

2019

## Assessing Mechanistic Reasoning: Supporting Systems Tracing

Paul J. Weinberg

Oakland University, [weinberg@oakland.edu](mailto:weinberg@oakland.edu)

Follow this and additional works at: <https://docs.lib.purdue.edu/jpeer>



Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Elementary Education Commons](#), and the [Science and Mathematics Education Commons](#)

---

### Recommended Citation

Weinberg, P. J. (2018). Assessing Mechanistic Reasoning: Supporting Systems Tracing. *Journal of Pre-College Engineering Education Research (J-PEER)*, 9(1), Article 3.

<https://doi.org/10.7771/2157-9288.1182>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact [epubs@purdue.edu](mailto:epubs@purdue.edu) for additional information.

This is an Open Access journal. This means that it uses a funding model that does not charge readers or their institutions for access. Readers may freely read, download, copy, distribute, print, search, or link to the full texts of articles. This journal is covered under the [CC BY-NC-ND license](#).

---

## Assessing Mechanistic Reasoning: Supporting Systems Tracing

### Abstract

Reasoning about mechanism is central to disciplined inquiry in science and engineering and should thus be one of the foundations of a science, technology, engineering, and mathematics education. In addition, mechanistic reasoning is one of the core competencies listed in the Next Generation Science Standards (NGSS) Engineering Concepts and Practices (NGSS Lead States, 2013). Mechanistic explanations focus on the processes that underlie cause–effect relationships and consider how the activities of system components affect one another.

While some assessment work has been accomplished in engineering education, to date mechanistic reasoning is an area where limited assessment development has been accomplished for pre-college populations. The data in this study come from the calibration of the Assessment of Mechanistic Reasoning Project (AMRP) (Weinberg, 2012), designed to diagnose individuals' mechanistic reasoning about systems of levers. This assessment presents a domain-specific characterization of mechanistic reasoning and illuminates what is easy and difficult about this form of reasoning. The study participants included elementary, middle, and high school students as well as college undergraduates and adults without higher education. Within this calibration study, item analyses, reliability, and validity measures were conducted using item response theory modeling; the AMRP assessment was found to have high reliability and validity. In addition, this study shows that machine characteristics such as number of levers, lever type, and arrangement of levers can affect the difficulty of mechanistic reasoning.

### Keywords

engineering education, assessment, science education

### Document Type

Article



## Assessing Mechanistic Reasoning: Supporting Systems Tracing

Paul J. Weinberg

*Oakland University*

---

### Abstract

Reasoning about mechanism is central to disciplined inquiry in science and engineering and should thus be one of the foundations of a science, technology, engineering, and mathematics education. In addition, mechanistic reasoning is one of the core competencies listed in the Next Generation Science Standards (NGSS) Engineering Concepts and Practices (NGSS Lead States, 2013). Mechanistic explanations focus on the processes that underlie cause–effect relationships and consider how the activities of system components affect one another.

While some assessment work has been accomplished in engineering education, to date mechanistic reasoning is an area where limited assessment development has been accomplished for pre-college populations. The data in this study come from the calibration of the Assessment of Mechanistic Reasoning Project (AMRP) (Weinberg, 2012), designed to diagnose individuals' mechanistic reasoning about systems of levers. This assessment presents a domain-specific characterization of mechanistic reasoning and illuminates what is easy and difficult about this form of reasoning. The study participants included elementary, middle, and high school students as well as college undergraduates and adults without higher education. Within this calibration study, item analyses, reliability, and validity measures were conducted using item response theory modeling; the AMRP assessment was found to have high reliability and validity. In addition, this study shows that machine characteristics such as number of levers, lever type, and arrangement of levers can affect the difficulty of mechanistic reasoning.

*Keywords:* engineering education, assessment, science education

---

### Introduction

Reasoning about mechanism is foundational to disciplined inquiry in science and engineering; thus, it should be one of the foundations of a science, technology, engineering, and mathematics (STEM) education (Bolger, Kobiela, Weinberg, & Lehrer, 2012; National Research Council [NRC], 2011; Russ, Scherr, Hammer, & Mikeska, 2008; Weinberg, 2017a, 2017b). The NRC (2009) indicates the tight connection between engineering principles, disciplinary knowledge, and disciplinary practices (e.g., mechanistic reasoning). For example, K–12 engineering education should emphasize engineering design. They indicate that the design process is “open to the idea that a problem may have many possible solutions ... [and provide] a meaningful context for learning scientific, mathematical, and technological concepts” (p. 4). Supporting students to engage in mechanistic reasoning requires opportunities for students to reason in varied and diverse ways, consider multiple forms of reasoning (e.g., reasoning about components, structure, and mechanisms), and construct multiple solutions. In addition, the engineering design process further supports the development of mechanistic reasoning within the learning of scientific phenomena. Accordingly, the Next Generation Science Standards (NGSS) Engineering Concepts and Practices (NGSS Lead States, 2013) include mechanistic reasoning as one of their core competencies and describe the “commitment to integrate engineering design into the structure of science education by raising engineering design to the same level as scientific inquiry” (p. 1). Moreover, the NRC nominates systems thinking as an important engineering habit-of-mind. While the development of mechanistic reasoning begins with a context-specific focus on

individual systems (e.g., simple levers, gears), as individuals have opportunities for further reasoning about these systems, mechanistic reasoning may become less localized and more systematized.

Although the nation seems to be developing a new emphasis around engineering education for K–12 students, a consensus around what that should entail has not yet emerged. While some significant assessment work has been accomplished in engineering education to date (e.g., Marra & Bogue, 2006; Purzer, Douglas, Folkerts, & Williams, 2017), further opportunities for assessment development for pre-college populations will occur as states continue to adopt and implement the NGSS. Moreover, advances in the understanding and measurement of learning bring new assumptions into play and offer the potential for richer and more coherent assessments (e.g., Gearhart & Saxe, 2004; Lehrer, Kim, Ayers, & Wilson, 2014; Pellegrino, Baxter, & Glaser, 1999). Specifically, mechanistic reasoning is an area where limited assessment work has been accomplished to date, especially in engineering contexts.

Mechanistic explanations focus on the processes that undergird causal relationships and consider how the elements (and the relations between these elements) of system components affect one another. Machamer, Darden, and Craver (2000) note that “[c]omplete descriptions of mechanisms exhibit productive continuity without gaps from the set up to terminal conditions” (p. 3). Lehrer and Schauble (1998) interviewed second- and fifth-grade students, within engineering tasks, to assess their reasoning about the mechanics of gears. These researchers characterized mechanistic explanations of these systems as those that described the transmission of motion as occurring through the interaction of the gear teeth. Although the majority of participants did not engage in mechanistic explanations, fewer were able to engage in systems thinking. However, those who did were able to use their reasoning about simple systems of gears to describe the operation of an eggbeater and a bicycle.

Russ et al. (2008) have reported on mechanistic reasoning from student verbal and written explanations within classroom activity and flexible interviews, where codes are applied to student conversational turns. Russ and colleagues’ framework and discourse analysis tool take a domain-general perspective on characterizing mechanistic reasoning within student explanations of scientific phenomenon. Domain-general theories of development suggest that individuals are born with cognitive mechanisms that support and guide generalized learning, regardless of the type of information being learned. However, domain-specific theories argue that many aspects of cognition are supported by specialized learning devices. The Assessment of Mechanistic Reasoning Project (AMRP) is developed from a domain-specific perspective on knowledge development because learners do not simply apply general forms of reasoning algorithmically; these forms of reasoning are tuned

to and affected by qualities of the devices that individuals are diagnosing. Accordingly, the affordances of diagnosing the mechanisms of systems of levers are not just that they are ubiquitous, but that they are open and inspectable. This quasi-transparency suggests that individuals of all ages and educational backgrounds are likely to have the capacity to reason about them.

Bolger et al. (2012) developed and verified mechanistic elements that were specific to systems of levers, from children’s explanations of their motion and their workings, as well as those of professionals in engineering and physics. The data were taken from one-on-one flexible interviews (Ginsburg, Jacobs, & Lopez, 1998). The present study reports on data from the AMRP (Weinberg, 2012), an assessment developed using item response theory (IRT) modeling that leverages children’s early capacities to reason mechanistically about properties of mechanical objects. In psychometrics, IRT is used in the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables. It is based on the application of mathematical models to testing data.

The AMRP assesses individuals’ capacities to mechanistically parse simple systems of levers, while characterizing their forms of reasoning. This assessment introduces students to mechanical principles through the mechanistic tracing of these simple systems; this provides a foundation for the building of knowledge about mechanical systems.

There are presently few assessments that leverage children’s early capacities to reason about properties of mechanical objects, promote domain-specific reasoning about mechanism, and support the engineering design process as well as engineering habits-of-mind. The American Association for the Advancement of Science (2011), through Project 2061, has developed an item bank that is aligned with current science standards and informed by the “misconceptions” literature; however, none of the items focus on mechanistic reasoning. The misconceptions literature states that adults hold consistent and erroneous beliefs about the physical world and that many of these beliefs are highly resistant to change by instruction (e.g., Caramazza, McCloskey, & Green, 1981). However, Smith III, diSessa, and Roschelle (1993) note that viewing novice forms of reasoning as misconceptions may be misleading. For example, diSessa (1993) 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. Accordingly, the items in the AMRP attend to the resources that students *do* have available to reason about this content.

Currently, the most widely used assessment of ideas about force and motion is the Force Concept Inventory (FCI) (Hestenes, Wells, & Swackhamer, 1992). This instrument 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 between a large truck and a compact car in terms of a "conflict metaphor" in action. In contrast, the assessment described here 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 domain-specific mechanistic tracing may provide a foundation for the building of important knowledge about mechanical systems. In addition, the AMRP supports the engineering design process. For example, items in this assessment (a) have multiple solutions, (b) provide meaningful contexts for learning scientific and mathematical concepts, and (c) promote systems thinking, modeling, and analysis (NRC, 2009). In addition, this assessment incorporates numerous engineering habits-of-mind, including systems thinking, creativity, collaboration, and communication.

### Research Questions

Through the development and administration of this assessment, the following research questions are addressed. (1) How can mechanistic reasoning be characterized with a standard paper-and-pencil assessment? (2) Can this assessment provide insight into what aspects of mechanistic reasoning are difficult?

