In addition, it was more difficult for students to causally trace through the machines (from input to output), citing all the mechanistic elements, than to simply mention each of the elements, without fully explaining how motion was transmitted through the elements to predict the motion of the machines.

The relative difficulty of mechanistic elements. This construct, which is called “causal mechanistic tracing,” indicates the differential difficulty of diagnosing machine motion with respect to the four mechanistic elements, based on the frequency with which they were cited within student explanations as they described and explained the motion of the machines (Bolger et al., 2012). In addition, the ordering of the construct levels is based on theoretical considerations concerning the machines’ workings. These two considerations for the ordering of the mechanistic elements will be presented. The mechanistic elements are ordered from least to most difficult as follows: *linked direction*, *rotation*, *lever arms*, and *constraint via the fixed pivot*. The mechanistic elements that are most related to correct prediction were not necessarily those cited most frequently by students. For example, *constraint via the fixed pivot* is less frequently used in explanations than the other mechanistic elements, but is highly associated with correct prediction. This is because the fixed pivot is the principal mechanism responsible

for lever motion. If a participant understands how the fixed pivot constrains the motion of the link, predicting its direction requires only noticing where the input is being applied.

Frequency of mechanistic elements cited within children's explanations. The construct levels follow from the overall frequency of mechanistic elements in children's explanations in our previous work (Bolger et al., 2012; Bolger et al., 2011) and are determined by considering both each element's salience and its difficulty. In these studies, elements could be cited in the explanation of every machine. More frequently cited elements were presumed to be easier to notice than less frequently cited elements; we did not count the citation of a specific mechanistic element more than once per machine. Of course, it is possible that failure to mention a mechanistic element may have been due to considering it obvious. In addition, it could be argued that a mechanistic element may be less salient, but not necessarily more difficult. In these prior studies, these possibilities were minimized by the design of the flexible interviews and the machines. For example, students were probed in order to elicit unspoken understandings about mechanistic elements. In addition, all of the machines were composed of visibly connected levers so that all their mechanistic elements were easily inspectable.

Linked direction was the element most frequently cited for describing or explaining machine motion (Bolger et al., 2012). It was used to explain the motion in 46% of the machines. *Rotation, lever arms, and constraint via the fixed pivot* were cited in fewer than 20% of machines. This order of relative frequency was also reflected in data from a second previous study (Bolger et al., 2011) in which pretest data were analyzed from three students who exhibited a range in their ability to predict the output direction, given a specified input. These cases similarly showed that *linked direction* was the most frequently cited element, followed by *rotation, lever arms, and constraint via the fixed pivot*.

Frequency of mechanistic elements cited within children's predictions. The ordering of the mechanistic elements within the construct (that is, our conjectures about their relative difficulties) is also informed by the frequency with which they were cited by children to correctly predict machine motion in previous research. Our previous work showed that students' reasoning about different mechanistic elements (Bolger et al., 2012) enabled them to make correct predictions of machine motion (i.e., the direction of the output for a given input). For instance, if a child's explanation (either before or after moving the machine) did not include any mechanistic elements, he was unlikely to predict correctly, because these mechanistic elements constitute the components of a mechanistic explanation of the motion of these machines. Without an understanding of any of a machine's mechanisms, it is unlikely (though not impossible) that a participant would be able to correctly make predictions about its output motion. It is also unlikely that an individual would make incorrect predictions using multiple mechanistic elements. Figure 2-1 illustrates the percentage of student performances that corresponded with a correct or incorrect machine prediction. It shows that in the research that preceded this study, correct predictions were associated most strongly with mentioning two of the elements of mechanistic reasoning, *linked direction* and *constraint via the fixed pivot*. Reasoning about *rotation* was also observed more often with correct prediction (Bolger et al., 2012).

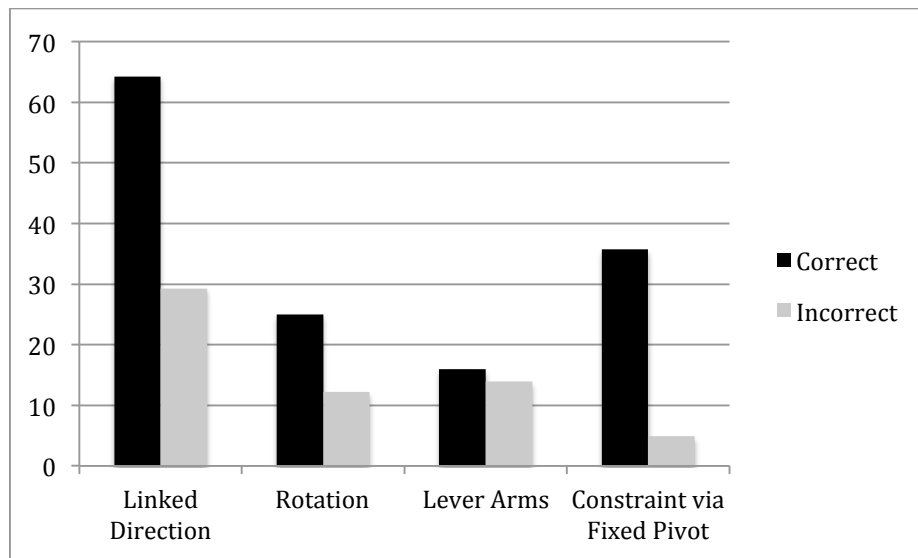


Figure 2-1. Correct prediction by mechanistic element.

Theoretical considerations. In addition to results from studies conducted with students in one-on-one interviews (Bolger et al., 2012; Bolger et al., 2011), the ordering of the elements is also influenced by a conceptual analysis of the cognitive demands of variants of these tasks. The difficulty orderings of the mechanistic elements were theorized as follows: *Linked direction* is the easiest element because it requires people to simply notice the direction and causal coordination of the input and output links, without referring to a specific path. Of the three mechanistic elements that require individuals to recognize the causal coordination of the links' motion within the system (i.e., *linked direction*, *lever arms*, and *constraint via the fixed pivot*), the cause-and-effect relationship between the direction of the input and output seems to be the simplest because the input and output are the machine's most salient components (Lehrer & Schauble, 1998; Metz, 1991). A child must simply notice the motion of these components and attribute a correct direction. To reason about the mechanistic element of *rotation*, participants need to notice the paths, and not just the endpoints (Piaget, Inhelder, &

Szeminska, 1960) of lever motion (i.e., that they are rotary and not linear). Perceiving a link's path may require greater focus and short-term memory than simply noticing inputs and outputs. It is more difficult to reason about *lever arms* than *rotation* because in order to reason about *lever arms*, participants must be able to view the causal coordination of the two lever arms (i.e., on both sides of the fixed pivot) from both a global and local perspective. It is difficult to view both the causally coordinated global motion of the lever and the local directed motion of each lever arm simultaneously. *Constraint via the fixed pivot* is hypothesized to be the most difficult mechanistic element to recognize. Those assessed at the level of *constraint via the fixed pivot* can do more than perceive regularities in machine motion; they understand the causal relationship between: (1) the presence of a fixed pivot, (2) the resultant constraint, and (3) the subsequent lever motion.

The leveling of the construct presumes that those participants who can diagnose the more difficult elements (e.g., *constraint via the fixed pivot*) should also be able to diagnose those that are less difficult (e.g., *linked direction*). For example, a participant who explains machine motion by referring to the *constraint of the fixed pivot* (i.e., the most difficult mechanistic element) should also be able to explain the role of the other three less difficult elements. This conjecture is consistent with findings from our previous research.

Tracing. To be assessed at the level of *tracing*, the highest level, participants must reason about machine motion by explaining the role of all of the mechanistic elements, from input to output. This does not mean simply diagnosing them, but also causally connecting them by tracing the motion of the machine from input to output; that is, indicating how pushes and pulls are transmitted across the machine from one element to the next. In previous research, we have observed only a small number of students reasoning like this (Bolger et al., 2012).

Once the progress variable (i.e., construct) has been specified, items can be developed.

Item Design

After the construct levels and associated performances were specified, twenty-seven assessment items were developed to elicit performances specific to each of the levels. Items are the smallest scored unit in an assessment. It is assumed that the performance on a specific pool of items should generalize to (i.e., serve as a sample of) a universe of items assessing the same construct levels (Kane, 1992). To more validly assess each construct level, an item pool should include multiple items at every level of the construct (Wilson, 2005).

Item format. Item format is critical to item design. Multiple-choice items are preferable when the nature of anticipated responses is clear and limited in scope (Briggs, Alonzo, Schwab, & Wilson, 2006). In addition, multiple-choice items may be preferable when the respondents are not skilled at communicating their thinking through writing; this can often be the case with younger participants. On the other hand, multiple-choice items are also subject to guessing or testing strategies (Martinez, 1999). Open-ended, short-answer questions are more appropriate when the anticipated responses are more complex or less clear. Both multiple-choice and minimally demanding short-answer responses (e.g., which require respondents to draw predicted motion) have been used because of the greater ease younger individuals have in responding to these items (i.e., as opposed to free response items that require participants to express complicated ideas through extensive writing). Moreover, the short-answer response items were motivated by research showing that it is easier to distinguish between different construct levels in open-ended than in multiple-choice items (Heuvel-Panhuizen, 1994). In

multiple-choice items, the distinctions that can most easily be drawn are between knowing and not knowing. Assessments that use drawing to make inferences about learning, as most of the items in this study do, have not been fully explored or exploited in the field (Quellmalz, Timms, & Schneider, 2012). However, there is a tradition within developmental and cognitive psychology of using children's drawings as ways of understanding how they are thinking.

Balanced, meaningful, and worthwhile assessments. It is ideal for different item responses to span multiple construct levels within the same item, so that a respondent's thinking can be more precisely diagnosed (Heuvel-Panhuizen, 1994). Each item can be designed to generate responses that span as many construct levels (at minimum two) as are present in the total number of constructs (Wilson, 2005). It is even possible to create items that generate responses that span the entire construct (i.e., apply to each level). For example, items have been developed (e.g., Causal Mechanistic Tracing Item- A1, see Appendix B) that can be classified on all five levels on the construct map, based on: (1) which element(s) a participant diagnoses and (2) whether he causally connects the mechanistic elements from input to output. The use of a construct map to guide item design ensures adequate construct coverage.

Heuvel-Panhuizen (1994) describes the importance of including a variety of problem formats (e.g., multiple-choice, short free-response) for a balanced assessment.

An assessment is judged to be "meaningful and worthwhile" based on item content and presentation. Heuvel-Panhuizen described problem format (e.g., interview item, free response item, multiple choice item) as a relevant consideration when designing an item. The item format may make reasoning tasks more or less accessible to participants. The available item formats and task contexts are further constrained by the assessment media (e.g., paper and pencil items). One might wonder whether paper and pencil items are an effective media for

assessing participants' causal mechanistic reasoning about the motion of pegboard levered machines. Research by Hagerty (1992; 2004) and Schwartz (1995; 1999; Schwartz & Black, 1996a; 1996b; 1999) seems to suggest that they can be. In research that asked respondents to reason about diagrams drawn on paper, Hagerty and Schwartz found that mechanical reasoning by mental simulation is analogous to the physical processes that are being simulated. However, in their work they were not demanding written responses. In the present research, care has been taken to preserve within the paper and pencil context the important elements of the levered machines about which the participants will reason. However, the problem context has undeniably changed, to some extent, what the participant is reasoning about, and therefore, there may be effects that are unforeseen.

Cognitive Interviews

Many of the paper and pencil items were adapted from interview questions used in the three studies previously conducted (Bolger et al., 2012; Bolger et al., 2011) with pegboard linkages. In some cases, flexible interviews (Ginsburg, Jacobs, & Lopez, 1998) established that the claims that can be made about responses to these interview items could also be extended to the current paper and pencil format. For example, when individuals drew rotary paths around fixed pivots on paper and pencil items, their talk and gesture were typically consistent with responses that would have been coded in our previous studies as reflecting an understanding of the mechanistic element of *rotation*.

In spite of their advantages (e.g., efficiency, scalability), paper and pencil tasks are limited in the information they can provide about an individual's knowledge. In some cases, it

may be difficult to determine how participant reasoning motivates the responses. To gain a better understanding of what the items assess, pilot cognitive interviews were conducted as participants worked with these items to increase our understanding of what the items assessed. For example, if a participant drew an arced path that was approximately centered around a fixed pivot, it was nonetheless important to confirm that he was reasoning about rotary motion. The purpose of the interviews was to provide evidence to confirm the construct validity for each mechanistic element.

In these interviews, participants explained their reasoning as they responded to the assessment items (i.e., think-alouds). After they had completed the item, they were asked again to explain their reasoning. Then any remaining clarifying questions were asked. The conduct of these and cognitive interviews conducted during the study is further described in the Method section.

Earlier small scale pilot studies suggest that paper and pencil items can be reliable predictors of how individuals would predict the output motion for a given input on actual pegboard machines. In these pilot studies, participants were presented with one sheet of paper showing representations of six pegboard machines. The directions read: “For each of the mechanisms below, draw an arrow showing how each little person [a small figure drawn to mark the machine’s output] would move if you PUSHED UP on the blue handle [tape indicating the machine’s input] (just like the red arrow shows).” A key indicated the difference between fixed and floating pivots. After students made these paper and pencil predictions, they went on to predict the motion of the pegboard machines in the one-on-one interviews. Of the 266 total predictions made, 77% (n=204) were consistent across paper items and machine predictions. The mean consistency across students was 76% (median=82%, mode=83%,

$SD=21\%$). This suggests that these paper items provide reliable data about how students perform when reasoning about real pegboard machines within flexible interviews.

Developing Scoring Exemplars

Once items were developed, scoring “exemplars,” (Wilson, 2005) (i.e., scoring guides that relate item responses to the construct map) were created. They qualitatively describe and provide concrete examples of all potential types of participant responses for each item, and associate the responses with different construct levels. An item’s exemplar, structured like the construct map, is ordered from the least to most sophisticated response. However, in the exemplar, only those construct levels relevant to that particular item are represented.

These exemplars contain a minimum of three scoring categories: (1) the construct linkage code (i.e., scores for responses that link to the construct); (2) the no link code (i.e., scores for responses that do not link to the construct), and (3) a missing code (i.e., a score for missing responses). The exemplars for constructed response items often include more than one construct linkage code (i.e., multiple different responses that map to different construct levels). There are 25 items in which *linked direction* could be scored, 23 items in which *rotation* could be scored, 15 items in which *lever arms* could be scored, 16 items in which *constraint via the fixed pivot* could be scored, and 14 items in which *tracing* could be scored. Table 2-2, in Appendix C, shows the construct coverage for each developed item.

CHAPTER III

METHOD

Participants

Table 3-1 shows the participant groups that comprise the sample. College undergraduates and private high school students were originally included to ensure complete construct coverage. However, as is explained in greater detail below, there was little variability in many of these individuals' responses, because of a ceiling effect with the assessment instrument. All these students scored at the highest level on at least all but 2 items. Removing these individuals was desirable because these twenty-eight participants, 20% of the total sample, scored at the highest construct level on almost every item where this was possible. This made the most difficult mechanistic element appear easier than other elements because of the high frequency of responses at this level. In order to increase the variability of responses, these 28 participants were excluded from analysis.

Table 3-1 <i>Participants</i>		
Participants	Number	Number Included
	Enrolled	in Analysis
Elementary School Students	28 (female=17)	28 (female=17)
Middle School Students	25 (female=16)	25 (female=16)
High School Students	23 (female=5)	20 (female=4)
University Undergraduates (Non-Science majors)	26 (female=19)	16 (female=13)
University Undergraduates (Engineering majors)	28 (female=8)	13 (female=5)
Adults (without college education)	10 (female=8)	10 (female=8)

The elementary, middle, and high school students come from public and private schools in the southeastern United States. The university undergraduates come from three universities, two in the southeastern (one is private) and one in the mid-western (which is private) United States. Of the two universities in the southeastern United States, one is a highly ranked private university and the other is a large lower ranked public university. The university in the mid-western United States is a highly ranked private liberal arts college.

These public elementary, middle, and high schools belong to Centennial Public School District (a pseudonym). Demographic information is presented in Table 3-2. The percent of children attending these three schools qualifying for free or reduced lunch ranges between 60 to 90 from year to year.

Table 3-2 <i>Demographic information for Centennial Public School District.</i>	
Race/Ethnicity	Percentage
Caucasian	34%
African-American	48%
Hispanic	14%
Asian	3%
Other	1%

The demographic information for one of the two private schools, Wordsworth Academy (a pseudonym), is presented in Table 3-3.

Table 3-3 <i>Demographic information for Wordsworth Academy.</i>	
Race/Ethnicity	Percentage
Caucasian	92%
African-American	7%
Hispanic	1%
Asian	1%
Other	0%

The demographic information for the other private school, University Lab Academy (a pseudonym), is presented in Table 3-4.

Table 3-4 <i>Demographic information for University Lab Academy.</i>	
Race/Ethnicity	Percentage
Caucasian	75%
African-American	11%
Hispanic	1%
Asian	7%
Other	6%

The undergraduate majors come from courses at the three universities as well as through snowball recruitment. The demographic information is presented in Tables 3-5, 3-6, and 3-7.

Table 3-5 <i>Demographic information for Private Research University.</i>	
Race/Ethnicity	Percentage
Caucasian	60%
African-American	8%
Hispanic	4%

Table 3-6	
<i>Demographic information for Private Liberal Arts College.</i>	
Race/Ethnicity	Percentage
Caucasian	76%
African-American	5%
Hispanic	5%
Asian	8%
Other	6%

Table 3-7	
<i>Demographic information for Public University.</i>	
Race/Ethnicity	Percentage
Caucasian	78%
African-American	15%
Hispanic	2%
Asian	3%
Other	1%

The adults without college degrees (n=10) are 10% Caucasian and 90% African-American.

Individuals in the study represent various ethnic backgrounds and life experiences. Participants were deliberately chosen to represent a very wide variety of ages and experience levels. The purpose was to ensure that the entire range of the conjectured construct would be adequately represented with participant responses. This is important because it is impossible to assess the difficulty of the highest construct levels without responses at these levels. The participants in elementary, middle, and high school, according to their teachers, represent a

wide spectrum of academic achievement. The undergraduates, both engineering majors and non-science majors, represent students along a continuum of academic success (i.e., from less to more highly rated universities). I conjectured that the engineering majors would perform well on this assessment because of the benefit of their academic engineering training. The adults without college degrees are likely from different populations than the other adults in the study (i.e., college undergraduates) and their experiences are also likely different. This population should increase the diversity of item responses as well as the resources they draw upon as they reason about the assessment items.

Twenty-eight private high school students (n=3) and undergraduate adults (n=25; engineering majors=15) were removed from the sample before IRT analysis was conducted because there was little variability in their responses. All these students scored at the highest level on at least all but 2 items. Removing these individuals was necessary because these twenty-eight participants, 20% of the total sample, scored at the highest construct level on almost every item where this was possible. This made *tracing* appear easier than other elements because of the high frequency of responses at this level. In order to increase the variability of responses, these 28 participants were excluded from analysis.

Procedure

Some of the third grade students (n=13) took the assessment after participating in instruction oriented around the design and construction of levered toys. These students were included in order to populate the sample with individuals who had some experience reasoning

about levered machines. The rest of the sample took the assessment uncoupled from any particular kind of instructional treatment.

Each participant responded to a cognitive interview while they worked on each item. The total interview (i.e., assessment administration) was completed during one day and lasted an average of 37.5 minutes (ranging from 17 minutes to 78 minutes). Interviews were recorded using one camera, with a table microphone. The camera was positioned at the side of the participant, about one half foot away from the table, angled down to capture what he was looking at as well as gestures he made over the paper; the camera view is shown in Figure 3-1. Interview sessions were digitally rendered for further analysis.

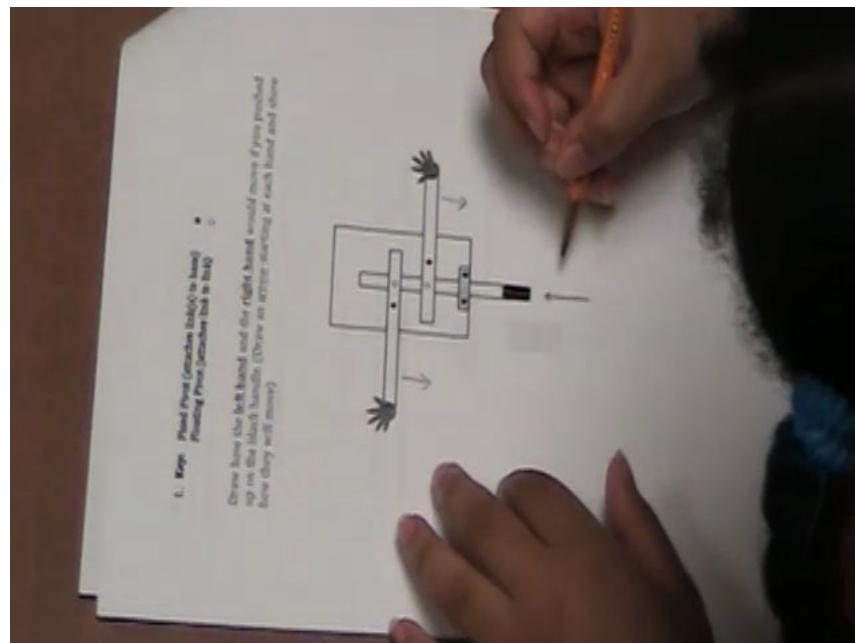


Figure 3-1. Camera positioning for cognitive interviews.

Participants who did not take part in the instructional sequence about the design and construction of levered toys were given a five-minute introduction. The respondents were shown how a levered machine could be built with brads and linkages made from pegboard

(with researcher assistance). Participants were provided with two links, a pegboard, and brads, and then guided through the process of making a fixed and floating pivot. The purpose was to ensure that participants were familiar with the relevant materials and vocabulary (e.g., fixed pivot, floating pivot) before proceeding to the next phase of the interview, in which they responded to paper and pencil items that were based on the pegboard linkages.

The paper and pencil items were presented to participants across seven forms to maximize the number of items to which participants could respond. There were three common items (or link items) across all forms so that scores and item difficulties across the forms could be compared. Link items are central to the process of test equating in IRT. Herein, the interchangeable use of alternate test forms is built to the same content and statistical specifications because scores and item difficulties based on different sets of items must be placed on a common scale. Various data collection designs and analytical procedures have been used for test equating, but for this assessment, the method relied on included items common to all test forms. This is an example of the “common-item nonequivalent groups design” (Kolen & Brennan, 1995).

Elementary and middle school students completed ten items per form, while high school students, undergraduates, and non-college educated adults completed fifteen items per form. Five items were indicated in each form that elementary and middle school students were instructed to skip.

Four item categories were included on each assessment: (1) Lever Arms Prediction (one or two items), (2) Rotation Constraint (one item), (3) Machine Prediction (six or seven items), and (4) Sequential Tracing (six or seven items). These categories were sampled, without replacement, to determine the items and their order on the assessment forms. Each

form began with the same relatively easy link item. This was done to check for independence of assessment items by comparing them to this first item.

Conduct of the Interview

While participants responded to each item, they were asked to: (1) read the problem aloud and (2) think aloud as they read and responded to each item. When the participant completed the item, he was asked to explain again, if necessary, the rationale for the observed item response based on interviewer probes. Finally, participants were asked to report any words that they found confusing, as well as whether there was any confusion about what the item was asking. The interview was conducted in this order to: (1) determine spontaneous thinking throughout participant experience with each item (i.e., think-aloud); (2) assess mechanistic reasoning that was present, but possibly not elicited during the think aloud (i.e., retrospective explanations) with interviewer probes; and (3) assess item validity by checking participant comprehension of both item instructions and the nature of item tasks.

Analysis

Scoring items. Each item was scored according to the exemplar developed for it. The exemplar levels were transformed into raw scores. Ten percent of the total items were scored by an outside researcher. The agreement was 85%.

Coding talk and gesture. In order to check the construct validity of the exemplars, participant talk and gesture were coded according to the analytic framework used in previous

studies to code these mechanistic elements as participants noticed, described, and explained the motion of the pegboard linkages (Bolger et al., 2012). The purpose was to determine whether the hypothesized item outcome space was consistent with the actual participant explanations of the items. More importantly, it indicated whether the exemplars were sufficient for assessing participants' mechanistic reasoning in a way that was consistent with our previous work.

Defining and coding episodes. A participant's work on one item is defined as a "performance." Each performance was broken up into "codable instances" according to the two interview phases (i.e., "think-alouds," conducted as participants worked with the materials and "retrospective reflections" determined after participants had made their initial judgments). The codeable instance is a logical unit of analysis because it captures the mechanistic elements each participant cites when responding to the two interview phases. Ericsson and Simon (1993) showed that providing simultaneous verbalizations (e.g., think-alouds) provides more consistent verbal reports of participant thinking than retrospective reports (e.g., student rationales for their response) because in retrospective reporting people generate inferences to fill out and generalize incomplete or missing memories. The written responses to the items were compared with what was coded during the think-alouds. Finally, the retrospective reflections provide insight into the resources participants may have for using mechanistic elements that might not have been captured by their written responses or the think-alouds. Thus, these responses provided data for an additional validity check of the exemplars. All codeable instances were coded using Nvivo 9.0 software.

Participants' assessments were coded according to the exemplars and compared with the coding of their talk and gesture (according to the analytic framework) while responding to

the items. This showed whether there were consistent patterns across item responses and talk and gesture.

Ten percent of the total instances were coded by an outside researcher. The agreement was 82%.

Analysis of Items

Item initial analysis provided substantial information for revising the items. The following was determined: (1) the extent to which item responses populated all of the construct levels and (2) the consistency between the mechanistic elements that individuals used to respond to the items and those hypothesized in the exemplars.

Construct coverage. The items elicited responses that covered the entire construct. However, six items were removed from the sample before IRT analysis (and after the data from the 28 high-responding students had been removed) because although they targeted important mechanistic elements, either: (1) the majority of the sample scored at the highest level(s) (4 items) or (2) the distribution of responses was bimodal, with the vast majority of responses divided between the highest and lowest levels (2 items). When the majority of responses are at the same level, items provide limited information about that level in relation to the entire scale. Similarly, in items where the majority of responses are at either the highest or the lowest level, these items provide little information about the intermediate levels. The omitted items are being redesigned for future administration.

Exemplar adequacy. All item responses could be scored according to the exemplar levels.

Descriptive Item Statistics

The means of the item scores and variances were calculated and reported.

Mean of item i . This is the average of the scores of an item for all respondents. It is calculated by dividing the sum of item scores of all respondents ($\sum_j^J y_{ij}$) by the total number of respondents (J) as follows:

$$\mu_i = \frac{\sum_j^J y_{ij}}{J}.$$

The item mean is a rough measure of performance on that item across all participants.

Variance of item i . The variance of the above mean can be calculated by summing the squares of differences between the item mean and the item score from each individual respondent, and then dividing that sum by the total number of respondents as follows:

$$\sigma_i^2 = \frac{\sum_j^J (y_{ij} - \mu_i)^2}{J}.$$

The variance tells us how variable participant performance was on this item, on the average.

Item Analysis with Classical Test Theory (CTT) and Item Response Theory (IRT)

To model the data from respondents, both CTT and IRT were considered. Although the two modeling approaches are generally consistent and complementary, there are a number of points of difference: (1) measurement errors can be obtained for each person in IRT (i.e., each person ability estimate has a standard error of measurement); (2) IRT provides several improvements in scaling items and people. IRT models scale the difficulty of items and the ability of people on the same latent continuum (thus, the difficulty of an item and the ability score of a person can be meaningfully compared); and (3) another improvement provided by IRT is that the parameters of IRT models are generally not sample- or test-dependent when the dimensionality is assumed to be the same across samples or tests, whereas true-score is defined in CTT in the context of a specific test. Embertson (1996) has compared important points of difference between CTT and IRT. These differences are summarized in Table 3-8. For the purpose of comparison, statistics were calculated from both CTT and IRT approaches.

Table 3-8

The Old and New Rules of Measurement (Embretson, 1996, p. 342)

CTT	IRT
1. The standard error of measurement applies to all scores in a particular population.	The standard error of measurement differs across scores, but generalizes across populations.
2. Comparing test scores across multiple forms depends on test parallelism or adequate equating.	Comparing scores from multiple forms is optimal when test difficulty levels vary across persons.
3. Unbiased assessment of item properties depends on representative samples from the population.	Unbiased estimates of item properties may be obtained from unrepresentative samples.
4. Meaningful scale scores are obtained by comparisons of position in a score distribution.	Meaningful scale scores are obtained by comparisons of distances from various items.

Classical test theory (CTT). Two item indices based on CTT are presented to analyze the item characteristics: (1) item difficulty and (2) item discrimination.

Item difficulty. This assessment contains both dichotomous and polytomous items. When an item is dichotomously scored its difficulty equals its mean item score (i.e., the proportion of respondents who answer the item correctly). When an item is polytomously scored its difficulty calculation is adjusted by dividing the mean item score by the difference between the possible maximum and minimum scores, so that its result will be on a scale similar to that of the dichotomously scored items:

$$p_i = \frac{\mu_i}{\mu_{\max} - \mu_{\min}}$$

Item discrimination. Item discrimination indicates how effectively an item discriminates between respondents who are relatively high on the criterion of interest and those who are relatively low. Different measures of item discrimination are available. Some, such as the index of discrimination, are applicable only to dichotomously scored items. The Pearson product moment correlation was used to measure the item discrimination because many polytomously scored items as well as dichotomously scored items were included in the assessment instrument:

$$r_{i(X-i)} = \frac{\sigma_{i(Y-i)}}{\sigma_i \sigma_{(Y-i)}}$$

Here σ_i is the sample variance of item i , $\sigma_{(Y-i)}$ is the sample variance of the total score excluding item i , and $\sigma_{i(Y-i)}$ is the covariance between individual item scores and total test scores.

Item response theory (IRT). Two Wright Maps obtained from a Rasch model for polytomous item responses, called a partial credit model (Masters, 1982), were generated: (1) the Item Map and (2) the Item-step Map. These Wright maps are used to characterize item difficulty decomposed into the average item location across steps and steps for each item. The item map presents the person's ability score on the same scale as the average item location across steps for each item; the item-step map presents the person's ability score on the same scale as each step for each item. In addition, the mean squared (MNSQ) statistics were calculated in order to investigate the item fit.

Although Rasch analysis characterizes items with respect to their difficulty, the notion of item difficulty in IRT analysis differs from that in classical item analysis based on CTT. In the classical item analysis, the item difficulty is essentially a characteristic of the observed scores; in IRT, the item difficulty is parameterized in the model. In IRT, the item difficulty and person ability score are placed on the same latent continuum. For example, an item's difficulty corresponds with the ability score of persons who have a 0.5 probability of correctly answering the item.

A one-dimensional PCM was used to fit the data. This model was specified based on the following considerations:

1. Dimensionality: The domain of causal mechanistic tracing is hypothesized to be one dimension, as shown in the construct map.

2. Scoring categories: This assessment contains polytomous items. The polytomous categories are ordered, but without the assumption of equal distance between adjacent categories.

The PCM was conducted using the ConQuest software (Wu, Adams, Wilson, & Haldane, 2007).