### Assessment Development

An assessment design begins with the specification of the construct. The construct for the AMRP was developed according to the research literature on causal reasoning in infancy (Borton, 1979; Leslie & Keeble, 1987), early childhood (Bullock, Gelman, & Baillargeon, 1982; Gopnik, Sobel, Schulz, & Glymour, 2001; Nazzi & Gopnik, 2000), and adulthood (Schauble, 1990, 1996). The construct was also informed by literature about the difficulty individuals have reconciling their intuitions about causality with forms of mechanistic explanation valued by disciplines (Abrams & Southerland, 2001; Chin & Brown, 2000; Hmelo-Silver & Pfeffer, 2004; Talanquer, 2010). Finally, the construct addressed what participants find difficult when reasoning about simple mechanical systems (Bolger et al., 2012; Lehrer & Schauble, 1998; Metz, 1985; Weinberg, 2017b). This analysis resulted in a distinct construct focusing on mechanistic reasoning about systems of levers; this construct level has been called *mechanistic tracing*. This construct

measures an individual's parsing of simple systems of levers, diagnosing the mechanistic elements described by Bolger and colleagues (2012): (a) related direction (i.e., attention to the coordinated direction of the input and output of a linkage), (b) rotation (i.e., attention to the rotary motion of the levers), (c) lever arms (i.e., attention to the coordinated opposite motion of the two lever arms), 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). Finally, *tracing* was determined according to the following criteria: each mechanistic element within a system was correctly diagnosed in sequence. These mechanistic elements are ordered as levels in the construct (Table 1). In addition, the five construct levels, descriptions, and examples are presented.

Construct levels are ordered according to their hypothesized difficulty. The actual difficulty ordering will be determined through IRT analysis. This difficulty ordering was developed by Bolger et al. (2012) according to the frequency with which each mechanistic element was cited within student explanations. In addition, the ordering was also based on theoretical considerations concerning the machines' workings described by Weinberg (2012). The mechanistic elements are ordered according to the following hypothesized difficulty (from least to most difficult): *related direction*, *rotation*, *lever arms*, *constraint via the fixed pivot*, and *tracing*.

### Item Design

After the construct levels and associated performances were specified (Table 1), 21 assessment items were developed. The AMRP is composed of items with short-answer questions that require participants to draw predicted lever motion (e.g., Appendix A, Item).

### Developing Scoring Exemplars

Once items were developed, *scoring exemplars* (i.e., scoring guides that relate item responses to the construct map) were created. Scoring exemplars describe and provide examples of potential participant responses for each item; these responses are aligned with construct levels (Appendix A, Exemplar). An item's exemplar is structured like the construct map; it is ordered from the least to most sophisticated (or difficult) response. However, in an exemplar only those construct levels relevant to that particular item are represented.

### Method

#### Participants

The participant groups that comprise the sample are presented in Table 2.

Table 1  
Construct map: mechanistic tracing.

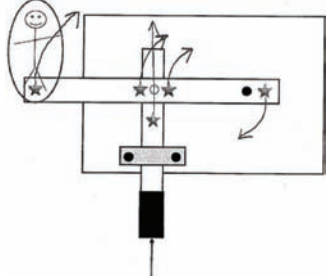
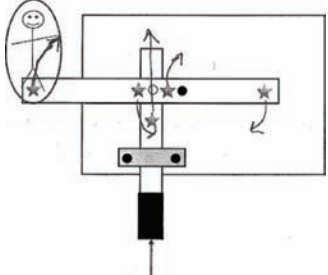
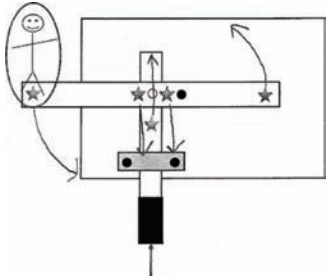
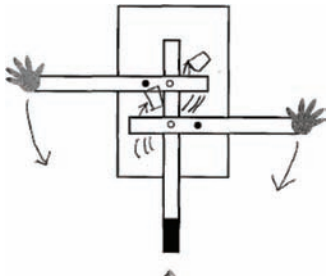
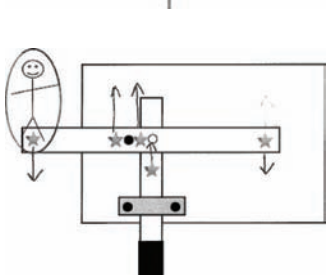
Level	Mechanistic element	Mechanistic element descriptions	Mechanistic element example
5	Tracing	Participant predicts all the mechanistic elements sequentially from input to output.	
4	Constraint via the fixed pivot	Participant draws the opposite and/or rotary motion of the two closest points on opposite 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. However, the location of these paths must reasonably approximate fractions of circles centered around either the fixed or floating pivot(s).	
1	Related directions	Participant draws the coordinated motion of input and output(s).	

Table 2  
Participants.

Participants	Number Included in analysis
Elementary school students	28 (female = 17)
Middle school students	25 (female = 16)
High school students	20 (female = 4)
University undergraduates (non-science majors)	16 (female = 13)
University undergraduates (engineering majors)	13 (female = 5)
Adults (without college education)	10 (female = 8)

As described in Weinberg (2017b), the elementary, middle, and high school students came from public and private schools in the southeastern USA. The university undergraduates came from three universities, two in the southeastern and one in the midwestern USA. Of the two universities in the southeastern USA, one is a highly ranked private university and the other is a large public university. The university in the midwestern USA is a highly ranked private liberal arts college. The public elementary, middle, and high schools belong to Centennial Public School District (a pseudonym). The percent of children attending these three schools qualifying for free or reduced lunch ranges between 60 to 90 from year to year. The adults without college degrees were recruited from the custodial staff at the highly ranked university in the southeastern USA. Study participants represent various ethnic backgrounds and life experiences. The diverse sample was chosen in order to assess and characterize mechanistic reasoning across age, socio-economic status, and experience. Although assessment items are typically calibrated with more homogenous groups, it was important to populate the construct map with diverse forms of reasoning. This served to show greater applicability of the instrument.

### Procedure

Each assessment administration was completed during one day and lasted an average of 37.5 minutes (ranging from 17 minutes to 78 minutes). The elementary school students averaged 34.9 minutes, the middle school students averaged 30.7 minutes, the high school students averaged 42.3 minutes, the college undergraduates averaged 38.9 minutes, and the adults averaged 48.7 minutes. These sessions were recorded using one camera, with a table microphone and were digitally rendered for further analysis.

The assessment was presented to participants on one of seven test forms. Elementary and middle school students were instructed to complete ten items per form, while high school students, undergraduates, and non-college educated adults were instructed to complete fifteen items per form. Five items were indicated in each form that elementary and middle school students were instructed to skip in order to avoid interview fatigue. Using item difficulty estimates from a previous study (Weinberg, 2017b), those items that were skipped by elementary school students did not

have different mean item difficulty estimates ( $M = -0.03$  logits) from those that were not skipped ( $M = -0.08$ , one-tailed  $t$ -test). Thus, the assessment was identically difficult for all age groups.

The AMRP items required respondents to draw predicted motion. There were 21 items in which *related direction* and *rotation* could be scored. In addition, there were 11 items in which *lever arms*, *constraint via the fixed pivot*, and *tracing* could be scored.

### Conduct of the Interview

While participants responded to each AMRP item, a clinical interview was conducted. The clinical interview was developed by Piaget (1951) to study individuals' knowledge and reasoning processes. Participants were asked to read the directions aloud for each item and think aloud as they responded. When the participant completed the item, s/he was asked for the rationale for the observed item response with interviewer probes. Finally, participants were asked to report any words that they found confusing as well as whether there was any confusion about the item. The interview was conducted in this manner to determine spontaneous thinking throughout participant interaction with each item as well as to assess mechanistic reasoning that was present, but possibly not elicited during the think aloud with interviewer probes.

### Analysis

#### Scoring items

Each item was scored according to its exemplar. A demonstration of how one item was scored is presented in Appendix A (Item & Exemplar). Exemplars contain three scoring categories: (1) the missing code (i.e., scores for missing responses), (2) the non-linking code (i.e., scores for responses that do not link to the construct map), and (3) construct linking codes (i.e., scores for responses that link to the construct map).

*Missing* The "missing" code was assessed when participant responses were not present (Appendix A, Exemplar). However, in this study participants responded to all items.

*No linkage* The "no link" code was assessed when participant responses provided evidence that they did not understand the nature of the task. This is seen in responses like "I don't know" (Appendix A, Exemplar).

*Construct linking codes* The construct linking codes include participant responses that (a) do not reason about mechanism, (b) reason about the four mechanistic elements (i.e., *related direction*, *rotation*, *lever arms*, *constraint via the fixed pivot*), and (c) causally connect all mechanistic elements from input to output (i.e., *tracing*) (Appendix A, Exemplar).

*No mechanistic elements are shown* These participant responses are not mechanistic. These may indicate participant reasoning about individual system components, machine structure, or idiosyncratic rules about machine motion. This is seen in the participant response scored at level 0 (Appendix A, Exemplar).

*Related direction* These participant responses indicate identification of the coordinated motion of lever input and output. This is seen in the participant response scored at level 1 (Appendix A, Exemplar).

*Rotation* These participant responses indicate identification of the rotary paths of the systems' levers. This is seen in the participant response scored at level 2 (Appendix A, Exemplar).

*Lever arms* These participant responses indicate identification of the coordinated opposite direction of the lever's arms. This is seen in the participant response scored at level 3 (Appendix A, Exemplar).

*Constraint via the fixed pivot* These participant responses indicate identification of the coordinated motion around the

fixed pivot. This is seen in the participant response scored at level 4 (Appendix A, Exemplar)

*Tracing* These participant responses indicate (a) identification of all mechanistic elements and (b) the sequential coordination of these elements from input to output. This is seen in the participant response scored at level 5 (Appendix A, Exemplar).

Participants were scored at the highest level (i.e., most difficult mechanistic element) where they achieved competency. For example, if a participant was assessed at the levels of both *rotation* and *related direction* on an item, they were assessed at the level of *rotation*. To be scored at the level of *tracing*, participants must have indicated a causal coordination of all elements. An outside researcher scored 10% of the total items. The agreement was 95% (Table 3). Wilson and Case (2000) indicate, in a study of a similarly formatted assessment instrument, that this is within the standard of 85%. The range of interrater agreement, disaggregated by item, ranges from 88% (Sequential Tracing-D1) to 100% (e.g., Sequential Tracing-E2) (Figure 1).

#### *Coding the clinical interview*

In order to investigate participant thinking during the AMRP administration, talk and gesture were coded according to an analytic framework used in a previous study (Bolger et al., 2012). A participant's work on one item is defined as a "performance." Participants were coded at the highest level where they achieved competency within each performance. For example, if a participant was coded at the levels of both *constraint via the fixed pivot* and *tracing* within the same instance, they were reported at the level of

Table 3  
Interrater agreement for scored items and coded interviews, by item.

Item	Scorer IRR (%)	Instances	Coded IRR (%)	Instances
Hands Fixed Pivot-Opposite	100	48	100	20
Machine Prediction-A2	92	48	94	35
Sequential Tracing-D1	88	74	90	49
Sequential Tracing-E2	100	49	86	7
Hands Fixed Pivot-Same	88	16	75	8
Machine Prediction-A1	100	4	75	8
Machine Prediction-A3	100	16	100	8
Machine Prediction-A3'	92	24	100	16
Machine Prediction-B2	93	28	94	16
Machine Prediction-B2'	100	20	100	16
Machine Prediction-D1	100	21	N/A	N/A
Machine Prediction-D1'	100	15	83	12
Sequential Tracing-A1	100	21	89	35
Sequential Tracing-A3	94	35	93	14
Sequential Tracing-A3'	100	28	91	35
Sequential Tracing-B1	93	28	89	28
Sequential Tracing-B1'	90	42	90	21
Sequential Tracing-B2	96	49	82	28
Sequential Tracing-D1'	100	21	89	28
Sequential Tracing-E1	89	35	86	28
Sequential Tracing-CMT	100	35	89	28
Total	95	667	90	440



**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 handle. (Draw an **arrow** starting at each hand and show how they will move)

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)

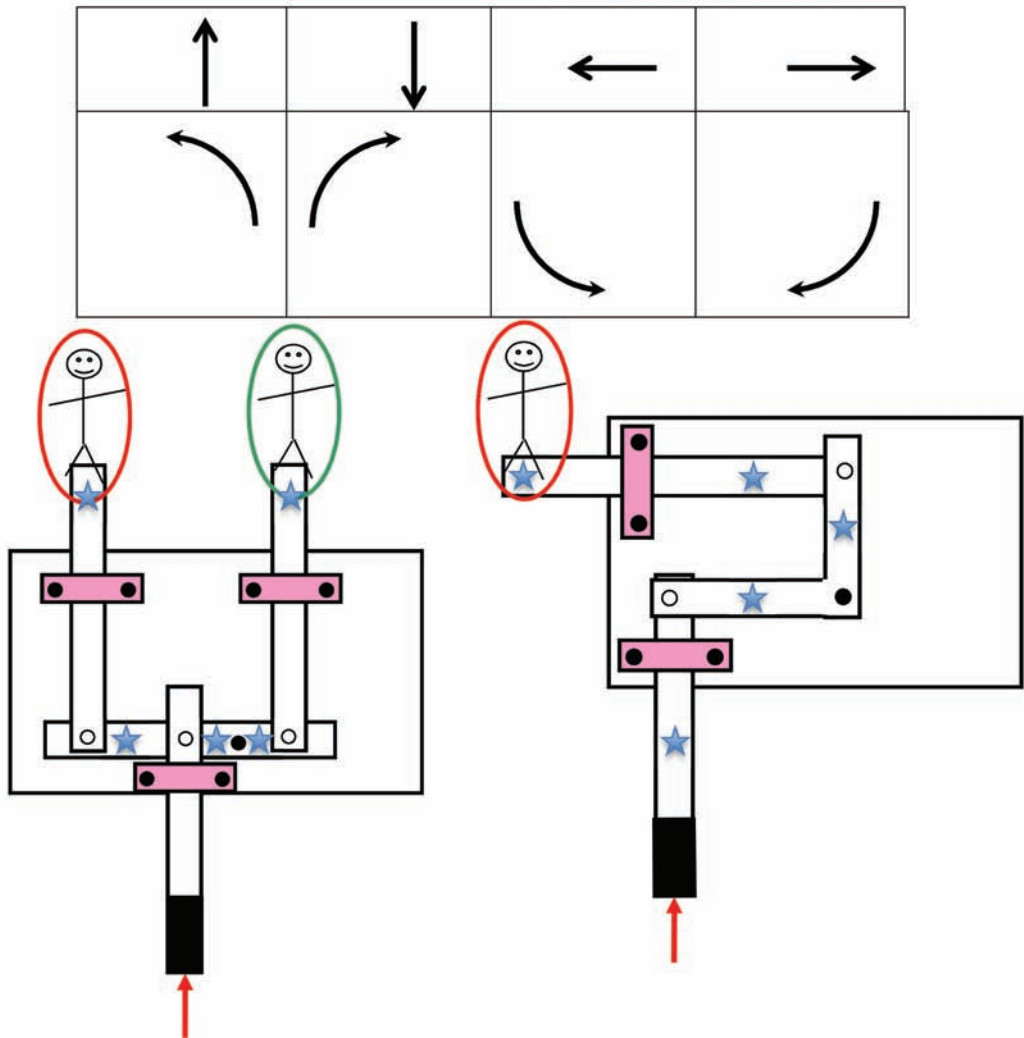


Figure 1. The IRR of all assessment ranges from 88% (Sequential Tracing-D1, left) to 100% (e.g., Sequential Tracing-E2, right).

tracing. All performances were coded using NVivo 11.0 software. An outside researcher coded 10% of the total instances. The agreement was 90% (Table 3).

*Item Response Theory Modeling*

To model the data from respondents, IRT was used and a partial credit model (PCM) was developed (Boone, 2016). A one-dimensional PCM was used to fit the data because the domain of mechanistic tracing is hypothesized to be one-dimensional, as shown in the construct map (Table 1). In addition, this assessment contains polytomous items; polytomous models contain items that have more than two

possible scores; common examples are Likert items (e.g., rated on an ordinal scale of 1 to 5) and partial credit items (e.g., an essay, which will typically be scored on an ordinal scale). IRT models typically assume that the item scores are integers. The polytomous categories are ordered, but without the assumption of equal distance between adjacent categories.

*Item analysis*

First, the item Wright Map is presented in order to analyze the behavior of the assessment items. 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 mean person ability and item difficulty are equal (i.e.,  $\theta_j - \beta_i = 0$ ). A person's ability in logits is his or her 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. For example, if an item has an item difficulty estimate of 1 logit, this indicates that those participants with person ability estimates of 1 logit have a 0.5 probability of correctly answering the item.

*Mean-square statistic*

The mean-square (MNSQ) statistics are presented in order to determine item fit. 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*

This section describes how it was determined that the assessment 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. In addition, the standard error of measurement (SEM) was calculated.

*Separation reliability*

In Rasch measurement, the person separation index is a summary of the separation as a ratio scale index comparing the spread of the measures with their measurement error. It indicates the measure of spread of the sample participants (or test items) in units of the test error in their measures. 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 has a maximum of one and a minimum of zero. Moreover, separation reliability indicates the extent to which the observed total variance,  $Var(\theta)$ , is accounted for by the model variance,  $Var(\hat{\theta})$ , which can be indicated by the following formula:

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

*Standard error of measurement*

An important difference between classical test theory (CTT) and IRT is the treatment of measurement error,

Table 4  
Item Wright Map results: mean item difficulty estimates, standard errors, and mechanistic elements assessed.

Item	Mean item difficulty estimate (logits)	Standard error	Mechanistic elements assessed
Hands Fixed Pivot-Opposite	0.587	0.115	RD, R
Machine Prediction-A2	-0.426	0.114	RD, R
Sequential Tracing-D1	0.171	0.079	RD, R, LA, CFP, T
Sequential Tracing-E2	0.323	0.109	RD, R, LA, CFP, T
Hands Fixed Pivot-Same	0.008	0.128	RD, R
Machine Prediction-A1	-0.547	0.133	RD, R
Machine Prediction-A3	-0.319	0.133	RD, R
Machine Prediction-A3'	0.259	0.133	RD, R
Machine Prediction-B2	0.286	0.131	RD, R
Machine Prediction-B2'	-0.391	0.135	RD, R
Machine Prediction-D1	0.711	0.144	RD, R
Machine Prediction-D1'	0.543	0.142	RD, R
Sequential Tracing-A1	-0.700	0.117	RD, R, LA, CFP, T
Sequential Tracing-A3	-0.760	0.115	RD, R, LA, CFP, T
Sequential Tracing-A3'	-0.169	0.120	RD, R, LA, CFP, T
Sequential Tracing-B1	-0.519	0.117	RD, R, LA, CFP, T
Sequential Tracing-B1'	0.134	0.105	RD, R, LA, CFP, T
Sequential Tracing-B2	-0.487	0.114	RD, R, LA, CFP, T
Sequential Tracing- D1'	0.578	0.113	RD, R, LA, CFP, T
Sequential Tracing-E1	0.923	0.113	RD, R, LA, CFP, T
Sequential Tracing-CMT	-0.205 <sup>a</sup>		RD, R, LA, CFP, T

Note. RD, related direction; R, rotation; LA, lever arms; CFP, constraint via the fixed pivot; T, tracing.

<sup>a</sup>Estimate is constrained.