Wright map. On a Wright Map, a vertical line is marked out in logits; person estimates and item locations are positioned on the left- and right-hand sides, respectively, of the vertical line. The zero point of the logit scale is where $\theta_j - \beta_i = 0$. A person's ability in logits is his natural log odds for succeeding on items that are chosen to define the "zero" point of the scale; and an item's difficulty in logits is its natural log odds for eliciting failure from persons with "zero" ability. The closer to the bottom of the Wright Map, the less capable the respondent and the less difficult the item; the reverse is true at the top of the Wright Map.

The item location is the item difficulty (i.e., the point at which the item difficulty corresponds to the ability score of persons who have 0.5 probability of correctly answering the item). For polytomously scored item, the item locations are thresholds for reaching successive response categories. In other words, the item locations for polytomously response categories indicate the ability score of persons who are more likely to reach level k once they reach level $k - 1$. Graphically, the item locations are the point at which the item response function curves of two adjacent response categories (e.g., 0 vs. 1, 1 vs. 2) cross.

The item thresholds that are plotted on the Wright Map, however, are Thurstone thresholds. Thurstone thresholds are cumulative; a threshold is the point at which the probability of responding below a category is equal to responding in or above that category. For example, for a four category item (i.e., 0, 1, 2, 3), the Thurstone threshold for score category 2 is the point at which participants are as likely to be observed below 2 as being observed in or above 2 (i.e., 0 & 1 vs. 2 & 3); the Thurstone threshold for score category 3 is the point at which participants have a 0.5 probability of responding in 3 and a 0.5 probability responding below 3 (i.e., 0&1&2 vs. 3).

Mean square statistic (MNSQ). In Rasch analysis, item fit indexes are reported for individual items. The MNSQ statistic is sensitive to response patterns of persons whose ability estimates match an item's difficulty estimate. Overfit indicates that the observations contain less variance than is predicted by the model; underfit indicates more variance in the observations than is predicted by the model (e.g., the presence of idiosyncratic groups). An item that equals 1 indicates perfect fit. In general, a value between 0.75 and 1.33 is considered to provide reasonable fit (Wilson, 2005).

Reliability and Validity

The reliability of this causal mechanistic tracing assessment will be demonstrated using evidence from several sources. The first source of evidence is Chronbach's alpha, from CTT; the last two sources are based on IRT.

Reliability

Classical test theory (CTT). In CTT reliability cannot be estimated directly, since that would require one to know the true scores (denoted by T) (defined as the expected number of correct scores (X) over an infinite number of independent administrations of the test), which, according to CTT, is impossible. In CTT, the use of various parallel forms can obtain estimates of reliability. However, the reliability coefficient, $\rho_{XX'} = \frac{\sigma_X^2}{\sigma_T^2}$, was not calculated because this would have required test-retest administrations using parallel forms for individual participants. Instead, Chronbach's alpha was reported. Cronbach's alpha can be shown to provide a lower bound for reliability under rather mild assumptions. Thus, the reliability of test scores in a population is always higher than the value of Cronbach's alpha in that population. Chronbach's alpha was calculated as a measure of reliability.

Cronbach's alpha. Chronbach's alpha measures the internal consistency of all items in a scale. Cronbach's alpha was computed for the construct as follows:

$$\hat{\alpha} = \frac{k}{k-1} \left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2} \right).$$

Here, k indicates the number of items.

Item response theory (IRT). This section describes ways to investigate whether the assessment instrument operates with sufficient consistency across individuals. In creating a

construct and developing an instrument, it is assumed that each respondent can be placed somewhere on that construct and measured reliably. The separation reliability was calculated. This measure indicates the proportion of the model variance that is accounted for by the total variance. In addition, the standard error of measurement (SEM) was calculated.

Separation reliability. In Rasch Measurement the person separation index is used instead of reliability indices (as used in CTT). The separation index is a summary of the genuine separation as a ratio to separation including measurement error. The amount of measurement error is not uniform across the range of a test, but is larger for more extreme scores (low and high). Separation reliability indicates how well the item parameters are separated; it has a maximum of one and a minimum of zero. ConQuest reports separation reliability, which indicates the extent to which observed total variance, $\text{Var}(\hat{\theta})$, is accounted for by the model variance, $\text{Var}(\theta)$:

$$\hat{\sigma}_p = \frac{\text{Var}(\theta)}{\text{Var}(\hat{\theta})}$$

$$= \frac{\frac{1}{J-1} \sum_{j=1}^J (\hat{\theta}_j - \bar{\theta}) - \frac{1}{J} \sum_{j=1}^J \text{SEM}(\theta_j)^2}{\frac{1}{J-1} \sum_{j=1}^J (\hat{\theta}_j - \bar{\theta})}$$

Standard error of measurement (SEM). An important difference between CTT and IRT is the treatment of measurement error, indexed by the standard error of measurement. All tests, questionnaires, and inventories are imprecise tools; we can never know a person's true score, but can only have an estimate, the observed score. There is some amount of random error that may push the observed score higher or lower than the true score. CTT assumes that the amount of error is the same for each examinee, but IRT allows it to vary; for this reason, the SEM will be calculated from the IRT analysis.

Construct Validity

This section describes evidence, obtained from IRT analysis, that the instrument targets the construct map. The item-step Wright map was used to determine whether the item responses are consistent with the hypotheses from the construct map. The item-step Wright map was used to empirically determine whether participant responses confirmed hypotheses about the difficulty of the mechanistic elements from the construct map.

CHAPTER IV

RESULTS

The Results section reflects the following structure: First, descriptive item statistics are presented. These data show that the majority (89%) of participants responded to the assessment items by citing at least one mechanistic element, suggesting they have resources to engage in causal mechanistic tracing. Next, item analyses based on CTT and IRT are shown. From the IRT analysis, an item Wright map is presented. Results show that the relative difficulty of items was based on three characteristics of the machines with which participants worked: (1) number of levers, (2) lever type, and (3) the presence of intermediate links. Next, reliability and validity measures are presented from both CTT and IRT. Here, an item-step Wright map is presented, which shows the difficulty ordering of the mechanistic elements for each item. As a next step, the differences between participants who scored many items at the construct map's top two levels (i.e., *constraint via the fixed pivot* and *tracing*) are reviewed. Although participants who scored in each category can diagnose a machine's motion according to all of its mechanistic elements, only those scored at the highest level can causally connect each mechanistic element from input to output. Then, an investigation of differences across members of the top construct level (i.e., *tracing*) is presented in order to determine whether increasing complexity of the machines that they work with can disrupt participants' diagnosis and causal connection of a machine's mechanisms from input to output. The section concludes with a discussion concerning the stability of causal mechanistic reasoning.

Descriptive Item Statistics

The means of the item scores and variances were calculated to present the typical scores as well as their spread for each item. In order to further investigate these items, the distribution of scores is presented for the three link items (because these items had the most responses and would likely have smaller variances and more stable distributions). In addition, the score distributions of the items with the most extreme variances, both small and large, were evaluated. This was all done in order to determine whether the distribution of item scores were consistent with the construct's hypothesized difficulty orderings.

Item responses were scored as follows: *linked direction*=1; *rotation*=2; *lever arms*=3; *constraint via the fixed pivot*=4; and *tracing*=5. This ordering was based on the hypothesized difficulty rankings for each mechanistic element, from easiest to most difficult.

Mean of item *i*. The mean item scores, across all respondents, ranged from 0.38 to 3.09. Table 4-1, in Appendix B, shows the means for each item. Other descriptive statistics (e.g., median, mode, and standard deviation) that present a clearer sense of the distributions of scores are also shown in this table. The average of the item means was 1.65; this is similar to the average of the item medians (1.60) and modes (1.19), with a mean item standard deviation of 1.27 (ranging from 0.49 to 2.01). On average, items were scored just above the hypothesized lowest level (i.e., *linked direction*=1); however, because of the large item variances, the 95% confidence interval likely locate the item means anywhere along the score distribution, across multiple test administrations.

Variance of item *i*. The item variance is the variance of each score across persons. These variances are small to large; they range from 0.24 to 4.04. However, this is not

unexpected due to the small sample of diverse populations. The variance of each item is displayed in Table 4-1 in Appendix B.

Figure 4-1 (a, b, and c) shows the distribution of scores for the three link items. These items best show trends in the score distributions because they have the most responses. On item HFPO (Figure 4-1a), respondents had a mean score of 0.69 and a variance of 0.58. This item shows the type of distribution one would expect for a well functioning item; the frequency of responses decreases as the hypothesized difficulty increases. For example, more respondents were scored at the level of *linked direction* than *rotation*. On item MPA2 (Figure 4-1b), respondents had a mean score of 1.00 and a variance of 0.61. This item similarly shows the expected distribution, with more responses for *linked direction* than for *rotation*. On item STD1 (Figure 4-1c), respondents had a mean score of 1.98 and a variance of 4.04 (the highest variance of all items). The distribution of scores on this item shows a bimodal distribution; forty-six percent (n=51) of the item responses were not scored on the construct map (i.e., indicating the absence of any mechanistic elements in their item responses). However, of those who were scored on the construct map, the pattern of responses was not consistent with the hypothesized construct level difficulty orderings. This item presents a machine built from four levers (the greatest number of levers for any item on the assessment) and has an intermediate lever. There are also two outputs, which move in opposite directions. It is clear that coordinating the direction of input and output (*linked direction*) is difficult on this item (and, thus, scored so infrequently). In addition, this item's numerous "holder constraints" (i.e., constraints that forced the outputs into approximately linear paths) may have disguised the machines rotary motion, making *rotation* less salient. However, it is not clear why *constraint via the fixed pivot* and *tracing* were scored so frequently. On this item, participants either failed

to diagnose any mechanistic elements or were able to diagnose them all (thirty-three percent of the sample was scored at *constraint via the fixed pivot* or *tracing*).

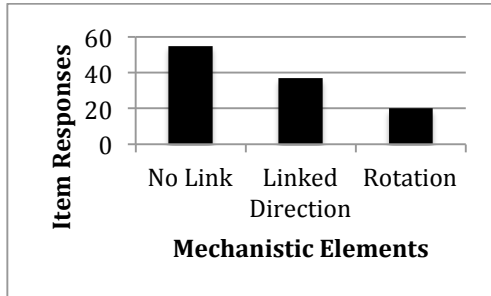


Figure 4-1a: Score distribution (HFPO).

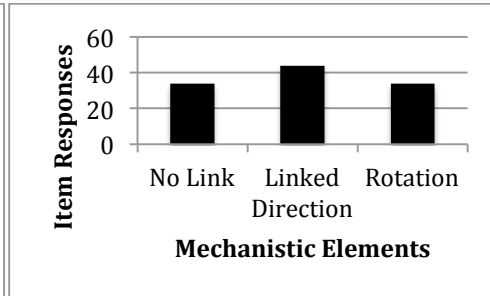


Figure 4-1b: Score distribution (MPA2).

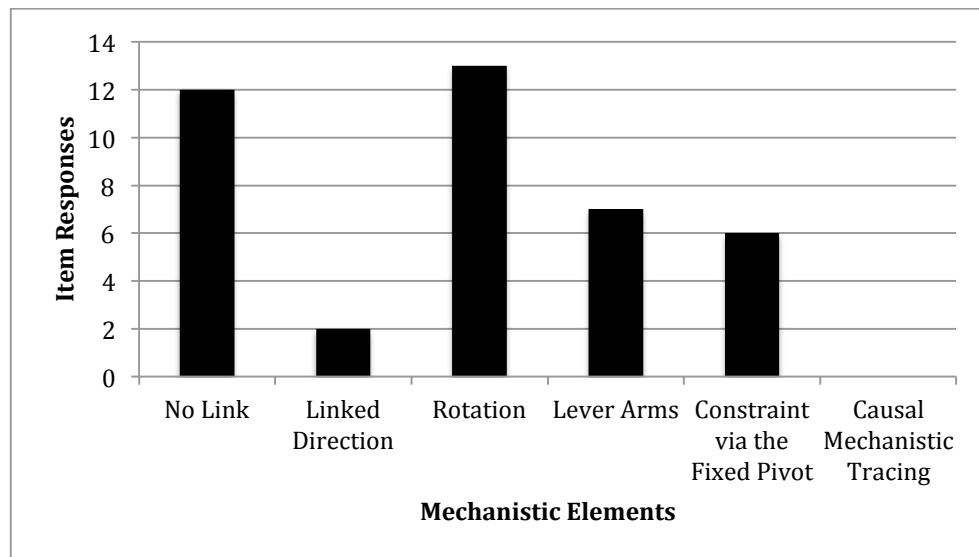


Figure 4-1c: Score distribution

On item MPD1 (Figure 4-1d), respondents had a mean score of 0.38 and a variance of 0.24 (the lowest variance of all items). This is a dichotomous item (i.e., participants were either scored at the level of *linked direction* or were not scored on the construct). On item STA3' (Figure 4-1e), respondents had a mean score of 1.83 and a variance of 2.05 (the lowest variance for an item with all six score categories). On this item, with the exception of *linked direction*,

the distribution of scores is consistent with the hypothesized construct levels. On this item, *linked direction* is scored less frequently than *constraint via the fixed pivot*. I conjecture that because this is a class 3 lever, where the input and output move in opposite directions; this made being assessed at the level of *linked direction* difficult unless participants had an understanding of fixed pivot constraint, in which case they would have been scored at the level of *constraint via the fixed pivot*.

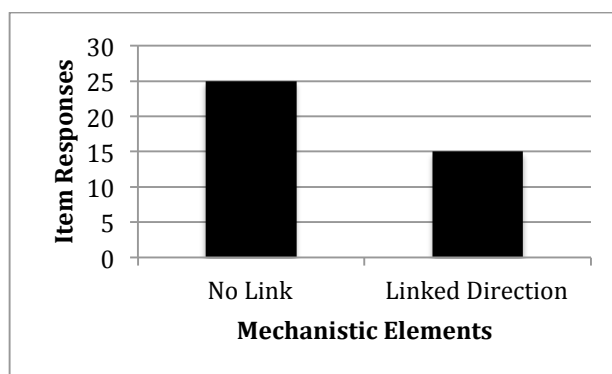


Figure 4-1d: Score distribution (MPD1).

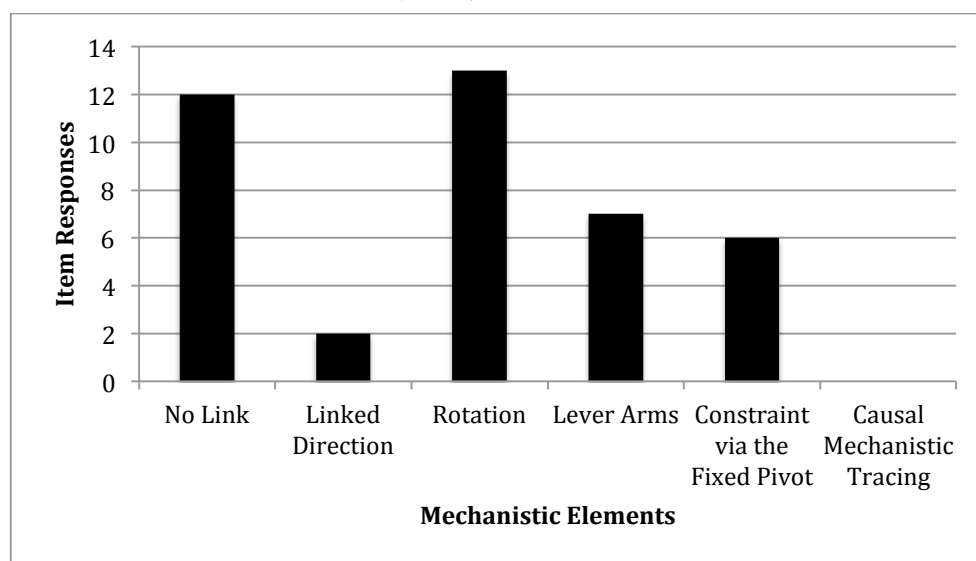


Figure 4-1e: Score distribution (STA3').