**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)

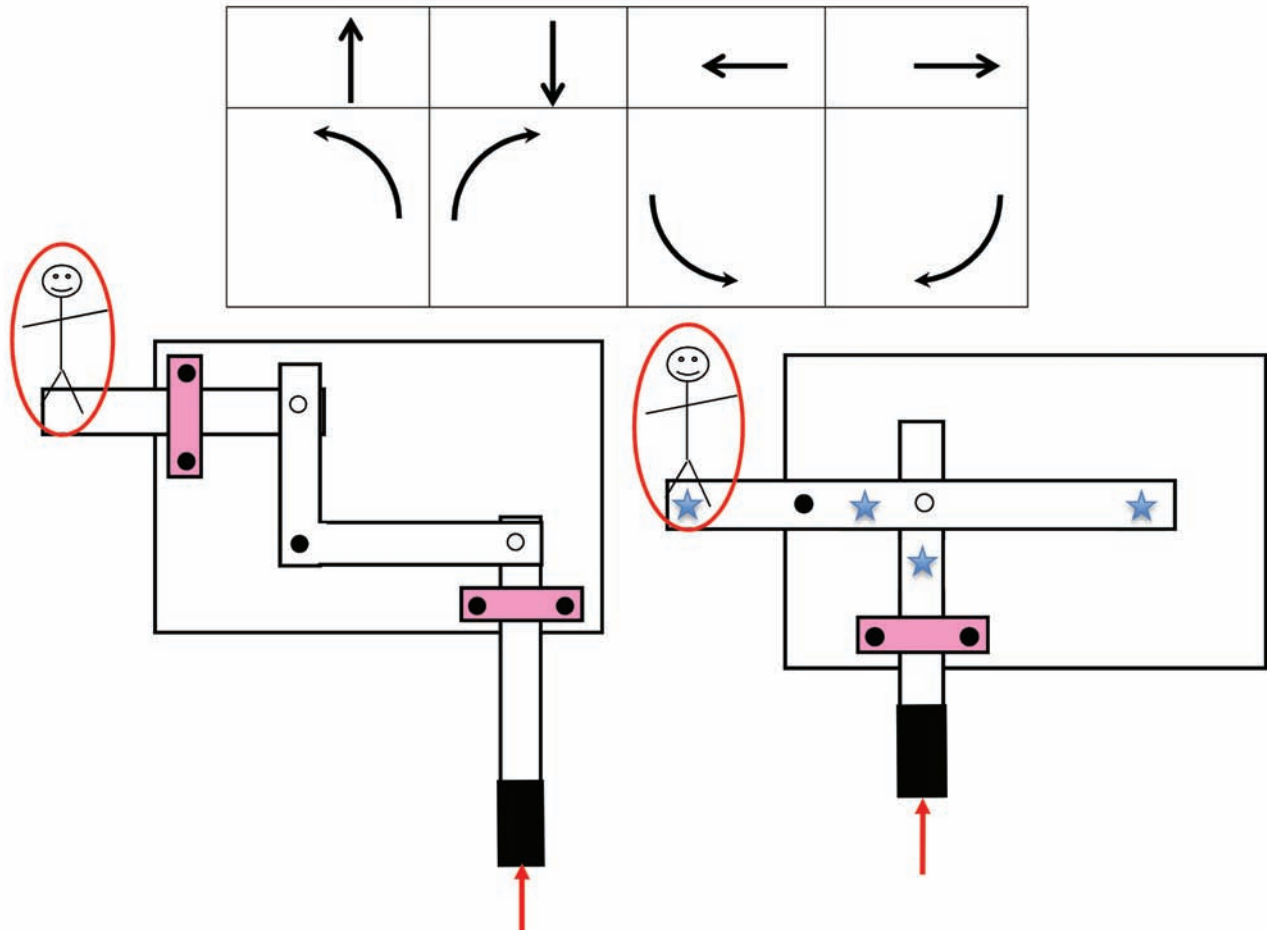


Figure 3. According to the item Wright Map, Sequential Tracing-E1 (left) is the most difficult item on the assessment (0.92 logits), while Sequential Tracing-A3' (right) is the easiest (-0.76 logits).

The item-step Wright Map presents the difficulty of endorsing each mechanistic element for each item. 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  $k-1$ . For example, they indicate the probability of a participant being diagnosed (on a particular item) at the level of *rotation* once s/he has been assessed at the level of *related direction*. Graphically, the item locations are the point at which the item response curves of two adjacent response categories (e.g., *constraint via the fixed pivot* vs. *tracing*) cross.

The item thresholds that are plotted on the Wright Map 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 five-category item,

the Thurstone threshold for score category 3 (*lever arms*) is the point at which participants are as likely to be observed below 3 (*rotation, related direction*) as being observed at or above 3 (*lever arms, constraint via the fixed pivot, or tracing*).

The item-step Wright Map is used to assess construct validity by empirically determining whether participant responses confirm hypotheses about the difficulty of the mechanistic elements from the construct map (Boone, 2016).

### Results

All AMRP items elicited responses that covered the entire construct. In addition, all item responses could be scored according to the exemplar levels. First, item analyses are shown. Then, reliability and validity measures are presented.

**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 handle. (Draw an arrow starting at each hand and show how they will move)

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)

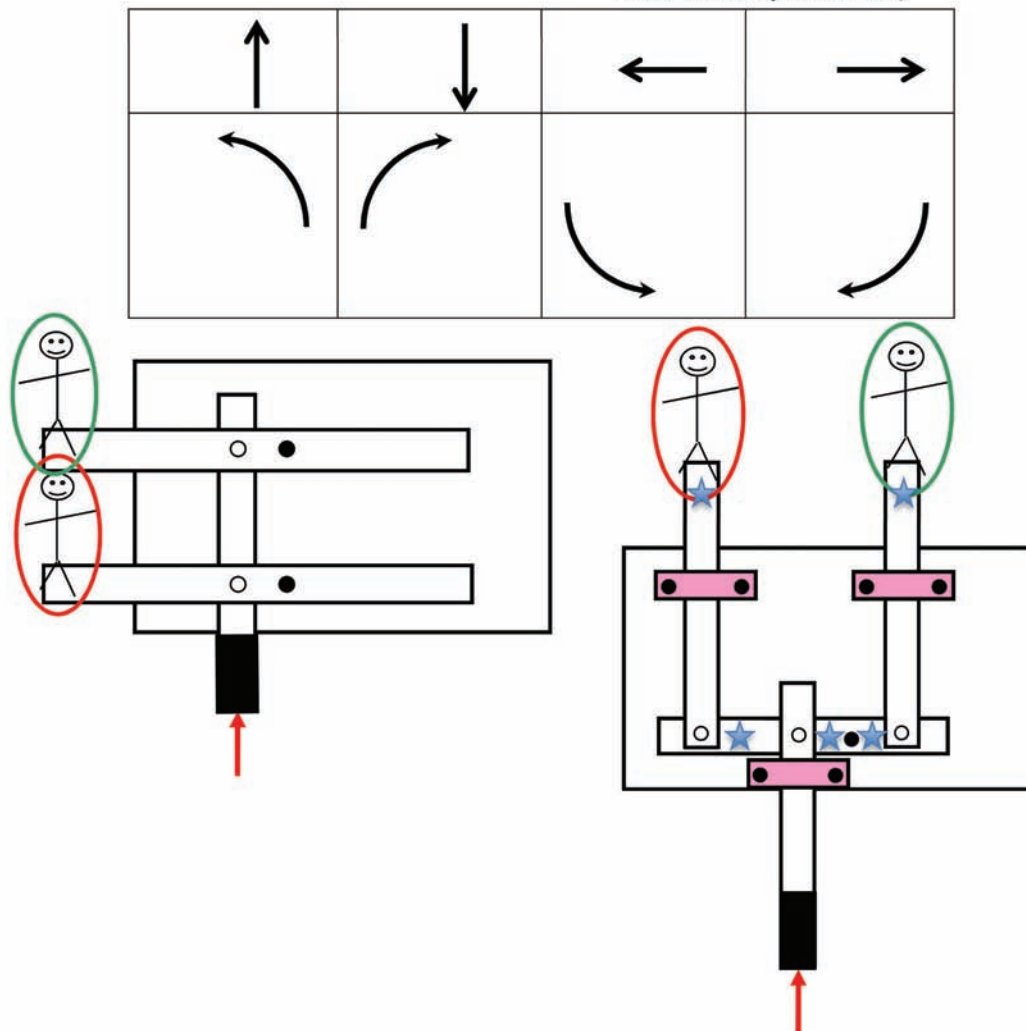


Figure 4. Machine Prediction-B2 (left) is a machine prediction item. Sequential Tracing-D1 (right) is a sequential tracing item.

*Item Analysis*

First, the item Wright Map is presented in order to analyze the behavior of the items. Then, the MNSQ statistics are presented in order to determine item fit.

*Item Wright Map*

The item Wright Map is presented in Figure 2. Results from this Wright Map (Table 4) make it possible to compare the mean difficulty of each item across the sample. The standard errors indicate the precision of the estimates. Sequential Tracing-E1 (STE1) (Figure 3) is the most difficult item, with a mean item difficulty of 0.92 logits.

The easiest item is Sequential Tracing-A3' (STA3') (Figure 3), with a mean item difficulty of -0.76 logits. Weinberg (2017b) reported that the following machine characteristics impact participants' diagnosis and causal connection of a machine's mechanisms: (a) item type, (b) number of levers, (c) lever type (e.g., class 1 levers), (d) arrangement of levers, and (e) 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 singular effect each has on participant mechanistic reasoning.