Eighty-nine percent (n=100) of respondents scored on the construct map (i.e., showed an element of mechanistic reasoning) on at least one item. This result is consistent with Bolger

and colleagues (2012) as well as Shultz’s (1982) findings that individuals show competencies in reasoning about mechanism from early ages.

Item Analysis with Classical Test Theory (CTT) and Item Response Theory (IRT)

Classical test theory (CTT)

Two CTT statistics are presented to analyze the item characteristics: (1) item difficulty and (2) item discrimination.

Item difficulty. The items range in difficulty from 0.99 (easy item) to 0.24 (difficult item). However, all the items have a mean (and median) item difficulty of 0.50, with a 90% confidence interval ranging from 0.24-0.76. Thus, the items have average (or typical) high to moderate difficulty. The item difficulty indices for all items can be seen in Table 4-2.

Assessments may be rejected as unreliable if item difficulty is not consistent with person ability. Tests that are too difficult or too easy for the respondents who take them often show low reliability (Henning, 1987). Thus, because the items are predominantly (90%) in the medium difficulty range, the assessment is likely to have good reliability.

Table 4-2		
<i>Item difficulty.</i>		
High	Medium	Low
(Difficult)	(Moderate)	(Easy)
<0.30	0.30-0.80	>0.80

Item discrimination. Item discrimination refers to the ability of an item to discriminate

between those participants who are relatively high on the criterion of interest and those who are relatively low. It provides an estimate of whether each individual item is measuring the same criteria as all other assessment items. These assessment items range from 0.58 (fair) to 0.91 (good). All item discrimination values can be seen in Table 4-1 in Appendix C. Table 4-3 classifies an item's ability to discriminate between participants at different ability levels by its item discrimination estimate. The mean of the item discrimination estimates is 0.77 (good). Nineteen of the twenty-one items (90%) have a good ability to discriminate between participants' reasoning about different mechanistic elements; the remaining two items have a fair ability to make this discrimination.

Table 4-3		
<i>Item discrimination estimates.</i>		
Poor	Fair	Good
<0	0-0.60	>0.60

Item response theory (IRT)

Two Wright Maps were generated by the item analysis: (1) the Item Map and (2) the Item-step Map. The Wright Map shows the distribution of items/item thresholds and respondents along a unidimensional logit scale. The item map presents the persons' scores on the same scale as the average item location across steps for each item; the item-step map presents the persons' scores on the same scale as each step (i.e., Thurston threshold) for each item. Together these maps provide insight into what is difficult about reasoning about groups of items as well as reasoning about the mechanistic elements, across all items. Standard errors

of item difficulty estimates are large because of the small sample size. Therefore, it can be less reliable to check if common items function similarly across test forms with IRT equating procedures. Thus, item difficulty parameters have been calibrated with the assumption that common items function equally across test forms.

This section first presents results that show that the item responses are consistent with IRT scoring assumptions. Next, the item Wright map is presented in order to analyze the behavior of the items. Then, the MNSQs are presented in order to determine the item fit. Finally, the item-step Wright Map is presented in order to characterize item difficulty with respect to the Thurston steps for each item.

IRT scoring assumption. According to IRT modeling, participants who are scored at one level should be capable of performances specified at all lower levels. In these data, across all items, 82% percent of all respondents (n=92) were scored at all levels of the construct map easier than the highest level at which they were scored. This finding indicates that participants who can reason about an item by diagnosing one mechanistic element (or a combination of mechanistic elements) can, in most cases, also diagnose easier elements.

Item Wright map. The item Wright Map, shown in Figure 4-2, makes it possible to compare the mean difficulty of each item across the sample. For example, Sequential Tracing E1 (STE1) is the most difficult item, with a mean item difficulty of 0.92 logits. The easiest item is STA3, with a mean item difficulty of -0.76 logits. Table 4-4 presents all item estimates and their corresponding standard errors. The standard errors indicate the precision of the estimates.

5						
		X				
4						
		X				
3		X				
		XX				
		XX				
		XX				
		XX				
2		XX				
		X				
		XXXX				
		XXXXXXX				
1		XXXXX	STE1			
		XXXXXXX	MPD1			
		XXXXX	HFPO	MPD1'	STD1'	
		XXXXXXXXX	STE2	MPA3'	MPB2	
0		XXXXXXXXXX	STD1	HFPS	STB1'	
		XXXXXXXXXX	STA3'	STCMT		
		XXXXXXXXXX	MPA2	MPA3	MPB2'	STB1
		XXXXXXX	MPA1	STA1	STA3	
		XXXXXX				
-1		XXXXXXX				
		XXXXX				
		XXXXX				
		XXXXX				
-2		XXXXX				
		XXX				
		XXXX				
		XXX				
-3		XX				
		XX				
		XX				
		XX				
-4		XX				
		XX				
-5		X				
		X				
-6		X				

Figure 4-2. Item Wright Map.

Table 4-4*Item difficulty estimates and standard errors.*

Item	Item Difficulty Estimate (logits)	Standard Error
Hands Fixed Pivot-Opposite	0.587	0.115
Machine Prediction-A2	-0.426	0.114
Sequential Tracing-D1	0.171	0.079
Sequential Tracing-E2	0.323	0.109
Hands Fixed Pivot-Same	0.008	0.128
Machine Prediction-A1	-0.547	0.133
Machine Prediction-A3	-0.319	0.133
Machine Prediction-A3'	0.259	0.133
Machine Prediction-B2	0.286	0.131
Machine Prediction-B2'	-0.391	0.135
Machine Prediction-D1	0.711	0.144
Machine Prediction-D1'	0.543	0.142
Sequential Tracing-A1	-0.700	0.117
Sequential Tracing-A3	-0.760	0.115
Sequential Tracing-A3'	-0.169	0.120
Sequential Tracing-B1	-0.519	0.117
Sequential Tracing-B1'	0.134	0.105
Sequential Tracing-B2	-0.487	0.114
Sequential Tracing- D1'	0.578	0.113
Sequential Tracing-E1	0.923	0.113
Sequential Tracing-CMT	-0.205*	
<i>Note: *Estimate is constrained</i>		

This item Wright map helps us consider the specific properties of these items that make causally tracing from input to output more or less difficult. In order to trace, participants need to reason about all of a machine's mechanistic elements and their causal coordination. A participant's ability to do this may be dependent on item type (e.g., are participants better at predicting the motions of the output lever than they are at predicting the motion of any of a machine's internal levers?). This section reports how the following machine characteristics impact participants' diagnosis and causal connection of a machine's mechanisms: (1) number of levers, (2) the arrangement of levers, (3) lever type (e.g., class one levers), and (4) the presence of specialized and unfamiliar levers (e.g., a bent crank). The "number of levers," "arrangement of levers," and inclusion of a "bent crank" are not independent machine characteristics. However, each is included in this analysis in order to determine the effect each singularly has on causal mechanistic tracing.

Item Type. There were two item types used in the final version of this assessment: (1) machine prediction items and (2) sequential tracing items. Machine prediction items ask respondents to predict the motion of machine outputs, whereas sequential tracing items ask respondents to predict the motion of all the different machine parts from input to output. There was no difference in item difficulty estimates between the two item types. Thus, it is not more difficult to predict motion of the output than to predict the motion of any other machine lever.

Number of levers. The number of the levers in a machine contributes to its visual complexity and impacts participants' ability to recognize mechanistic elements and causally trace from input to output. This is because added links require that respondents diagnose more mechanistic elements in multiple places. Participants had greater difficulty in diagnosing

machines composed of three or more levers ($M = 0.19$ logits) than those with two or fewer ($M = -0.38$ logits; $p=0.003$, one-tailed).

Lever type. The type of lever in the machines impacts difficulty as well. Five items include machines composed of class one levers; five items feature machines composed of class three levers. Participants had greater difficulty with class three levers ($M = -0.03$) than the corresponding class one levers ($M = -0.41$; $p=0.08$, one-tailed). In class one levers the input and output move in the same direction. It appears easier to imagine this simple translation than the opposite directed motion of the input and output that is characteristic of the class 3 lever.

Arrangement of levers. In addition to the number of levers in a machine, their arrangement is also important. Of the twenty-one items, seven were constructed with one or more intermediate link(s) between the input and output. These seven items were more difficult ($M = 0.43$ logits) than the remaining fourteen ($M = -0.22$; $p=0.001$, one-tailed), which had no intermediate links between the input and output.

Bent Crank. Participants also had difficulty diagnosing the mechanistic elements of, as well as causally tracing through, machines that used intermediate links that were not standard levers (e.g., bent cranks). The most difficult item was STE1; this is shown in Figure 4-2. This item contains an input link, an output link, and an intermediate link that is a bent crank. The cognitive interviews suggested that the motion of this intermediate piece was confusing to many participants. They found it difficult to predict the rotary path or the coordinated motion of the lever arms of the bent crank. In addition, it was not obvious how this intermediate link transmitted motion from the input to the output. Many participants who could predict the correct motion of the bent crank were unable to trace that motion correctly to the output.

Another item with a bent crank as the intermediate link, STE2, was also one of the most difficult items on the assessment.

Mean square statistic (MNSQ). In Rasch analysis, item fit indices can be reported for individual items. An item that has a mean squared statistic equal to 1 indicates perfect fit. In general, a value between 0.75 and 1.33 indicates good fit (Wilson, 2005). Table 4-5 shows the mean squared statistic for all of the items. Of the twenty-one items, seventeen (81%) are good fits. Two items, Hands Fixed Pivot- Opposite and Sequential Tracing-B1' are slightly out of the good fit range. An additional two items are farther out of this range: Machine Prediction-B2' (0.60) and Sequential Tracing-D1' (MNSQ=1.66). Wright and Linacre (1994) suggest that only Sequential Tracing-D1' (MNSQ= 1.66) would produce a misfit that would be unproductive for assessment, but would not degrade the assessment. This is shown in Table 4-6. Moreover, the inclusion of only one item (5% of the total items) with a problematic fit statistic does not compromise the assessment. Practice dictates that if more than 10% of the values are outside the 0.75-1.33 range, the most misfitting items should be reworded, omitted, or placed in a sub scale (Wright & Masters, 1982).

Table 4-5	
<i>MNSQ fit statistic for each item.</i>	
Item	Mean Squared Statistic (MNSQ)
Hands Fixed Pivot- Opposite	1.34*
Machine Prediction-A2	1.22
Hands Fixed Pivot- Same	1.13
Machine Prediction-A1	1.23
Machine Prediction-A3	1.16
Machine Prediction-A3'	0.90
Machine Prediction-B2	0.97
Machine Prediction-B2'	0.60**
Machine Prediction-D1	0.98
Machine Prediction-D1'	0.94
Sequential Tracing- A1	1.10
Sequential Tracing- A3	1.02
Sequential Tracing- A3'	0.78
Sequential Tracing-	0.78

B1	
Sequential Tracing-	1.37*
B1'	
Sequential Tracing-	1.10
B2	
Sequential Tracing-	1.07
D1	
Sequential Tracing-	1.66**
D1'	
Sequential Tracing-	1.21
E1	
Sequential Tracing-	1.03
E2	
Sequential Tracing-	1.15
CMT	
Note: **Not considered a good fit, *Marginally out of good fit range	

Table 4-6

Interpretation of parameter-level mean-square fit statistics (Wright & Linacre, 1994).

Mean-square Value	Implication for Measurement
> 2.0	Distorts or degrades the measurement system. May be caused by only one or two observations.
1.5 - 2.0	Unproductive for construction of measurement, but not degrading.
0.5 - 1.5	Productive for measurement.
< 0.5	Less productive for measurement, but not degrading. May produce misleadingly high reliability and separation coefficients.

Wilson (2005) states that one way to judge the severity of the misfit is to compare the expected proportion (based on the estimated item parameters) scored at each construct level to the actual proportion. Figure 4-3 shows (at top) an item with a fit at the extreme of the .75-1.33 range (MNSQ=1.3) and (at bottom) an item with perfect fit (1.0). The continuous line shows the expected cumulative probabilities and the dots show observed cumulative proportions who responded at each level. Where the dots lie along the line the fit is good; as they depart from the line the fit gets somewhat worse. It can be valuable to explore the degree of misfit in order to determine the severity of the misfit. Figure 4-4 presents the fit plot for Sequential Tracing-D1', where the proportion of responses, per ability level, are greatly inconsistent with that expected by the model. This is significantly worse than a misfit, as in Figure 4-3 (top), where the proportion of responses slightly deviate from the model in many places.

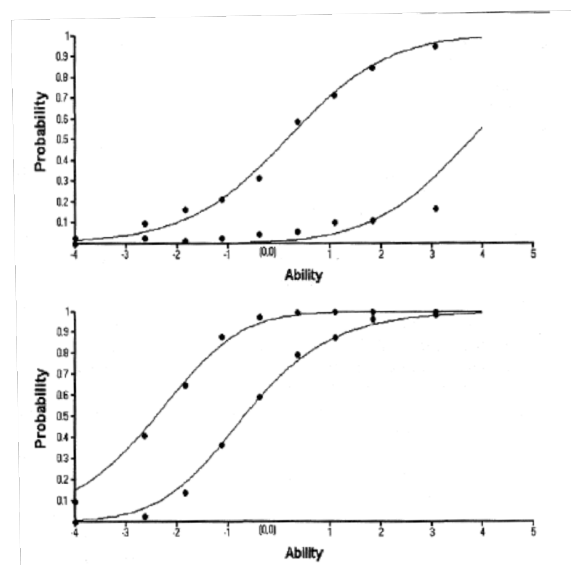


Figure 4-3. Fit plots for two items. This figure illustrates an item fit (upper panel) that is just within the good fit range and another item (bottom panel) that is a perfect fit (Wilson, 2005; p. 131).

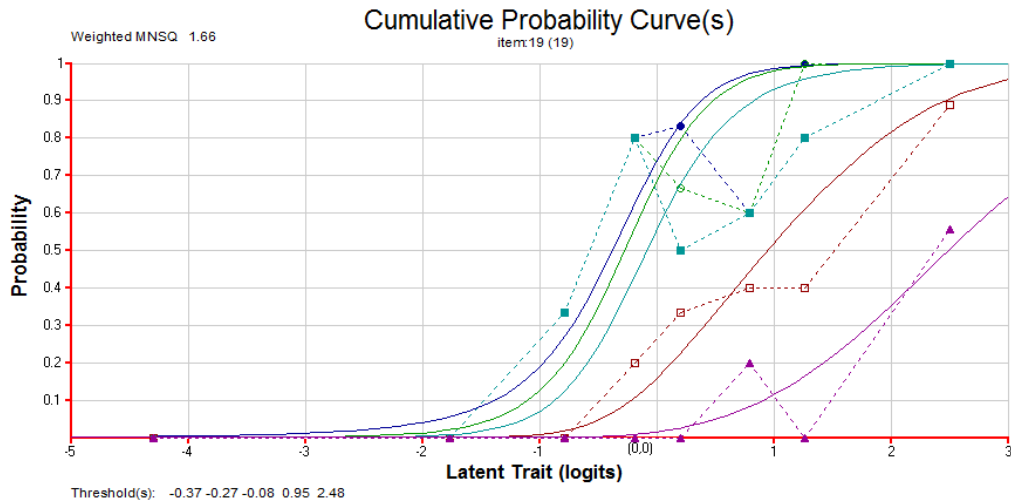


Figure 4-4. Fit plot for item Sequential Tracing-D1'. This figure illustrates a misfit.

Sequential Tracing-D1'. In Figure 4-5 the actual variance is greater than expected.

According to the Wright Map that is presented in Figure 4-2, this item was one of the most difficult. The probability of scoring at the level of *tracing* (i.e., this item's most difficult level) is consistent with the model (probability = 0) from the person ability level of -3 logits to 0 logits (i.e., 0, 0). At approximately (0.8, 0.2), the respondents have a significantly higher probability (i.e., 0.2) of scoring at this level than the model predicts. However, at approximately ability level 1.2 logits, the respondents again have a 0 probability that they will score at the level of *tracing*. This shows that on this item, as ability increases, the probability of being scored at *tracing* decreases. This indicates item misfit.

The three easiest mechanistic elements for this item, *lever arms*, *rotation*, and *constraint via the fixed pivot*, substantially deviate from the model from person ability estimate 0.4 logits to 2.5 logits. *Lever arms* dips from (0.4, 0.9) to (0.8, 0.6), below the model value for

rotation. This shows that participants with a person ability estimate of 0.4 logits have a probability of 0.9 of scoring at the level of *lever arms* on this item (this item's easiest level), while those with a person ability estimate of 0.8 logits have a probability of only 0.6 of scoring at the level of *lever arms* on this item. In addition, *rotation* dips from (-0.4, 0.8) to (0.4, 0.5). This shows that participants with person ability estimates of -0.4 logits have a probability of 0.8 of scoring at the level of *rotation*, but participants with person ability estimates of 0.4 have a probability of only 0.5 of scoring at this level. This indicates misfit.

The item-step Wright map. Figure 5-6 presents an item-step Wright map that places respondent ability and each item Thurston threshold (i.e., difficulty for each mechanistic element by item) on the same latent continuum. For instance, the element *tracing* has an item difficulty estimate of 3.04 logits for the item Sequential Tracing-E1 (an item with a bent crank as an intermediate link). This indicates that those respondents who have person ability estimates of 3.04 logits will have a 0.5 probability of being scored at this level for this item. Table 4-6 presents the item-step estimates for each item with its corresponding standard error. The standard error indicates the precision of the estimates.

5									
	X								
4									
	X								
3	X	STE1.T							
	XX								
	XX	STB1'.T							
	XX	STD1'.T							
	XX								
2	XX	STD1.T	STB1.T	STT.T					
	X	HFPO.R							
	XXXX	STE2.CFP	MPA3'.R	STA1.CFP	STA3'.CFP	STB2.T			
	XXXXXXXX	STE1.LA							
1	XXXXX	STA3.CFP	STD1'.LD	STE1.LD					
	XXXXXXXX	MPA2.R	MPB2.R	MPB2'.R	MPD1.LD				
	XXXXX	MPD1'.LD	STA1.T	STB2.CFP					
	XXXXXXXXXX	STD1.LD	STE2.LD	HFPS.LD	MPA3.R	STA3.T	STA3'.LA	STB1.CFP	STT.CFP
0	XXXXXXXXXX	MPA1.R	STE1.CFP						
	XXXXXXXXXX	STE2.LA	MPB2.LD	STB1'.LD	STD1'.R	STD1'.CFP			
	XXXXXXXXXXXX	HFPO.LD	STD1.R	STD1.CFP	STE2.R	HFPS.R	STD1'.LA		
	XXXXXXX	STD1.LA	STB1'.R	STB1'.CFP	STE1.R				
	XXXXXX	MPA3.LD	MPA3'.LD	STB1'.LA					
-1	XXXXXXX	MPA1.LD	STA3.LA	STA3'.LD	STB1.LA	STB2.LD	STT.R	STT.LD	
	XXXXX	STA3'.R	STB2.R	STT.LA					
	XXXXX	MPA2.LD	MPB2'.LD	STA1.R	STB1.R				
	XXXXX	STA1.LD	STA3.R	STB2.LA					
-2	XXXX	STA1.LA							
	XXX	STB1.LD							
	XXXX	STA3.LD							
	XXX								
-3	XX								
	XX								
	XX								
	XX								
-4	XX								
	XX								
-5	X								
	X								
-6	X								

Tracing:
Mean = 1.78 logits

Constraint via the Fixed Pivot:
Mean = 0.24 logits

Linked Direction:
Mean = -0.56 logits

Rotation:
Mean = -0.31 logits

Lever Arms:
Mean = -0.82 logits

Figure 4-5. Item-step Wright

Reliability and Validity

Reliability

This section describes ways to investigate whether the assessment instrument operates with sufficient consistency across individuals. In creating a construct and developing an instrument it is assumed that each respondent can be placed somewhere on that construct and measured reliably.

Classical test theory (CTT)

Cronbach's alpha. Chronbach's alpha is a measure of internal consistency; it determines how closely related a set of items is as a group. A "high" value of alpha is often used (along with substantive arguments and possibly other statistical measures) as evidence that the items measure an underlying (or latent) construct. As a rule of thumb, many professionals require a value of at least 0.70 before they will use an instrument. Here Cronbach's alpha (α) is equal to 0.54; however, though these items did not reach the 0.70 threshold, this criteria should be loosened with items that are not altogether correlated. For example, ten of the items were not capable of assessing the three highest construct levels (*lever arms, constraint via the fixed pivot, and tracing*), whereas eleven were. Clearly, responses across these items would not correlate highly. In addition, it should be considered that the sample size was necessarily small in order to accommodate the cognitive interviews conducted. These two factors account for this low reliability measure. Thus, Chronbach's alpha is not the ideal measure of reliability for this assessment. In IRT, the standard error of measure provides a measure of reliability that is better suited.

Item response theory (IRT)

Separation reliability. In Rasch Measurement separation reliability indicates how well the item parameters are separated; it has a maximum of one and a minimum of zero. This value is typically high and increases with increasing sample sizes. These items have a separation reliability equal to 0.94; suggesting that most observed total variance, $\text{Var}(\hat{\theta})$, is accounted for by the model variance, $\text{Var}(\theta)$. There is no professional standard; however, the state of California has accepted the separation reliability value of 0.90 as a minimum for achievement tests used in schools for individual testing. However, this level has not been consistently applied (Wilson, 2005).

Standard error of measurement (SEM). In IRT, unlike CTT, person abilities are reported with standard errors of measurement (SEM) that indicate how reliable a person's ability estimate is. Figure 4-4 shows that for this assessment a participant whose ability estimate is in the middle of the logit scale tends to have smaller SEM values, whereas those on the two extremes tend to have larger SEM values. The smaller the SEM, the more reliable the ability estimates. The mean SEM for these items is equal to 0.49, with a range from 0.27 to 1.10. The person estimates range from -3.69 logits to 3.29 logits. The highest SEM values may be too high to precisely indicate the person ability estimates; however, the person ability estimates are reasonably good approximations for all but the most extreme ability estimates. The relationship between person ability estimate and standard error of measurement (SEM) in Figure 4-6 indicates reliability.

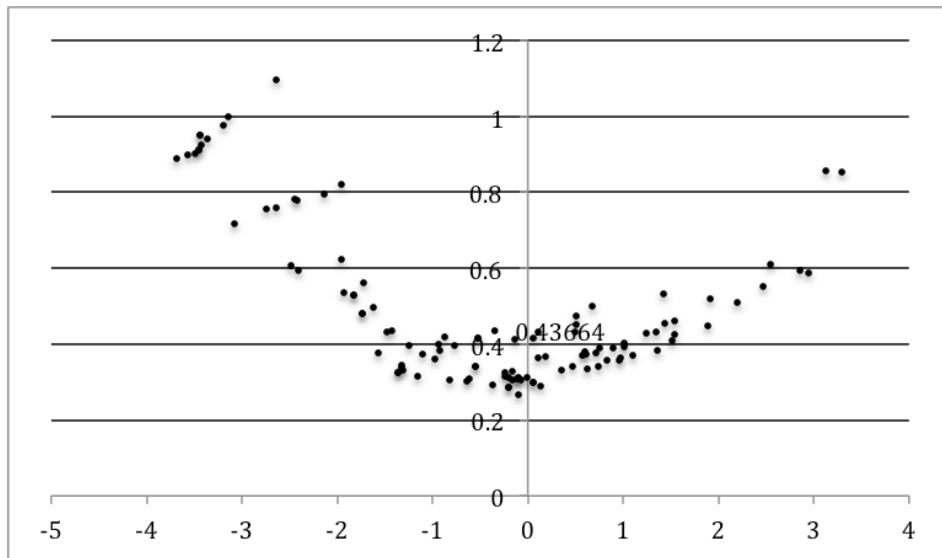


Figure 4-6. Scatter plot: Person ability estimates v. standard error of measurement (SEM).

Validity. This section describes evidence that the instrument targets the construct map.

First, results from the item-step Wright map are compared with the hypotheses from the construct map. The item-step Wright map is used to empirically determine whether participant responses confirm hypotheses about the difficulty of the mechanistic elements from the construct map (i.e., Table 2-1). The item-step Wright map makes it possible to consider the specific properties of these machines that make causally tracing from input to output more and less difficult. In order to trace, participants need to reason about all of a machine’s mechanistic elements and their coordination. This section reports how the following machine characteristics impacted participants’ diagnosis and causal connection of a machine’s mechanistic elements: (1) lever type (e.g., class one levers) and (2) the presence of specialized and unfamiliar levers (e.g., a bent crank).

Correspondence between item responses and participant talk and gesture. Items were scored according to the exemplars, whereas participant talk and gesture (while responding to the items) were coded independently according to the analytic framework developed in earlier

research (Bolger et al., 2012). This coding was completed for 715 items (across participants) that were scored on the construct; the talk and gesture for item responses that were not scored on the construct were not considered. Because item responses are nested in items and persons, a one-way chi-square goodness-of-fit test compared the mechanistic elements scored for each item with the corresponding interview coding. For example, 219 items were coded as *linked direction* using the exemplar. Of those 219 items, during the interview 44 (20%) were coded as “no mechanistic elements,” 136 (62%) were coded as “*linked direction*,” 38 (17%) were coded as “*rotation*,” 1 (0%) was coded as “*lever arms*,” and none (0%) were coded as either “*constraint via the fixed pivot*” or “*tracing*.” This distribution of coded mechanistic elements for *linked direction* is different ($p < 0.0001$) from the expected proportions, based on how all 715 items were scored according to Bolger and colleagues. Table 4-7 shows how all items were both scored and coded. For instance, Table 4-7 shows that seventy-four percent of those items scored at the level of *rotation* on the exemplar were also coded at that level in the cognitive interview. This relationship is consistent across all the mechanistic elements: (1) *linked direction* (62%); (2) *rotation* (74%); (3) *lever arms* (61%); (4) *constraint via the fixed pivot* (45%); and (5) *tracing* (46%). This suggests that when participants are responding to the paper and pencil items they are reasoning about physical levered machines. This is an indication of construct validity and shows that the assessment is measuring participants’ capacity to reason about the targeted mechanistic elements. Some items, according to the exemplars, could not be scored at all levels even though their explanations, according to Bolger and colleagues (2012) could. In these cases, participants were coded at the highest level scoreable on the exemplar.

Table 4-7

Percentage of mechanistic elements (scored on items) compared with how they were coded in the cognitive interviews.

		Exemplars					
Analytic Framework (Bolger et al., 2012)		Linked Direction*	Rotation*	Lever Arms*	Constraint via the	Tracing*	Total
		(n=219)	(n=199)	(n=114)	Fixed Pivot*	(n=74)	(N=715)
					(n=109)		
	No Mechanistic Elements	20%	13%	14%	4%	3%	13%
	Linked Direction	62%	8%	0%	1%	0%	21%
	Rotation	17%	74%	4%	4%	3%	27%
	Lever Arms	0%	5%	61%	33%	22%	19%
Constraint via the Fixed Pivot	0%	0%	11%	45%	27%	11%	
Tracing	0%	1%	9%	14%	46%	8%	

Note: Chi-squared goodness-of-fit (*p<0.0001, non-directional)

Ordering the mechanistic elements according to difficulty. The item thresholds for each of the five mechanistic elements are presented in Table 4-8 and graphically represented in Figure 4-5 (the item-step Wright Map). The means of these thresholds for each mechanistic element are presented and are rank ordered according to difficulty as follows (from the easiest to most difficult mechanistic elements across the twenty-one items): (1) *lever arms*, (2) *linked direction*, (3) *rotation*, (4) *constraint via the fixed pivot*, and (5) *tracing*. There were mean differences in difficulty between *rotation* ($M = -0.36$) and *constraint via the fixed pivot* ($M = 0.52$; $p < 0.1$, one-tailed), as well as *constraint via the fixed pivot* and *causal mechanistic tracing* ($M = 1.80$; $p < 0.0001$, one-tailed). Thus, *tracing* is the most difficult mechanistic element, more difficult than both *constraint via the fixed pivot* and *rotation*. There is no difference between the three easiest levels. I conjecture that this is a result of the diverse and small sample.

The following section explains how participants diagnosed each of the mechanistic elements. Then the section examines differences between participants who can diagnose all of the mechanistic elements (i.e., assessed at the level of *constraint via the fixed pivot*) and those who can, further, causally connect them (i.e., assessed at the level of *tracing*). The section concludes with a discussion of those machine characteristics that seem to disrupt a participant's ability to diagnose and causally connect all of a machine's mechanistic elements. That is, why does a participant who can diagnose and causally connect all mechanistic elements on one item fail to do so consistently?

Table 4-8

Item thresholds.

Item	Lever Arms	Linked Direction	Rotation	Constraint via the Fixed Pivot	Tracing
Hands Fixed Pivot- Opposite		-0.41	1.59		
Machine Prediction-A2		-1.55	0.70		
Sequential Tracing-D1	-0.59	0.25	-0.45	-0.36	1.83
Hands Fixed Pivot-Same		0.38	-0.37		
Machine Prediction-A1		-1.17	-0.08		
Machine Prediction-A3		-0.84	0.20		
Machine Prediction-A3'		-0.98	1.49		
Machine Prediction-B2		-0.20	0.77		
Machine Prediction-B2'		-1.59	0.80		
Machine Prediction-D1		0.71			
Machine Prediction-D1'		0.55			
Sequential Tracing-A1	-2.12	-1.79	-1.68	1.51	0.61
Sequential Tracing-A3	-1.07	-1.73	-2.41	1.04	0.38
Sequential Tracing-A3'	-1.44	-1.22	0.38	1.54	
Sequential Tracing-B1	-1.00	-2.18	-1.52	0.20	1.89
Sequential Tracing-B1'	-0.83	-0.07	-0.69	-0.55	2.65
Sequential Tracing-B2	-1.87	-1.11	-1.46	0.50	1.48
Sequential Tracing- D1'	-0.37	0.95	-0.27	-0.08	2.48
Sequential Tracing-E1	1.19	0.96	-0.60	0.01	3.04
Sequential Tracing-E2	0.35	-0.46	-0.29	1.57	
Sequential Tracing-CMT	-1.24	-1.00	-1.07	0.33	1.86
Mean	-0.82	-0.60	-0.36	0.52*	1.80**

Note: **p<0.01, *p<0.1; Machine Prediction and Hands items can only assess *linked direction* and *rotation*. Levels that could not be scored are highlighted in black; levels with no scores are highlighted in gray.

Lever arms. I hypothesized that lever arms would be more difficult than both *linked direction* and *rotation*. *Lever arms* was the most frequently scored mechanistic element according to the exemplars. To explain why, I refer back to the interview data. These data show that two dissimilar groups of participants were scored at the level of *lever arms* on the items: the first group was participants able to recognize few or no mechanistic elements (i.e., in the cognitive interview) and the second included those able to recognize most, if not all, mechanistic elements (i.e., in the interview). For example, of those scored at the level of *lever arms* on the exemplar, 14% were coded as using no mechanistic elements in their explanations in the interview, while 20% were coded as using either *constraint via the fixed pivot* or *tracing*. The first group may have been assessed at the level of *lever arms* on the items, but not in the interviews because participants with little understanding of the machines' mechanisms could conceive of a lever being "like a seesaw." Participants alluded to "weight," "gravity," or "pressure" on one side being greater than on the other to justify their predictions. However, their explanations did not describe the causally coordinated motion of the two lever arms. These participants primarily described one lever arm being up and the other being down, without indicating: (1) a causal coordination (i.e., that the motion of one arm would cause the motion of the other) or (2) a dynamic system (i.e., the presence of motion). For example, one participant noted: "this side is kind of tilted up, so it will go up *and* this side is kind of tilting down, so it will go down." This example was characteristic of those not invoking a causal coordination.