**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)

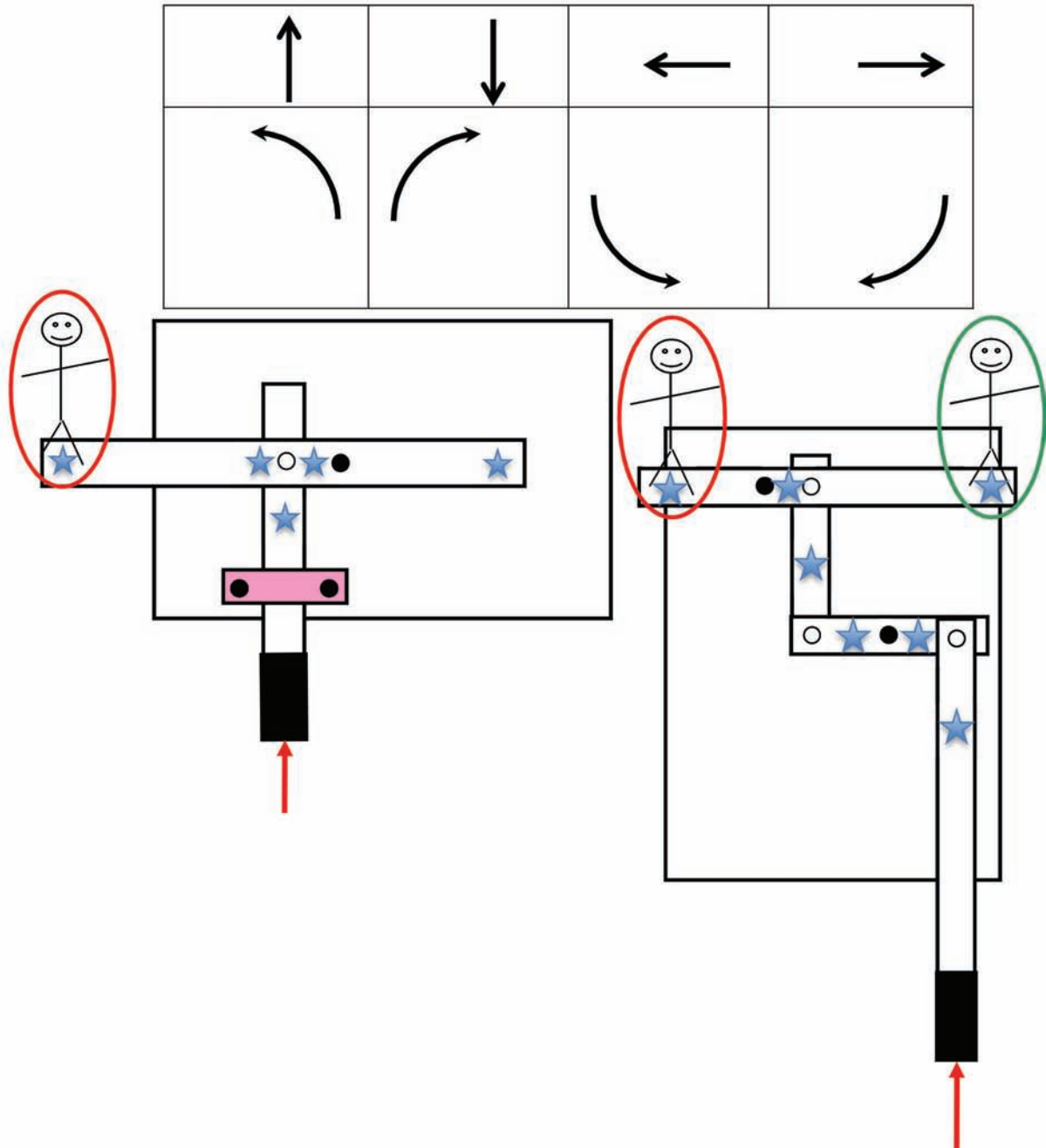


Figure 5. Sequential Tracing-A1 (left) is composed of two levers, while Sequential Tracing-CMT (right) is composed of four levers.



**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)

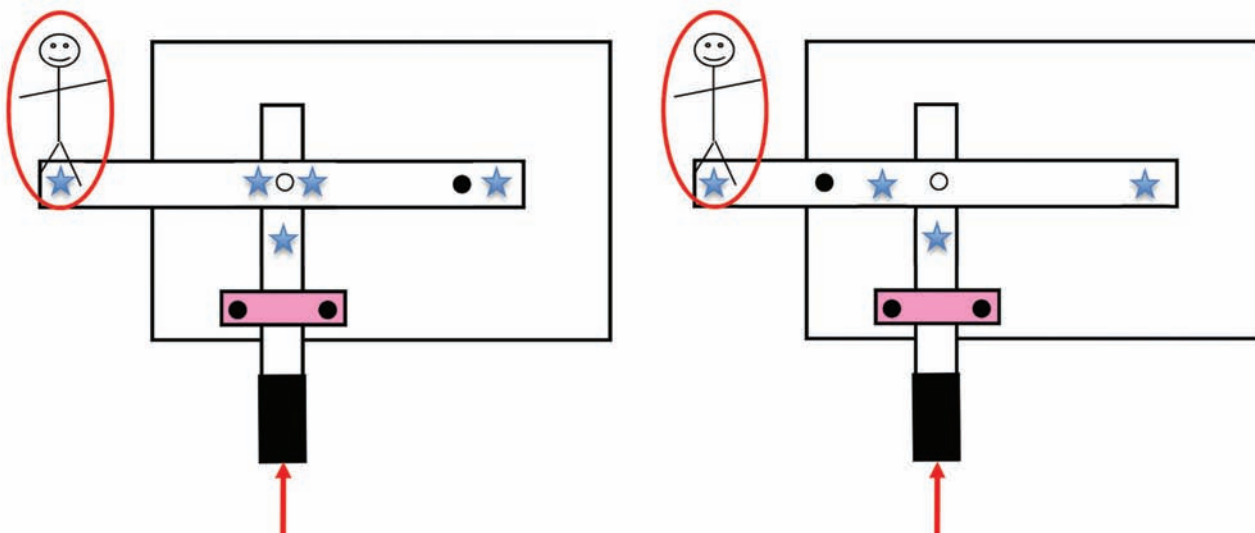
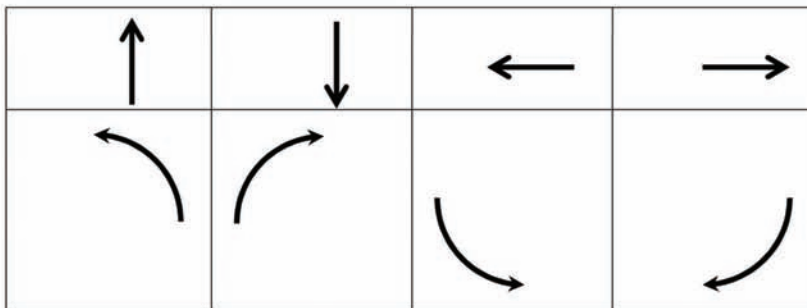


Figure 6. Sequential Tracing-A3 (left) is composed of a class 1 lever, while Sequential Tracing-A3' (right) is composed of a class 3 lever.

*Item type* Two different item formats were used on the AMRP: machine prediction and sequential tracing items (Figure 4). Machine prediction items challenge respondents to predict the motion of machine outputs; in addition, sequential tracing items challenge respondents to predict the motion of all the different machine parts from input to output. There was no difference in item difficulty estimates between these two types of items.

*Number of levers* Participants showed greater difficulty 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) (Figure 5).

*Lever type* Five items include machines composed of class 1 levers; in addition, five items include machines composed of class 3 levers (Figure 6). With class 1 levers,

the input and output move in the same direction; whereas, with class 3 levers the input and output move in opposite directions. Participants had more difficulty diagnosing mechanisms of class 3 levers ( $M = -0.03$  logits) than of class 1 levers ( $M = -0.41$  logits;  $p = 0.08$ , one-tailed).

*Arrangement of levers* Seven items were composed of one or more intermediate levers between the input and output (Figure 7), while fourteen items did not include any levers between the input and output. These seven items were more difficult ( $M = 0.43$  logits) to diagnose than the fourteen ( $M = -0.22$  logits;  $p = 0.001$ , one-tailed).

*Bent crank* Participants had difficulty diagnosing machines that used non-standard intermediate levers. One intermediate lever included on the AMRP was a bent crank. The most difficult item on the assessment was Sequential

**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)

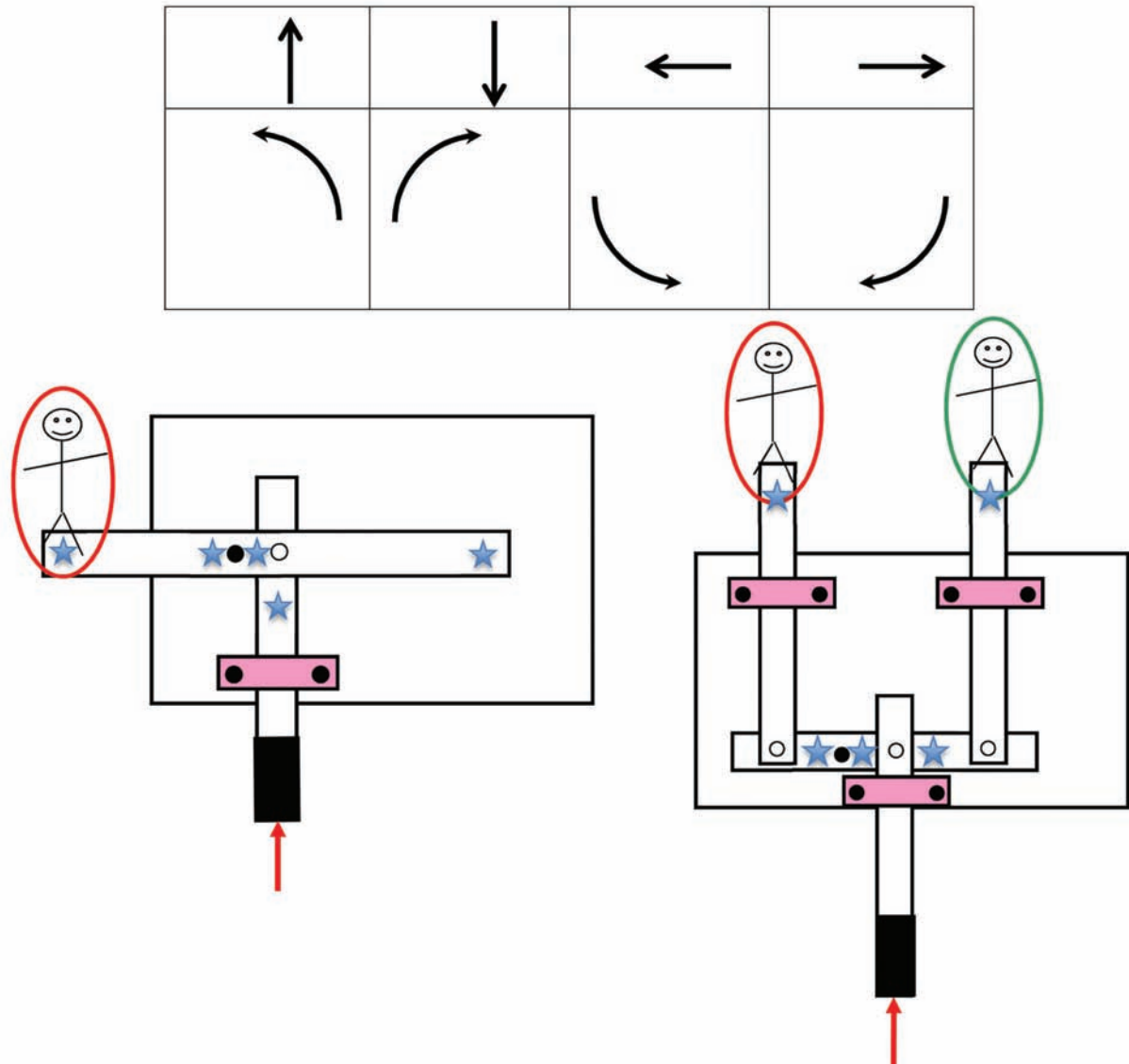


Figure 7. Sequential Tracing-A2 (left) is composed of no intermediate levers, while Sequential Tracing-D1' (right) is composed of one intermediate lever between input and output.

Tracing-E1 (STE1) (Table 4; Figure 3, left), which included a bent crank. Sequential Tracing-E2 (STE2) was another item that used a bent crank as an intermediate link; this was also one of the most difficult items on the AMRP.

*Mean-square statistic*

In Rasch analysis, item fit indices are reported for individual items. An item that has a MNSQ statistic equal to 1 indicates perfect fit. In general, a value between 0.75 and 1.33 indicates

good fit. The MNSQ statistic for all of the items is presented in Table 5. Of the 21 items, 17 (81%) are good fits. Two items, Hands Fixed Pivot-Opposite and Sequential Tracing-B1' (Figure 8) 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) (Figure 9). Wright, Linacre, and Gustafson (2009) present a professional standard for the interpretation of MNSQ statistics. They indicate that only Sequential Tracing-D1'



(MNSQ = 1.66) would produce a misfit that would be unproductive for assessment, but would not degrade the assessment. Thus, none of the items would compromise the assessment.

### Reliability

This section describes ways to investigate whether the assessment operates with sufficient consistency across individuals.

#### 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. The items on the AMRP have a separation reliability equal to 0.94, suggesting that most observed total variance is accounted for by the model variance.

#### Standard error of measurement

The SEM 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 (Figure 10). 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 relationship between the person ability estimate and the SEM indicates high reliability of the assessment.

### Validity

This section describes evidence that the AMRP targets the construct map. The correspondence between the item-step Wright Map and the construct map is discussed.

#### Item response and clinical interview

Items were scored according to the exemplars, whereas participant talk and gesture were coded independently according to an analytic framework (Bolger et al., 2012). This coding was completed for 715 items (across participants) (Table 6). There were 219 items that were scored at the level of *related direction* using the exemplar; during the interview 136 (62%) of those items were coded as *related direction*. There were 199 items that were scored at the level of *rotation* using the exemplar; during the interview 136 (74%) of those items were coded as *rotation*. There were 114 items that were scored at the level of *lever arms* using the exemplar; during the interview 70 (61%) of those items were coded at the level of *lever arms*. There were 109 items that were scored at the level of *constraint via the fixed pivot* using the exemplar; during the interview 49 (45%) of those items were coded at the level of *constraint via the fixed pivot*. Finally, there were 74 items that were scored at the level of *tracing* using the exemplar; during the interview 35 (47%) of

Table 5  
MNSQ fit statistic for each item.

Item	MNSQ statistic
Hands Fixed Pivot-Opposite	1.34 <sup>a</sup>
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 <sup>b</sup>
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-B1	0.78
Sequential Tracing-B1'	1.37 <sup>a</sup>
Sequential Tracing-B2	1.10
Sequential Tracing-D1	1.07
Sequential Tracing-D1'	1.66 <sup>b</sup>
Sequential Tracing-E1	1.21
Sequential Tracing-E2	1.03
Sequential Tracing-CMT	1.15

<sup>a</sup>Not considered a good fit.

<sup>b</sup>Marginally out of good fit range.

those items were coded as *tracing*. The proportion of coded mechanistic elements was different from the proportion expected by chance ( $p < 0.0001$ ; Chi-squared test). This indicates a correspondence between responses to items and cognitive interviews.

#### Item-step Wright Map

The item-step Wright Map is presented in Figure 11. Results from the item-step Wright Map provide estimates for each mechanistic element, by item, with corresponding standard errors (Table 7). The standard error indicates the precision of each estimate. For example, *tracing* has an item difficulty estimate of 3.04 logits for the item Sequential Tracing-E1 (an item with a bent crank). This indicates that respondents who have a person ability estimate of 3.04 logits will have a probability of 0.5 of being scored at the level of *tracing* on this item.

This section describes machine characteristics that seem to disrupt a participant's propensity to trace. Twenty-five participants showed the propensity to trace on at least one item. However, two machine characteristics ("lever type" and "bent crank") made a difference in these participants' propensities to consistently apply *tracing*. There were eleven items in which *tracing* could be assessed. The number of items per form where this level could be assessed ranged from three to eight, with a mean of six (median = 6).

#### Lever type

Of those participants who had scored at the level of *tracing*, 0% did so on items with machines with class 3 levers; whereas, 80% had scored at the level of *tracing* for

**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)

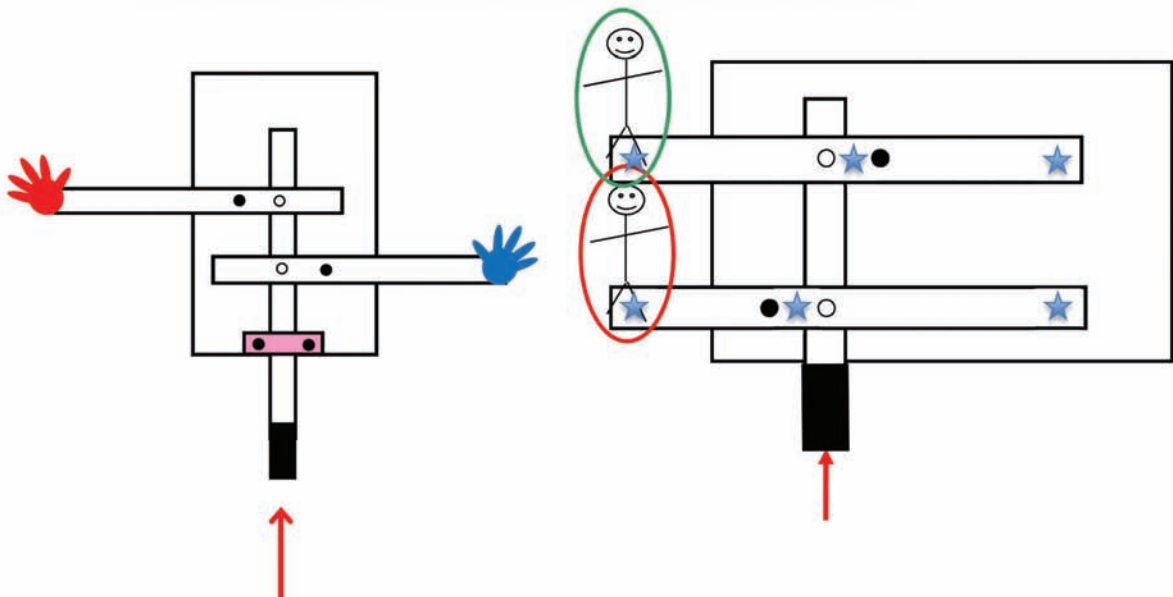
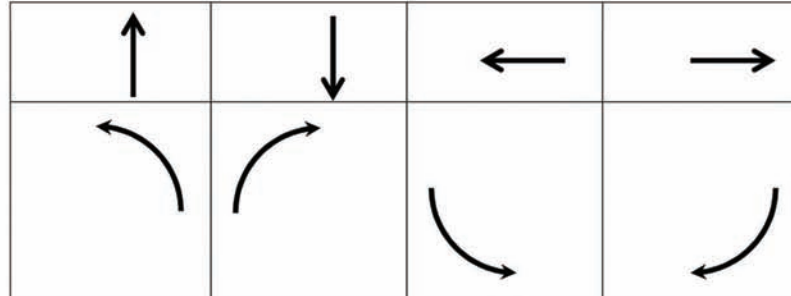


Figure 8. Hands Fixed Pivot-Opposite (left) and Sequential Tracing-B1' (right) are two items that that are slightly out of the MNSQ "good fit" range.

items with class 1 levers (Table 8). There is a difference in the proportion of participants able to apply *tracing* on machines with class 3 levers and those able to apply *tracing* on machines with class 1 levers ( $p = 0.0005$ , sign test).

**Bent cranks**

Of those participants who had scored at the level of *tracing*, 26% did so on items with machines with bent cranks; whereas, 71% had scored at the level of *tracing* for items with machines without bent cranks (Table 8). There is a difference in the proportion of participants able to apply *tracing* on machines with bent cranks and those able to apply *tracing* on machines without bent cranks ( $p = 0.01$ , sign test).

**Rank Ordering the Mechanistic Elements**

The item difficulties for each mechanistic element, by item, are presented in Table 7. The item difficulty means

for all of the mechanistic elements have been rank ordered as follows (from the easiest to most difficult): (1) *lever arms*, (2) *related 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 *t*-test), as well as between *constraint via the fixed pivot* and *tracing* ( $M = 1.80$ ;  $p < 0.0001$ , one-tailed *t*-test). Thus, there is no difference between the three easiest levels and this confirms to the hypothesized construct level (Table 1) ordering.