Alternatively, other participants who were scored at the level of *lever arms*, but coded at the level of *constraint via the fixed pivot* or *tracing* (i.e., in the cognitive interview), were typically those who had ideas about constraint, but failed to specify (on the items) that the

entire lever would be revolving around the fixed pivot. These participants were predominantly college undergraduates (50%, n=9; engineering majors=4). Figure 5-3a shows how a participant, Sarah, responded to an item that was scored at the level of *lever arms*, but coded at the level of *constraint via the fixed pivot* in the cognitive interview. Sarah was scored at *lever arms* on the item because she indicated that stars C and E would move in coordinated opposite directions. Note, however, that she did not indicate that A and B would move in coordinated opposite directions; in addition, she did not indicate the direction of the motion of star D. According to the exemplar, this suggests that she is not capable of diagnosing *constraint via the fixed pivot*. However, the interview data does not support this conclusion.

As Sarah began diagnosing the machine (see Figure 5-3a), she first noted that when the input lever was pushed up, “this thing (indicates right end of right lever arm, A’) turns that way (indicates rotary motion in the counterclockwise direction). This side (indicates left end of left lever arm, B’) goes that way (indicates rotary motion in the counterclockwise direction). So this star (A) goes that way (indicates the counterclockwise direction) and this star (B) goes that way (indicates counterclockwise direction, but does not indicate rotation around the fixed pivot).” Next, the participant correctly predicts the direction of the rotary motion of stars E and C, without attending to the motion of D. She then provides a mechanistic explanation for the predicted motion of A’ and B’: “If I push this up (if A’ is pushed up with the input), it pulls this (B’) down because it’s fixed right there (she points at the fixed pivot between A and B; Figure 4-7 shows her indication of the fixed pivot as the mechanism that generates the coordinated opposite motion from each lever arm).”

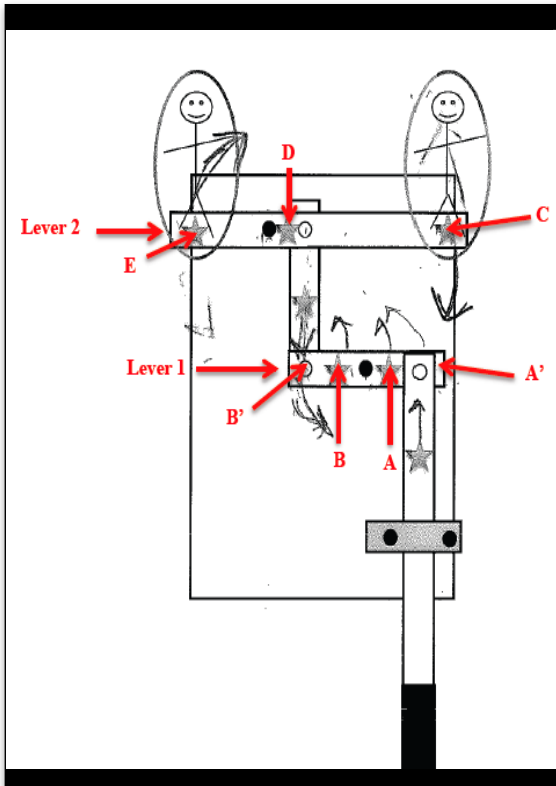


Figure 4-7a. Item response. This presents the item response from a student scored at the level of lever arms.

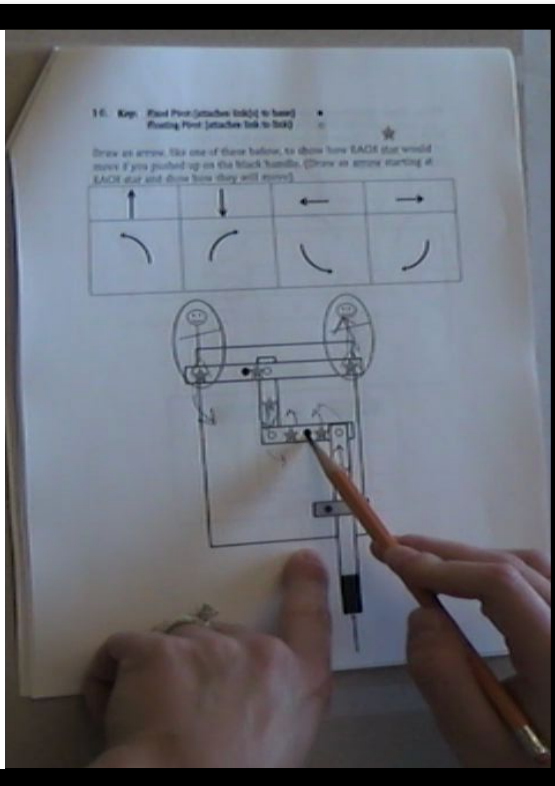


Figure 4-7b. Interview gesture. Student indicates that the fixed pivot is responsible for lever motion

How did this participant diagnose *constraint via the fixed pivot*, but not correctly characterize the direction of motion on opposite sides of the fixed pivot? The way that Sarah diagnosed this machine provides insight into what makes causal mechanistic tracing difficult. She diagnosed this machine according to the global motion of its parts (e.g., entire levers). She began by predicting the motion of A'; she then moved to diagnosing B'. She correctly diagnosed the motion of the entire lever. Considering these correct predictions and her subsequent recognition of fixed pivot constraint, she clearly understood the relationship between *constraint via the fixed pivot* and the general motion of the lever. However, after diagnosing A', she retraced her steps and attempted to predict the motion of A and B. Here, Sarah's tracing seems to have been disrupted. She was unable to replicate the same causal coordination of the lever arms around the fixed pivot. Sarah's correct prediction of the coordinated opposite rotary motion of the lever arms of lever two (i.e., points C and E) supports the hypothesis that Sarah can diagnose the coordinated motion of the lever around the fixed pivot globally. However, she seems to have difficulty coordinating the local lever arm motion around the fixed pivot. A total of 18 participants had similar difficulties diagnosing the directions of the stars surrounding the fixed pivot across twenty-three items.

Even though *lever arms* was the easiest mechanistic element to diagnose, its diagnosis is essential to causal mechanistic tracing. A participant may diagnose this mechanistic element by seeing the machine globally and recognizing that the motion of one lever arm implies opposite directed motion on the opposite side of the fulcrum. However, in order to causally trace, participants must recognize the importance of the fulcrum, because its location determines the path and direction of the lever (for a given input).

The other mechanistic elements. Linked direction, rotation, constraint via the fixed pivot, and tracing were rank ordered as was hypothesized. However, there were only mean differences between rotation, constraint via the fixed pivot, and tracing. Next, what accounts for the difficulty of the remaining mechanistic elements and how they contribute to causal mechanistic tracing are addressed.

Linked direction. After *lever arms*, *linked direction* was the easiest element to diagnose because it simply requires participants to notice the direction and causal coordination of the input and output links without referring to a specific path. Although this element is necessary for causal mechanistic tracing, it is not sufficient, because one can simply notice these relations without understanding the direction of the lever motion produced (i.e., rotary).

Rotation. After *linked direction*, *rotation* was the most difficult mechanistic element (according to a rank ordering of means across items). In order to identify *rotation*, participants must notice that the paths of the levers are rotary. Even having mastered the combination of *rotation* and *linked direction*, one might not necessarily be capable of causally tracing. Taken together, being able to diagnose these two mechanistic elements enables one to recognize the direction of specific links, but does not ensure that participants are thinking about what constrains the system to determine lever motion (i.e., the fixed pivot).

Constraint via the fixed pivot. After *rotation*, *constraint via the fixed pivot* was the most difficult mechanistic element. To recognize the mechanistic element of *constraint via the fixed pivot*, an individual must understand that the way in which the pivot is fixed to the board determines lever motion. This understanding is exemplified in the following explanation, coded at the level of *constraint via the fixed pivot* during a cognitive interview: “The fixed pivot is keeping it [the output link] from going up straight and it’s going around in a circle.”

Scoring *constraint via the fixed pivot* indicated that a participant recognized all of the easier mechanistic elements and thus should be prepared to causally trace through every machine, correctly diagnosing its mechanistic elements. However, being able to diagnose all of a machine's mechanisms does not seem to be sufficient to predict an individual's propensity to apply *tracing*. Across all items, there is a mean difference between *constraint via the fixed pivot* and *tracing*, showing that being able to diagnose and causally connect each mechanistic element is harder than simply being able to diagnose them.

Tracing. The most difficult level on the construct map, *tracing*, requires participants to be able to diagnose all of the easier elements, from input to output. Doing so on even one occasion is difficult; it is even more difficult to do so consistently across items, suggesting that mechanistic reasoning may vary with context (i.e., that is, specific features they are asked to respond to or features of the machines themselves).

Causal Mechanistic Tracing and Machine Characteristics

Twenty-five participants showed the ability to causally connect all four mechanistic elements on at least one item. However, two of the machine characteristics (lever type and the inclusion of a bent crank), discussed earlier, made a significant difference in these participants' ability to consistently causally coordinate all four mechanistic elements when responding to an item. There were a total of 11 items in which *tracing* could be assessed. The number of items per form where this level could be assessed ranged from 3 to 8, with a mean of 6 (median = 6).

Lever type. The mean percentage of items scored at the level of *tracing* (across all respondents who had scored at least one item at this level) was 0% on items with machines

with class 3 levers, in contrast to 80% for items with class 1 levers. Table 4-9 shows the mean percentage scores for this group across these two machine characteristics. The number of participants who increased the percent of items scored at the level of *tracing* across these machine features ($p=0.0005$, sign test) shows that the lever type has an impact on whether participants who have shown the capacity at least on one item to recognize and causally connect all four mechanistic elements, will do so on other items.

Bent cranks. The mean percentage of items scored at the level of *tracing* was 26% on items with machines with bent cranks, in contrast to 71% for items without bent cranks This is shown in Table 4-9. The number of participants who increased the percent of items scored at the level of *tracing* across these machine features ($p=0.01$, sign test) shows that the presence of a bent crank has an impact on whether participants who can recognize and causally connect all four mechanistic elements will do so.

Table 4-9

Tracing by machine characteristics.

Machine Characteristics		Scored at the level of <i>tracing</i>
Lever Type	Class 3 lever(s)	0%
	Class 1 lever(s)	80%**
Bent Crank	With Bent Crank	26%
	Without Bent Crank	71%*
Note: Sign test: **p<0.001; *p<0.01		

The presence of different lever types and bent cranks made a significant difference in the ability of those who could diagnose and trace all four mechanistic elements from input to output on one item, to do so on others. This suggests that mechanistic reasoning can be unstable across machines, even when people are reasoning about simple, inspectable mechanisms.

CHAPTER V

DISCUSSION

Children's Causal Mechanistic Reasoning

Mechanistic reasoning is fundamental for predicting and explaining the behavior of both designed and physical systems and, thus, is necessary for disciplinary practices (e.g., argumentation) in STEM fields (Bolger et al., 2012; Russ et al., 2009). Although intuitions about causes and effects emerges very early in life (e.g., Baillargeon, 1994; 1987a; 1987b; Baillargeon, Kim & Spelke, 1992; Baillargeon, Spelke, & Wasserman, 1985; Spelke, Katz, Purcell, Ehrlich, & Breinlinger, 1992), these resources do not necessarily translate into well-formed system reasoning later in life. In this study, 89% (n=100) of participants were able to diagnose at least one mechanistic element (i.e., one machine mechanism on at least one item), indicating that these early resources are present and functioning in this context. However, the mastery of this disciplined form of reasoning is not an all or nothing accomplishment, nor is its utility always transparent (Bolger et al., 2012; Bolger et al., 2011; Metz, 1991; Lehrer & Schauble, 1998).

Infants' rich naïve intuitions about cause within physical systems are apparently not systematically developed in schooling to assist adults (high school aged and older) in making consistent causal attributions about mechanisms within systems like those featured in this study (Carmazza, McCloskey, & Green, 1981; Clement, 1982; Minstrell, 1983). In this study, 78% (n=87) of all the participants failed on all the items to causally trace from input to output.

Making sense of the development of causal mechanistic reasoning requires more than an understanding of early resources and later “misconceptions.” The assessment reported here has characterized this form of reasoning about simple levered systems. In addition, it has helped to explain why this form of reasoning is difficult and what accounts for this difficulty. This study shows that machine characteristics such as number of levers, lever type, arrangement of levers, and inclusion of a bent crank can affect the difficulty of causal reasoning. In addition, even when participants do, on at least one occasion, trace pushes and pulls through a machine, inclusion of class 3 levers or bent cranks can disrupt their propensity to do so. During subsequent iterations of the design of this instrument, items that introduce additional machine features may lead to deeper understanding about features that tend to disrupt this kind of system reasoning.

Assessment Development (Research Question #1):
Can mechanistic reasoning be assessed via a standard assessment instrument?

This study reports the first iteration of the design of an assessment instrument for characterizing reasoning about basic mechanical systems. The instrument assesses individuals’ use of four mechanistic elements of levered machines: (1) *lever arms*, (2) *linked direction*, (3) *rotation*, and (4) *constraint via the fixed pivot*. The assessment instrument also assesses individuals’ ability to diagnose and causally trace all the mechanistic elements from input to output (i.e., *tracing*). The study has shown that when participants are responding to the paper and pencil items they are reasoning about the motion of actual physical levered systems. This is shown in Table 4-7.

This assessment showed good reliability and validity, using measures from both classical test theory (CTT) and item response theory (IRT). The large item variance is likely a result of the diverse and necessarily small sample to accommodate the interviews. I conjecture that this contributed to the low value of Chronbach's alpha and the absence of clear average difficulty difference between three easiest mechanistic elements. Small sample sizes produce large standard errors of item estimates.

However, it is also possible that the variance in Thurston threshold estimates across the three easiest mechanistic elements may be a consequence of the greater difficulty of being assessed at some construct levels (e.g., rotation) on items with certain machine characteristics (e.g., class 3 levers, intermediate links).

Next Design Iteration

This small and diverse sample was identified to assure, via cognitive interviews, that the items were assessing what they were intended to assess. Now that this has been established, it would be useful to administer the assessment to a larger sample to provide more evidence about item difficulty estimates, and to learn whether there are differences in the difficulty ordering of the three easiest mechanistic elements. A larger sample should provide more insight about the two different groups of participants that were scored at the level of *lever arms* (i.e., those citing no mechanistic elements and those citing *constraint via the fixed pivot* and *tracing* during the cognitive interviews). For instance: (1) will both of these groups continue to be scored at the level of *lever arms* with a larger sample; (2) can items be developed that will

differentiate between these groups; and (3) if so, how will this affect the difficulty of *lever arms* across the assessment items?

In addition, a larger sample will allow the further investigation of types of machine characteristics (e.g., number of levers, lever type, lever arrangement, and type of intermediate link) that affect individuals' ability to diagnose and causally trace a machine's mechanistic elements from input to output. This data can be modeled to determine the extent to which these machine characteristics contribute to item difficulty (e.g., a linear logistic latent trait model, Fischer).

This assessment administration provided substantial information with which to revise these items. Participant responses consistently populated all the construct levels and the mechanisms diagnosed in the item responses are consistent with those hypothesized in the exemplars. However, there were six items that were eliminated from the assessment before the analysis; these items should be revised, tested in cognitive interviews, and used in the next assessment administration. Moreover, in the next iteration of design, the assessment should be administered to elementary and middle school students to avoid the ceiling effect observed with twenty-eight participants from an elite private high school and university.

Additional Forms of Reasoning to Be Assessed

As the assessment is further developed, it may also be expandable to additional important learning targets. In a previous study of children's naïve mechanistic reasoning (Bolger et al., 2012), we noted that children rarely paid attention to how far links moved, even when a paired contrast was used to draw their attention to this feature. Moreover, no child's

explanation of this phenomenon went beyond the noticing of an empirical pattern (e.g., when the brads are closer to each other, the link moves more). This may be because explaining the relative input to output distances relies upon mathematical relationships that were not apparent to the children.

It could be valuable to develop items that target (at least qualitatively) the distance relationship between the amount of input and output. This relationship blends mechanistic and quantitative reasoning. Bolger and colleagues (2011) showed that by mathematizing these levered systems, participants can develop an understanding of both mechanism and mathematics. For instance, being able to map the mathematics of circles onto the physical systems may both focus students' mechanistic reasoning and lend it additional precision. In the present study, 68% (n=17) of those who were assessed at the level of *tracing* on at least one item made a reference to the mathematics of circles during the cognitive interview. For example, participants used the following terms to explain the machine motion: "circle," "center of the circle," "radius," "circumference," "axis of rotation." These findings strongly suggest that the mathematization (Freudenthal, 1973; Kline, 1982) of these systems makes their mechanisms more visible.

The Stability of Mechanistic Reasoning (Research Question #2):
Can this assessment provide insight into the features of machines that are most likely to disrupt an individual's capacity to reason mechanistically?

In their work with mechanistic reasoning about simple machines, Bolger and colleagues (2012), Metz (1985, 1991), and Lehrer and Schauble (1998) have not addressed the extent to which mechanistic reasoning generalizes across machines within a single system (e.g.,

gears, levered machines). For example, to what extent does a participant's ability to reason mechanistically about one levered machine generalize to other similar machines? What supports or disrupts the ability for an individual to see multiple levered machines as variants of those that they can diagnose and causally trace? In this study, some items disrupted participants' abilities to diagnose and causally trace a machine's mechanistic elements from input to output, when they had previously exhibited this ability on other items. For example, all individuals (n=25) who were assessed at least once at the level of *tracing*, showed a decrease in their ability to perform at this level when diagnosing machines with class 3 levers and bent cranks. This shows that certain machine characteristics can disrupt the coordination of all of these mechanistic elements. diSessa (1993) provides an example of how Newton's third law of motion can be understood differently in two different contexts. He notes that students are more likely to cite the relevant "equal and opposite forces" when a book is supported by a person's hand, rather than a table. In the assessment developed in this study, what mechanistic elements are cued (and causally connected) seems dependent on processes that could be further investigated in subsequent research.

System Tracing

In order to diagnose this system, individuals must recognize the push-pull interactions of the various components as they trace the transmission of force. Similar diagnosis is essential in systems across engineering, physics, as well as in the designed world. Forbus (1987) models the cognitive processes involved in making observations and inferences about physical systems. Qualitative Process (QP) theory may be used to model how individuals reason about the

motion of these simple levered systems. In QP theory, the way time is segmented (i.e., the frequency with which observations are made) determines what inferences will be drawn about phenomena being observed. Individuals similarly segment time when inspecting the levers. How they segment their inspections of the machines makes the difference between seeing endpoints of motion (i.e., *linked direction*) or complete paths of levers (i.e., *rotation*). Forbus describes how a computer program, FROB, fills in gaps in sparse data. This ability may account for those participants who were able to recognize the global motion of the lever around the fixed pivot, but were unable to replicate the same causal coordination of the lever arms closer to, but on opposite sides of, the fixed pivot. These participants were able to indicate the correct global direction of motion of the lever (i.e., *lever arms*), based on the fixed pivot, but were unable to “fill in the gaps” (i.e., recognize *lever arms*) closer to the fixed pivot. The capacity to impute mechanisms based on other visible mechanisms and an understanding of the system seems critical to tracing. This form of system tracing is productive when diagnosing mechanisms in systems where forces are transmitted through visible components. Simple levered machines make good candidates for systems in which individuals can gain access to mechanistic reasoning. However, this form of system tracing is also fundamental to diagnosing mechanisms in other systems where forces are transmitted through components, such as those featured in mechanics and engineering.

APPENDIX A

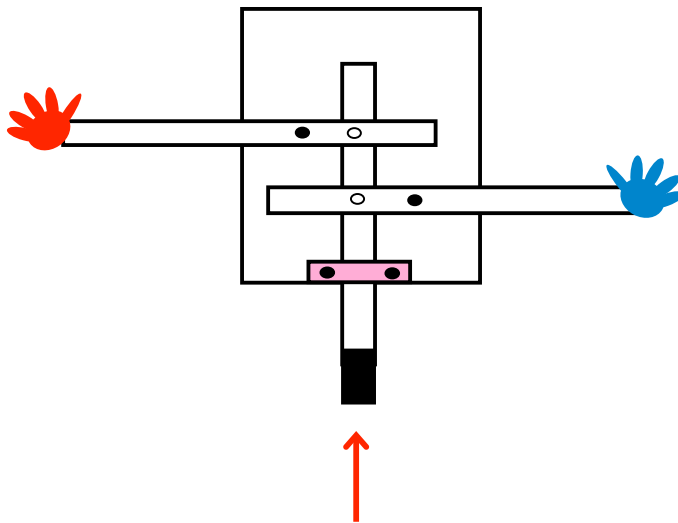
EXEMPLAR EXAMPLES

Hands Fixed Pivots -Opposite

Key: Fixed Pivot (attaches link(s) to base) ●

Floating Pivot (attaches link to link) ○

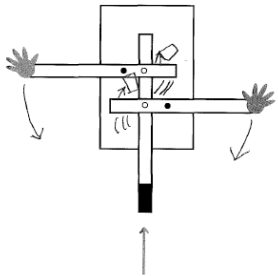
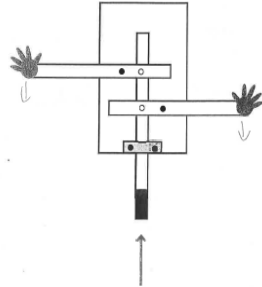
Draw how the **left hand** and the **right hand** would move if you pushed **UP** on the black part. (Draw an arrow starting at each hand and show how they will move).

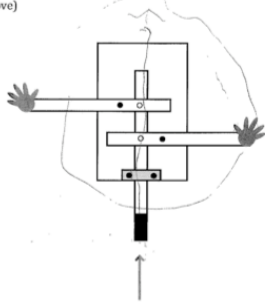


Item Exemplar

Table 5: Hands A- Fixed Pivots on Opposite Sides of the Input

This item assesses students' ability to use the mechanistic elements of *linked direction* and *rotation*. “No link” (NL) indicates an item response that does not provide any evidence of mechanistic reasoning (i.e., use of mechanistic elements). “Missing” indicates that item was left completely blank.

<p>1 Element</p>	<p>Rotation</p>	<p>Participant draws an arced path (they may show the incorrect direction). The location of this path must reasonably approximate fractions of circles either centered around the fixed or floating pivot.</p> <p><i>Note: Although these paths are actually centered around the fixed pivot, this element of mechanistic reasoning can be assessed without a clear understanding of the location of the center of rotation.</i></p>	 <p>The diagram shows a mechanical linkage with a central vertical input shaft and two horizontal output arms. A hand-drawn path is shown as a curved line starting from the top of the input shaft, curving to the right, then down, then left, and finally up, forming a partial circle. Arrows indicate the input motion is up and the output motions are down.</p>
	<p>Linked Direction</p>	<p>Participant draws the correct output motion of <u>both</u> outputs.</p>	 <p>The diagram shows the same mechanical linkage as above. The output arms are drawn with straight lines and arrows pointing downwards, indicating the correct output motion for both arms.</p>

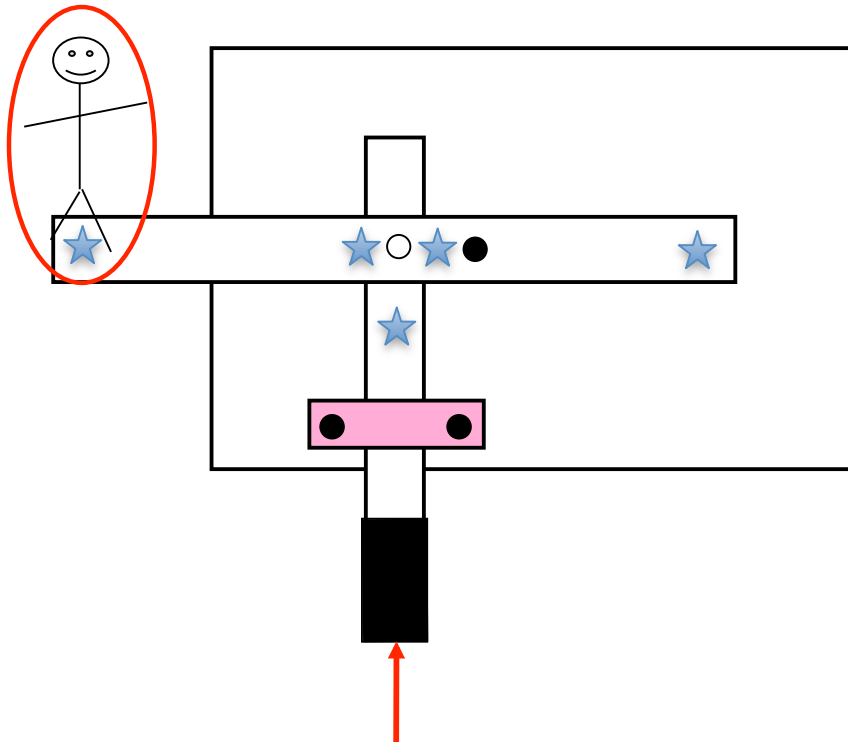
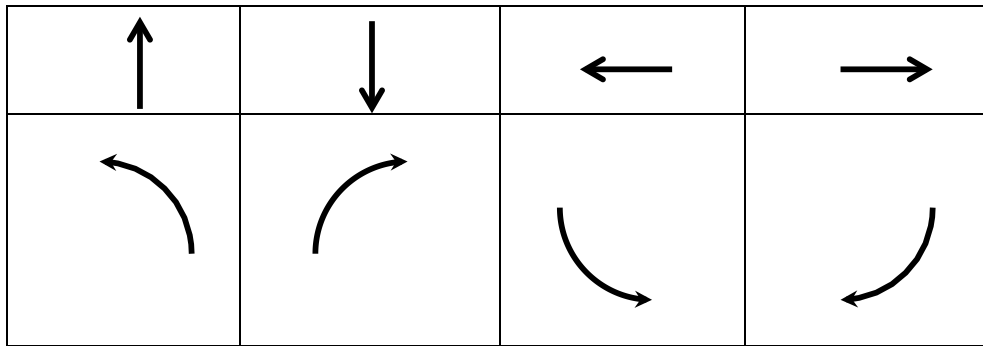
No Elements	Participant Uses No Mechanistic Element	No mechanistic elements are shown.	<p>black part (draw an arrow starting at each hand and show move)</p> 
No Link		It is unclear whether the participant understood the nature of the task.	"I don't know?"
Missing		Missing Response	

Sequential Tracing A1

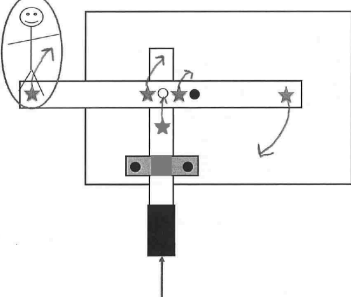
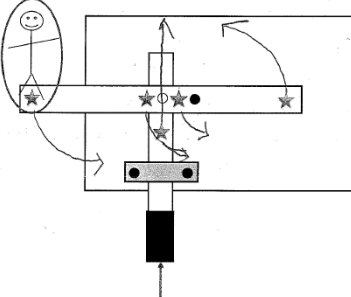
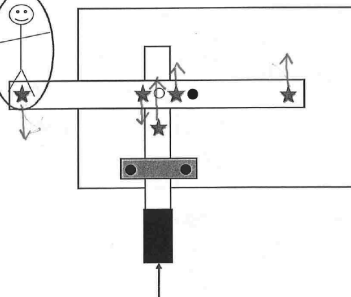
Key: Fixed Pivot (attaches link(s) to base) ●

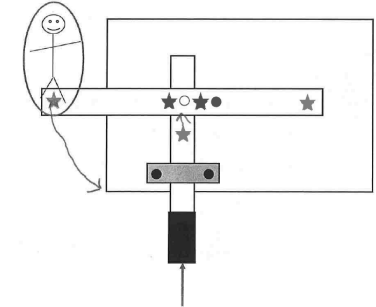
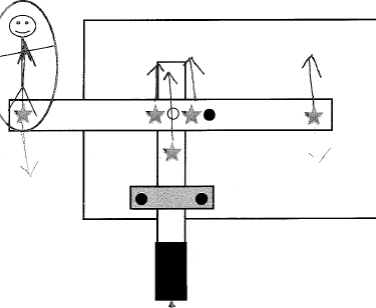
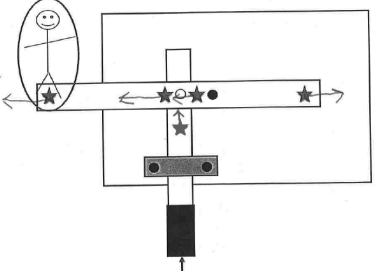
Floating Pivot (attaches link to link) ○

Draw an arrow, like one of these below, to show how each star would move if you pushed up on the black handle. (Draw an arrow starting at EACH star and show how they will move)



This item assesses students' ability to use the mechanistic elements of *linked direction*, *rotation*, *lever arms*, *constraint via the fixed pivot*, and *tracing*. No link (NL) indicates an item response that does not provide any evidence of mechanistic reasoning (i.e., use of mechanistic elements). "Missing" indicates that the item was left completely blank.