**Discussion**

Reasoning about mechanism is central to disciplined inquiry within STEM fields. The AMRP has characterized this form of reasoning about simple systems of levers. In addition, it has helped to explain why this form of

**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 handle. (Draw an **arrow** starting at each hand and show how they will move)

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)

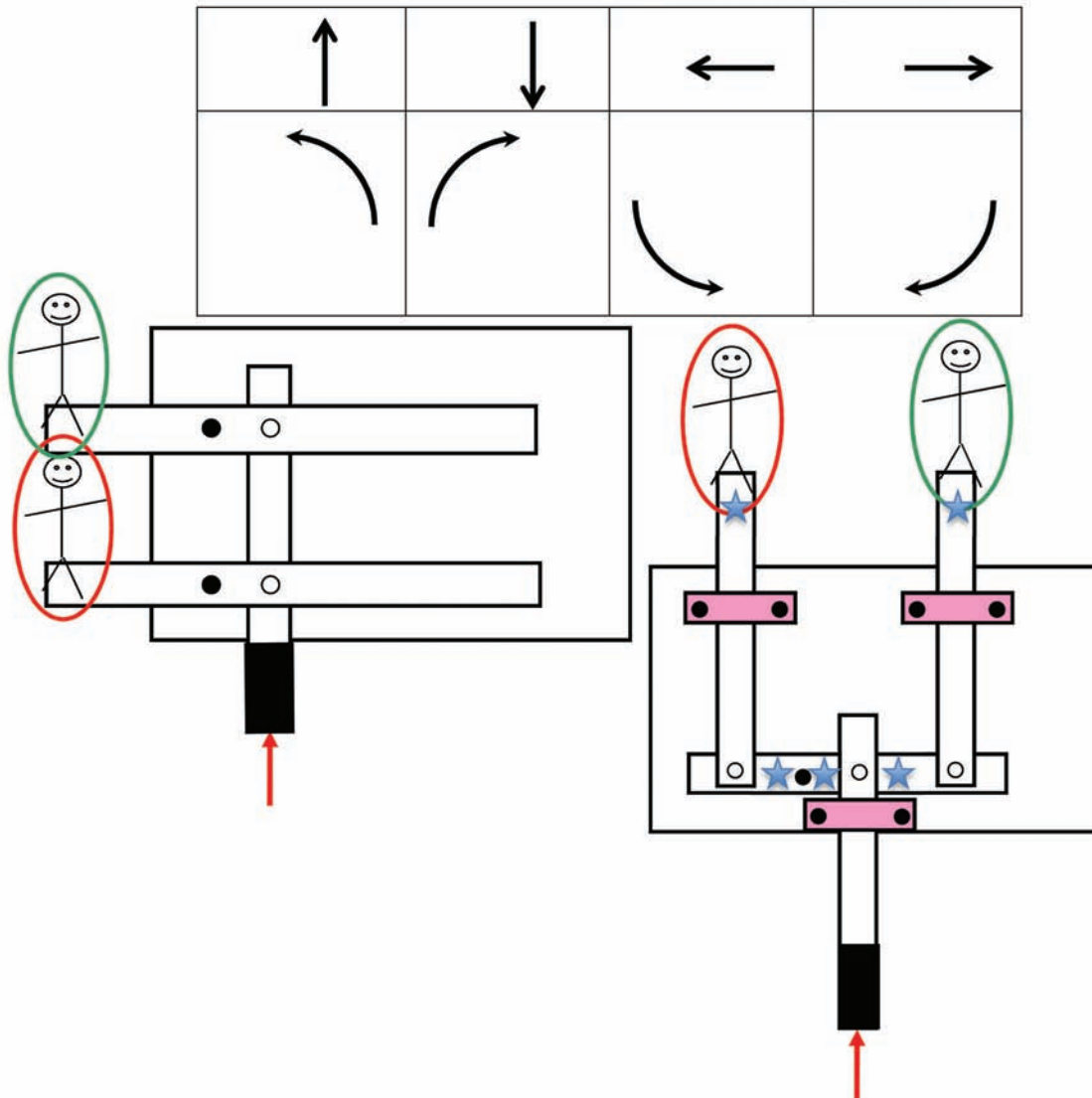


Figure 9. Machine Prediction-B2' (left) and Sequential Tracing-D1' (right) are two items that that are considered to be out of the MNSQ "good fit" range.

reasoning is difficult and what accounts for this difficulty. For example, this study shows that machine characteristics (e.g., lever type, bent crank) affect the difficulty of mechanistic reasoning.

*Assessment Validation*

Item analyses were conducted to characterize item difficulty, item fit, and item standard error. All items fell within appropriate parameters. This assessment shows high reliability by analyzing separation reliability and the SEM.

The separation reliability was equal to 0.94, indicating that the model variance accounts for most total variance. The SEM shows that on the AMRP 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 relationship between person ability estimate and SEM indicates high reliability. Validity measures were taken through use of the clinical interview and the item-step Wright Map. Participants were likely to be scored similarly on AMRP items and

Table 6

Relationship between how mechanistic elements were scored on exemplars and in cognitive interviews.

	Related direction*	Rotation*	Lever arms*	Constraint via the fixed pivot*	Tracing*	Total
Exemplars	219	199	114	109	74	715
Analytic framework (%)	62 ( $n = 136$ )	74 ( $n = 147$ )	61 ( $n = 70$ )	45 ( $n = 49$ )	47 ( $n = 35$ )	100
Mechanistic elements in sample (%)	31	28	16	15	10	100

Note. Chi-squared goodness-of-fit ( $*p < 0.0001$ , non-directional). The analytic framework used to code the cognitive interviews is presented in Bolger et al. (2012).

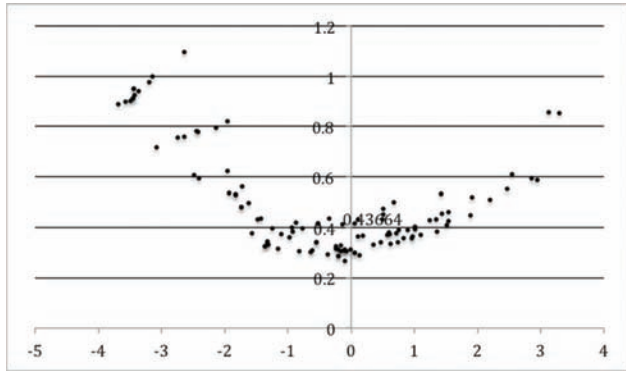


Figure 10. Scatter plot: person ability estimates versus standard error of measurement (SEM).

in the clinical interviews. This is an indication of construct validity and shows that the assessment is measuring participants' capacity to reason about the targeted mechanistic elements. Finally, the item-step Wright Map validated construct map levels (Table 1).

### Study Limitations

This study's diverse and necessarily small sample was identified to ensure that the items were correctly assessing construct levels; thus, the need to conduct clinical interviews during the AMRP's administration limited the sample size. Now that the AMRP has been calibrated, it will be important to administer it to a larger sample. Such an administration would provide more evidence about participant means to support mechanistic reasoning as well as those aspects of systems of levers that make mechanistic reasoning more or less difficult. This assessment administration provided substantial information with which to revise these items. For example, the MNSQ statistic indicated that Hands Fixed Pivot-Opposite, Sequential Tracing-B1', Machine Prediction-B2', and Sequential Tracing-D1' are slightly out of the good fit range (Figures 8 and 9). Thus, these items will be reconceptualized and redesigned. In addition, another administration to a larger sample may more clearly differentiate the difficulty of all of the mechanistic elements. Moreover, in the next iteration of design, the AMRP should only be administered to elementary and middle school students to

avoid potential ceiling effects with some high school and college students.

### Additional Forms of Reasoning to be Assessed

As the assessment is further developed, it may be expanded to assess additional learning targets. In a previous study of children's naïve mechanistic reasoning, Bolger and colleagues (2012) noted that children rarely paid attention to how far output levers moved for given inputs, 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 lever 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 relationship between input and output distance. This relationship blends mechanistic and quantitative reasoning.

### 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. The observations that are made of these mechanisms determine whether and how they may be causally coordinated. For example, how individuals inspect systems makes the difference between whether they see endpoints of motion (e.g., *related direction*) or complete rotary paths (e.g., *rotation*). In addition, the capacity to infer less visible mechanisms (e.g., *constraint via the fixed pivot*) based on other visible mechanisms and an understanding of the system seems critical to tracing.

### Perspectives for STEM Education

Because mechanistic reasoning depends on the development of domain-specific content and processes (Weinberg, 2017b), it is important that these are taught and learned across K-12 STEM education. The National





Table 7  
Item-step Wright Map: item thresholds.

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

Note. Machine prediction and hands items can only assess *related direction* and *rotation*.

\*\* $p < 0.01$ ,  
\* $p < 0.1$ .

Table 8  
Tracing by machine characteristics.

Machine characteristics	Scored at the level of tracing (%)
Lever type	
Class 3 lever(s)	0
Class 1 lever(s)	80**
Bent crank	
With bent crank	26
Without bent crank	71*

Sign test: \*\* $p < 0.001$ ,

of attachment between each two levers. This contrasts systems, like gears that have multiple points of attachments, and circuits, whose parts and mechanisms are not at all visible.

*Assessment Use*

Although this assessment is undergoing further iterations of design, the AMRP has potential uses for researchers, teacher educators, and teachers. Although mechanistic reasoning is domain-specific (Weinberg, 2017b), the AMRP can assess and support the development of prerequisite knowledge for and reasoning about diverse and complex systems. This assessment is intended to be used with students at the elementary level. This assessment has principally been used in studies focused on mechanistic reasoning about simple systems of levers (Weinberg, 2012, 2014, 2017a, 2017b;

Weinberg & Sorensen, 2018). Thus, this assessment would also have been useful in other research focused on children’s mechanistic reasoning about systems of linkages (Bolger et al., 2012; Bolger, Kobiela, Weinberg, & Lehrer, 2009; Bolger, Kobiela, Weinberg, & Lehrer, 2010; Bolger, Weinberg, Kobiela, Rouse, & Lehrer, 2011; Kobiela, Bolger, Weinberg, & Rouse, 2011).

In addition, the AMRP can assess prerequisite reasoning about many more simple and complex systems. For instance, scissors, bicycles, and eggbeaters are systems that rely on understandings of many of the same mechanisms present in levers. Two levers and a screw are the constituent parts of a pair of scissors; bicycles and eggbeaters are compound machines that include gears. Accordingly, this assessment would also have been useful in the characterization of mechanistic reasoning about similarly inspectable systems (e.g., systems of gears) (Lehrer & Schauble, 1998; Metz, 1985,1991).

Bryk, Gomez, Grunow, and LeMahieu (2015) draw a distinction between assessment tools that are used for improvement and those that are used for evaluation. In the above studies, the AMRP was used as a summative assessment. Participant mechanistic reasoning was assessed to make characterizations about how individuals and groups reasoned. In addition, the AMRP could further summatively assess how reasoning about linkages can aid children and adults in understanding mechanisms within many simple and compound systems. Rouse (2014) described

how he supported engineering practices through the design of paper pop-up books with 7th graders. The physical system in Rouse's work is similar to, but less inspectable than, the simple systems of levers in the AMRP. However, the AMRP can assess and support the prerequisite reasoning for these more complex systems, which principally rely on lever motion. In addition, English, Hudson, and Dawes (2013) supported students to reason about the design of trebuchet catapults; these catapults similarly rely on lever motion. Thus, the AMRP could also have assessed and supported prerequisite knowledge for reasoning about this system.

Instead of being used for evaluation, this assessment may also be used for improvement (Bryk et al., 2015). For example, teachers who are interested in supporting their students to reason mechanistically may use this assessment formatively. These teachers may choose to use the AMRP throughout the academic year to assess individual and group change in mechanistic reasoning over time. Moreover, science and engineering educators working in teacher education may utilize the AMRP to diagnose how effectively their interns can support mechanistic reasoning with their K–12 students in their field placements and classrooms. The purpose of measurement for improvement is to “inform efforts to change” (p. 8). Such work must include a theory of how pre- and in-service teachers' work with mechanistic reasoning changes and develops with time and experience. Weinberg and Sorensen (2018) have laid out a trajectory of development of mechanistic reasoning with third-grade students. The AMRP can more effectively be used for improvement when such a trajectory of development has been elaborated and articulated across K–12 STEM education.

## References

- Abrams, E., & Southerland, S. (2001). The how's and why's of biological change: How learners neglect physical mechanisms in their search for meaning. *International Journal of Science Education*, 23(12), 1271–1281.
- American Association for the Advancement of Science. (2011). *Vision and change: A call to action*. Washington, DC: Department of Education.
- Bolger, M., Kobiela, M., Weinberg, P., Lehrer, R. (2009, June). *Analysis of children's mechanistic reasoning about linkages and levers in the context of engineering design*. Paper presented at the meeting of the 2009 American Society for Engineering Education Annual Conference, Austin, TX.
- Bolger, M., Kobiela, M., Weinberg, P., Lehrer, R. (2010, July). *Embodied experiences within an engineering curriculum*. Paper presented at the meeting of the 2010 International Conference of the Learning Sciences, Chicago, IL.
- Bolger, M., Kobiela, M., Weinberg, P. J., Lehrer, R. (2012). Children's mechanistic reasoning. *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 meeting of the 2011 National Association for Research in Science Teaching Annual International Conference, Orlando, FL.
- Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, 15(4), rm4.
- Borton, R. W. (1979, March). *The perception of causality in infants*. Paper presented at the meeting of the Society for Research in Child Development, San Francisco, CA.
- Brophy, S., Klein, S., Portsmore, M., & Rogers, C. (2008). Advancing engineering education in P–12 classrooms. *Journal of Engineering Education*, 97(3), 369–387.
- Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.
- Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York, NY: Academic Press.
- Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in “sophisticated” subjects: Misconceptions about trajectories of objects. *Cognition*, 9(2), 117–123.
- Chin, C., & Brown, D. E. (2000). Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching*, 37(2), 109–138.
- Coyle, E. J., Jamieson, L. H., & Oakes, W. C. (2005). EPICS: Engineering projects in community service. *International Journal of Engineering Education*, 21(1), 139–150.
- Cunningham, C. M. (2009). Engineering is elementary. *The Bridge*, 30(3), 11–17.
- diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10(2–3), 105–225.
- English, L. D., Hudson, P. B., & Dawes, L. A. (2013). Engineering-based problem solving in the middle school: Design and construction with simple machines. *Journal of Pre-College Engineering Education Research*, 3(2), 1–113.
- Gearhart, M., & Saxe, G. B. (2004). When teachers know what students know: Integrating mathematics assessment. *Theory into Practice*, 43(4), 304–313.
- Ginsburg, H., 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.
- Gopnik, A., Sobel, M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37(5), 620–629.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141–158.
- Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors and functions. *Cognitive Science*, 28(1), 127–138.
- Hynes, M., Portsmore, M., Dare, E., Milto, E., Rogers, C., Hammer, D., & Carberry, A. (2011). *Infusing engineering design into high school STEM courses*. National Center for Engineering and Technology Education.
- Kobiela, M., Bolger, M., Weinberg, P. J., Rouse, R. (2011, June). *Mathematization and embodiment for reasoning about mechanism within an engineering curriculum*. Paper presented at the 41st Annual Meeting of the Jean Piaget Society, Berkeley, CA.
- Lachapelle, C. P., & Cunningham, C. M. (2014). Engineering in elementary schools. In S. Purzer, J. Strobel, M. Cardella (Eds.), *Engineering in pre-college settings: Synthesizing research, policy, and practices* (pp. 61–88) Lafayette, IN: Purdue University Press.
- Lehrer, R., Kim, M. J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In J. Confrey A. Malony (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. 31–60). Charlotte, NC: Information Age Publishers.
- 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., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3), 265–288.
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1–25.
- Marra, R. M., & Bogue, B. (2006). Women engineering students' self efficacy—A longitudinal multi-institution study. Women in Engineering ProActive Network.
- Marshall, J. A., & Berland, L. K. (2012). Developing a vision of pre-college engineering education. *Journal of Pre-College Engineering Education Research*, 2(2), 36–50.
- 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.
- 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.
- Moore, T. J., Glancy, A. W., Tank, K. M., Kersten, J. A., Smith, K. A., & Stohlmann, M. S. (2014). A framework for quality K–12 engineering education: Research and development. *Journal of Pre-College Engineering Education Research (J-PEER)*, 4(1), 2.
- Moore, T. J., Tank, K. M., Glancy, A. W., & Kersten, J. A. (2015). NGSS and the landscape of engineering in K–12 state science standards. *Journal of Research in Science Teaching*, 52(3), 296–318.
- National Research Council. (2009). *Engineering in K–12 education: Understanding the status and improving the prospects*. Washington, DC: National Academies Press.
- National Research Council. (2010). *Standards for K–12 engineering education*. Washington, DC: National Academies Press.
- National Research Council. (2011). *A framework for K–12 science education: Practices, crosscutting concepts and core ideas*. Washington, DC: National Academies Press.
- Nazzi, T., & Gopnik, A. (2000). A shift in children's use of perceptual and causal cues to categorization. *Developmental Science*, 3(4), 389–396.
- NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.
- Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the “two disciplines” problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, 24(1), 307–353.
- Piaget, J. (1951). *The child's conception of the world*. Lanham, MD: Rowman & Littlefield.
- Purzer, S., Douglas, K. A., Folkerts, J. A., & Williams, T. V. (2017). *An assessment framework for first-year introduction to engineering courses*.
- Roehrig, G. H., Moore, T. J., Wang, H. H., & Park, M. S. (2012). Is adding the E enough? Investigating the impact of K–12 engineering standards on the implementation of STEM integration. *School Science and Mathematics*, 112(1), 31–44.
- Rouse, R. J. (2014). *Investigating how K–12 students engage in engineering practices* (Doctoral dissertation). Vanderbilt University, Nashville, TN.
- Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, 93(2), 499–525.
- Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, 49(1), 31–57.
- Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, 32(1), 102.
- Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, 3(2), 115–163.
- Talanquer, V. (2010). Exploring dominant types of explanations built by general chemistry students. *International Journal of Science Education*, 32(18), 2393–2412.
- Weinberg, P. J. (2012). *Assessing mechanistic reasoning: Promoting system thinking* (Doctoral dissertation). Retrieved from Electronic Theses and Dissertations. (etd-07112012-004845)
- Weinberg, P. J. (2014, April). Mathematical description and mechanistic reasoning in STEM education: Learning through mathematizing levered systems. Paper presented at the meeting of the American Education Research Association, Philadelphia, PA.
- Weinberg, P. J. (2017a). Mathematical description and mechanistic reasoning: A pathway towards STEM integration. *Journal of Pre-College Engineering Education Research (J-PEER)*, 7(1), 90–107.
- Weinberg, P. J. (2017b). Supporting mechanistic reasoning in domain-specific contexts. *Journal of Pre-College Engineering Education Research (J-PEER)*, 7(2), 27–39.
- Weinberg, P. J., Sorensen, E. K. (April, 2018). Embodiment, mathematization, and mechanistic reasoning: Characterizing an after-school engineering program. Paper presented at the meeting of the American Education Research Association Conference, Toronto, ON, Canada.
- Wilson, M. (2005). *Constructing measures: An item response theory approach*. Mahwah, NJ: Lawrence Erlbaum.
- Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. *Objective Measurement: Theory into Practice*, 5, 113–133.
- Wright, B. D., Linacre, J. M., & Gustafson, J. E. (2009). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370.



Appendix A

**Item**

**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)

↑	↓	←	→
↶	↷	↵	↷

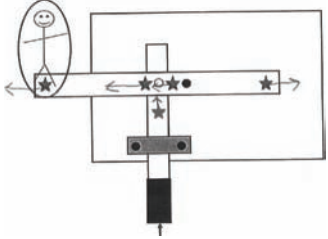
  

Figure A. Item Sequential Tracing-A1 (STA1).

**Exemplar.** Exemplar for Sequential Tracing A1 (STA1).

Level	Mechanistic element	Mechanistic element descriptions	Mechanistic element example
5	Tracing	Participant diagnoses all mechanistic elements (without gaps) from input to output.	
4	Constraint via the fixed pivot	Participant draws the opposite 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. However, the location of these paths must reasonably approximate fractions of circles centered around either the fixed or floating pivot. <i>Note: Although these paths are centered around the fixed pivot, this element of mechanistic reasoning does not make this distinction.</i>	
1	Related direction	Participant draws the coordinated input/output motion.	

**Exemplar.** Exemplar for Sequential Tracing A1 (STA1). (continued)

Level	Mechanistic element	Mechanistic element descriptions	Mechanistic element example
0	Student diagnoses no mechanistic elements	No mechanistic elements are shown.	
NL	No link	It is not clear if the participant understood the nature of the task.	"I don't know"
M		Missing response	

*Note.* This item assesses students' ability to diagnose the mechanistic elements of *related direction*, *rotation*, *lever arms*, and *constraint via the fixed pivot* as well as *tracing*. No link (NL) indicates an item response that does not provide any evidence of mechanistic reasoning (i.e., diagnosis of no mechanistic elements). "Missing" indicates that the item was left completely blank. The "stars" have been placed on the levers to allow participants to indicate lever motion. A "little person" has been included on the output lever to make the system output salient.