<p>4 Elements</p>	<p>Tracing</p>	<p>Student is assessed at the level of <i>constraint via the fixed pivot</i> and diagnoses motion correctly (and without gaps) on all stars from input to output.</p>	
<p>1 Element</p>	<p>Constraint via the Fixed Pivot</p>	<p>Participant correctly draws the opposite and/or rotary motion of the two closest points on opposite sides of the fixed pivot.</p>	
	<p>Lever Arms</p>	<p>Student draws arrows with opposite directions from stars on opposite sides of a lever's arms.</p> <p><i>To code lever arms alone the direction must be incorrect.</i></p>	

	<p>Rotation</p>	<p>Student draws arced paths (they may show the incorrect direction). However, the location of these paths must reasonably approximate fractions of circles either centered around the fixed or floating pivot.</p> <p><i>Note: Although these paths are centered around the fixed pivot, this element of mechanistic reasoning does not make this distinction.</i></p>	
	<p>Linked Direction</p>	<p>Student draws the correct input motion; the correct output motion is drawn at least once.</p>	
<p>No Elements</p>	<p>Student Uses No Mechanistic Elements</p>	<p>No mechanistic elements are shown.</p>	

NL		It is not clear if the student understood the nature of the task.	"I don't know"
Missing		Missing Response	

APPENDIX B

TABLES

Table 2-2					
<i>Item Coverage Matrix</i>					
Items\Levels	1	2	3	4	5
	Linked Direction (LD)	Rotation (R)	Lever Arms (LA)	Constraint Fixed Pivot (CFP)	Tracing (T)
1. Hands A- FP Opposite	1	1			
2. Hands A- FP Same	1	1			
3. Machine Prediction A1	1	1			
4. Machine Prediction A2	1	1			
5. Machine Prediction A3	1	1			
6. Machine Prediction A3'	1	1			
7. Machine Prediction B2	1	1			
8. Machine Prediction B2'	1	1			
9. Machine Prediction D1	1				
10. Machine Prediction D1'	1				

11.	Sequential Tracing A4	1	1	1	1	1
12.	Sequential Tracing A1	1	1	1	1	1
13.	Sequential Tracing A2	1	1	1	1	1
14.	Sequential Tracing A3	1	1	1	1	1
15.	Sequential Tracing A3'	1	1	1	1	1
16.	Sequential Tracing B1	1	1	1	1	1
17.	Sequential Tracing B1'	1	1	1	1	1
18.	Sequential Tracing B2	1	1	1	1	1
19.	Sequential Tracing C1	1	1	1	1	1
20.	Sequential Tracing D1	1	1	1	1	1
21.	Sequential Tracing D1'	1	1	1	1	1
22.	Sequential Tracing E1	1	1	1	1	1
23.	Sequential Tracing E2	1	1	1	1	1
24.	Sequential Tracing- Tracing Mechanism A	1	1	1	1	1
25.	Rotation- Constraint B				1	
26.	Constraint Fixed Pivot				1	

Item				
27.	Lever Arms	1	1	1
Prediction B				

Table 4-1*Descriptive and classical test theory (CTT) statistics.*

	HFP	MP	ST	ST	HFP	MP	MP	MP	MP	MP	MP	MP	ST	ST	ST	ST	ST	ST	ST	HFP	MP	ST
	O	A2	D1	E2	S	A1	A3	A3'	B2	B2'	D1	D1'	A1	A3	A3'	B1	B1'	B2	D1'	O	A2	D1
Mean of item	0.69	1	1.98	1.22	0.82	1.26	1.15	0.87	0.91	1.1	0.38	0.4	3.05	3.09	1.83	2.88	2.13	2.83	2.65	0.69	1	1.98
Median of item	1	1	1.5	0	0.5	1.5	1	1	1	1	0	0	3	3	2	3	3	3	3	1	1	1.5
Mode of item	0	1	0	0	0	2	2	1	0	1	0	0	3	3	2	4	0	3	0	0	1	0
SD	0.76	0.78	2.01	1.52	0.9	0.83	0.89	0.75	0.89	0.77	0.49	0.5	1.63	1.64	1.43	1.68	1.95	1.78	1.86	0.76	0.78	2.01
Variance	0.58	0.61	4.04	2.32	0.8	0.69	0.8	0.56	0.8	0.59	0.24	0.25	2.65	2.69	2.05	2.81	3.81	3.17	3.46	0.58	0.61	4.04
Item difficulty	0.34	0.5	0.4	0.24	0.34	0.5	0.99	0.61	0.41	0.63	0.58	0.43	0.61	0.62	0.37	0.58	0.43	0.57	0.53	0.34	0.5	0.4
Item discrimination	0.58	0.62	0.87	0.83	0.64	0.63	0.76	0.75	0.75	0.87	0.63	0.6	0.85	0.85	0.9	0.91	0.83	0.88	0.73	0.58	0.62	0.87

Table B-1*Item-step estimates and standard errors.*

Item	Item-step	Item-step Estimate	Standard Error
Hands Fixed Pivot- Opposite	Linked Direction	-0.859	0.216
	Rotation	0.859*	
Machine Prediction-A2	Linked Direction	-1.008	0.206
	Rotation	1.008*	
Sequential Tracing-D1	Linked Direction	0.843	0.236
	Rotation	-0.287	0.244
	Lever Arms	-2.005	0.248
	Constraint via the Fixed Pivot	-0.039	0.275
	Tracing	1.488*	
Sequential Tracing-E2	Linked Direction	0.845	0.330
	Rotation	-1.740	0.348
	Lever Arms	-0.102	0.429
	Constraint via the Fixed Pivot	0.997*	
Hands Fixed Pivot-Same	Linked Direction	0.257	0.402

Pivot-Same			
	Rotation	-0.257*	
Machine			
	Linked Direction	-0.291	0.382
Prediction-A1			
	Rotation	0.291*	
Machine			
	Linked Direction	-0.086	0.385
Prediction-A3			
	Rotation	0.086*	
Machine			
	Linked Direction	-1.144	0.318
Prediction-A3'			
	Rotation	1.144*	
Machine			
	Linked Direction	-0.011	0.387
Prediction-B2			
	Rotation	0.011*	
Machine			
	Linked Direction	-1.099	0.337
Prediction-B2'			
	Rotation	1.099*	
Sequential			
	Linked Direction	-0.476	0.395
Tracing-A1			
	Rotation	-0.217	0.395
	Lever Arms	-2.809	0.389

	Constraint via the Fixed Pivot	1.817	0.438
	Tracing	1.684*	
Sequential Tracing-A3	Linked Direction	-1.100	0.396
	Rotation	-0.999	0.384
	Lever Arms	-0.699	0.352
	Constraint via the Fixed Pivot	1.674	0.456
	Tracing	1.124*	
Sequential Tracing-A3'	Linked Direction	0.285	0.396
	Rotation	-2.468	0.393
	Lever Arms	0.823	0.451
	Constraint via the Fixed Pivot	1.361*	
Sequential Tracing-B1	Linked Direction	-1.125	0.387
	Rotation	-0.923	0.386
	Lever Arms	-0.871	0.382
	Constraint via the Fixed Pivot	0.684	0.395
	Tracing	2.235*	
Sequential Tracing-B1'	Linked Direction	0.634	0.395

	Rotation	-0.921	0.403
	Lever Arms	-1.529	0.399
	Constraint via the Fixed Pivot	-0.658	0.396
	Tracing	2.475*	
Sequential	Linked Direction	-0.566	0.368
Tracing-B2			
	Rotation	-0.854	0.375
	Lever Arms	-1.457	0.365
	Constraint via the Fixed Pivot	1.348	0.417
	Tracing	1.529*	
Sequential	Linked Direction	0.990	0.406
Tracing- D1'			
	Rotation	-1.438	0.412
	Lever Arms	-1.689	0.409
	Constraint via the Fixed Pivot	0.444	0.422
	Tracing	1.693*	
Sequential	Linked Direction	-0.831	0.380
Tracing-E1			
	Rotation	-1.416	0.397
	Lever Arms	1.158	0.462
	Constraint via the Fixed Pivot	-0.970	0.495
	Tracing	2.059*	

Sequential	Linked Direction	0.389	0.370
Tracing-			
CMT			
	Rotation	-0.056	0.374
	Lever Arms	-2.878	0.366
	Constraint via the Fixed Pivot	0.709	0.362
	Tracing	1.837*	
<i>Note: *Item-step is constrained</i>			

Table B-2*Person ability estimates and standard errors.*

Person ID	Person Ability Estimates	Standard Errors
1	-0.34705	0.43664
2	-3.14015	0.99836
3	3.29444	0.85285
4	-0.5222	0.417
5	-2.63867	1.09581
6	0.05227	0.41603
7	0.49562	0.4328
8	-2.13783	0.79435
9	-1.95691	0.82123
10	-1.71847	0.56025
11	-0.5222	0.417
12	-1.61741	0.4983
13	-0.91794	0.38413
14	-0.55382	0.3396
15	-0.23414	0.31514
16	-2.42174	0.77759
17	-0.55382	0.3396
18	-3.1926	0.97751
19	0.35095	0.33043
20	2.86093	0.59541
21	0.96899	0.36331
22	-2.40931	0.59396

23	-0.19906	0.28522
24	-1.30784	0.32993
25	-0.36524	0.29291
26	-0.63456	0.30351
27	1.10296	0.36985
28	0.4637	0.34105
29	-3.48725	0.90121
30	-0.19906	0.28522
31	-3.56483	0.89996
32	-0.23434	0.32648
33	3.12318	0.85686
34	-2.74831	0.7554
35	-0.93038	0.40107
36	-0.77105	0.39706
37	0.61006	0.37338
38	-0.07532	0.30691
39	-1.35772	0.32608
40	-1.35772	0.32608
41	-0.16845	0.30421
42	0.9013	0.38854
43	-3.08351	0.718
44	-1.15338	0.31515
45	-1.82964	0.52778
46	-0.16498	0.32687
47	-2.63626	0.75904
48	0.186	0.3663

49	-3.35792	0.94078
50	-1.82964	0.52778
51	1.54509	0.46096
52	-0.61478	0.30989
53	-1.57048	0.377
54	1.01019	0.39328
55	0.71296	0.3782
56	1.34581	0.43191
57	-2.48939	0.60853
58	-1.31926	0.33737
59	-1.47999	0.43153
60	-1.93539	0.53608
61	0.51076	0.45296
62	0.67939	0.50069
63	-3.42773	0.92511
64	-0.13788	0.41165
65	2.19211	0.51053
66	-0.82444	0.3067
67	0.06255	0.2983
68	0.06255	0.2983
69	0.82634	0.35589
70	1.54399	0.42665
71	2.47496	0.55265
72	0.95388	0.35849
73	-0.87111	0.41739
74	-3.43475	0.94907

75	1.42886	0.53126
76	-3.43475	0.94907
77	0.10547	0.36249
78	-0.20271	0.31082
79	0.56707	0.37168
80	-0.10696	0.30901
81	-0.01058	0.31272
82	1.01624	0.40165
83	-0.10696	0.30901
84	-1.73772	0.47972
85	0.1032	0.43185
86	0.514	0.47528
87	-1.73772	0.47972
88	-3.44816	0.91306
89	-3.44816	0.91306
90	0.75073	0.39061
91	2.54647	0.60956
92	-0.10304	0.31256
93	1.43749	0.45584
94	1.9125	0.51971
95	1.24266	0.42782
96	0.60166	0.38099
97	-1.32292	0.34599
98	-0.96936	0.36091
99	-1.42241	0.43669
100	-1.25071	0.39575

101	-1.96059	0.62312
102	-2.45173	0.78357
103	-1.10367	0.37326
104	2.94662	0.58677
105	-0.09256	0.26739
106	0.62268	0.33351
107	1.88477	0.44948
108	0.13829	0.2891
109	1.35768	0.38424
110	1.51493	0.40872
111	0.73614	0.33965
112	-3.68772	0.88744

REFERENCES

- Baillargeon, R. (1987a). Object permanence in 3.5 and 4.5-month-old infants. *Developmental Psychology*, 23(5), 655-664.
- Baillargeon, R. (1987b). Young infants' reasoning about the physical and spatial properties of a hidden object. *Cognitive Development*, 2(3), 179-200.
- Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, 3(5), 133-140.
- Baillargeon, R. (1995). A model of physical reasoning in infancy. In C. Rovee-Collier & L. Lipsitt (Eds.), *Advances in infancy research* (Vol. 9, pp. 305-371). Norwood: Ablex.
- Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, 20, 191-208.
- Bolger, M., Kobiela, M., Weinberg, P.J., & Lehrer, R. (2012). Analysis of children's mechanistic reasoning about linkages and levers in the context of engineering design. *Cognition and Instruction*, 30(2), 170-206.
- Bolger, M., Weinberg, P., Kobiela, M., Rouse, R., & Lehrer, R. (2011, April). Embodied experiences as a resource for children's mechanistic and mathematical reasoning in an engineering curriculum. Paper presented at the National Association for Research in Science Teaching Annual International Conference: Orlando, FL.
- Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple choice items. *Educational Assessment*, 11(1), 33-63.
- Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition*, 9, 117-123.
- Clement, J. (1981). Students' preconceptions in introductory mechanics. *American Journal of Physics*, 50, 66-71.

- diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10(2/3), 105-225.
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8(4), 341-349.
- Ericsson, K. A., & Simon, H. A. (1993). *Verbal Reports as Data*. Cambridge, MA: MIT Press.
- Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Forbus, K. D. (1987). Interpreting observations of physical systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-17(3), 350-359.
- Freudenthal, H. (1973). *Mathematics as an Educational Task*. Dordrecht-Holland: D. Reidel Publishing Company.
- Ginsburg, H. P., Jacobs, S. F., & Lopez, L. S. (1998). *The Teacher's Guide to Flexible Interviewing in the Classroom: Learning What Children Know about Math*. Boston, MA: Allyn & Bacon.
- Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, 69(S3), S342-S353.
- Hagerty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Neurosciences*, 8(6), 280-285.
- Hagerty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology*, 18(5), 1084-1102.
- Henning, G. (1987). *A guide to language testing*. Los Angeles: Newbury House.

- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141-158.
- Heuvel-Panhuizen, M. V. D. (1994). Improvement of (didactical) assessment by improvement of problems: An attempt with respect to percentage. *Educational Studies in Mathematics*, 27, 341-372.
- Ioannides, C., & Vosniadou, S. (2002). The changing meanings of force: From coherence to fragmentation. *Cognitive Science Quarterly*, 2(1), 5-62.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.
- Kiel, F. (1979). The development of the young child's ability to anticipate the outcomes of simple causal events. *Child Development*, 50(2), 455-462.
- Kim, K., & Spelke, E. S. (1992). Infants' sensitivity to effects of gravity on visible object motion. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2), 385-393.
- Kline, M. (1982). *Mathematics: The loss of certainty*. New York: Oxford University Press.
- Kobiela, M., Bolger, M., Weinberg, P., Rouse, R., & Lehrer, R. (2011, June). Mathematization and embodiment for reasoning about mechanism within an engineering curriculum. Paper presented at the Annual Meeting of the Jean Piaget Society: Berkeley, CA.
- Lehrer, R., & Schauble, L. (1998). Reasoning about structure and function: Children's conceptions of gears. *Journal of Research in Science Teaching*, 35(1), 3-25.
- Leslie, A. M. (1984). Infant perception of a manual pick-up event. *British Journal of Developmental Psychology*, 2(1), 19-32.
- Leslie, A. M. (1982). The perception of causality in infants. *Perception*, 11(2), 173-186.

- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1-25.
- Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34(4), 207-218.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Metz, K. E. (1991). Development of explanation: Incremental and fundamental change in children's physics knowledge. *Journal of Research in Science Teaching*, 28(9), 785-797.
- Metz, K. E. (1985). The development of children's problem solving in a gears task: A problem space perspective. *Cognitive Science*, 9(4), 431-471.
- Piaget, J., Inhelder, B., & Szeminska, S. (1960). *The Child's Conception of Geometry*. New York, NY: Basic Books.
- Schwartz, D. L. (1999). Physical imagery: Kinematic versus dynamic models. *Cognitive Psychology*, 38(3), 433-464.
- Schwartz, D. L. (1995). Reasoning about the referent of a picture versus reasoning about the picture as the referent: An effect of visual realism. *Memory & Cognition*, 23(6), 709-722.
- Schwartz, D. L., & Black, J. B. (1996a). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology*, 30(2), 154-219.
- Schwartz, D. L., & Black, J. B. (1996b). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science*, 20(4), 457-497.

- Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(1), 116-136.
- Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47(1), 1-51.
- Smith, J., diSessa, A., & Rochelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, 3, 115-163.
- Spelke, E. S. (1991). Physical knowledge in infancy: Reflections on Piaget's theory. In S. Carey & R. Gelman (Eds.), *The Epigenesis of Mind: Essays on Biology and Cognition* (pp. 133-169). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Spelke, E. S., Katz, G., Purcell, S. E., Ehrlich, S. M., & Breinlinger, K. (1994). Early knowledge of object motion: continuity and inertia. *Cognition*, 51, 131-176.
- Stevens, R., & Hall, R. (1998). Disciplined perception: Learning to see in technoscience. In M. Lampert & M. L. Blunk (Eds.), *Talking Mathematics in School*. Cambridge: Cambridge University Press.
- Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 37.
- Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest 2.0: General item response modeling software [computer program manual]* Camberwell, Vic: ACER Press.