Integrating heuristic and simulation-based reasoning in intuitive physics

Kevin A Smith

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Peter W Battaglia

DeepMind

Joshua B Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Abstract

The ability to predict, reason about, and act in the physical world is crucial for human survival, but the cognitive systems that underlie these capabilities have been the subject of intense debate. Some theories posit that reasoning about physical events is based on dynamic mental models that approximately simulate underlying physical mechanisms (e.g., the forces incident on objects that cause them to move); others argue for simpler heuristics that predict key physical outcomes (e.g.,“objects fall straight down when dropped”). We argue that general physical reasoning requires both simulation and rules, and propose a modeling framework for understanding the interactions and trade-offs between these cognitive systems as *resource-rational* computations to efficiently solve problems. We study these trade-offs using predictions about stability: judging whether a balance beam will fall, and if so how. While prior research suggests that people often use rules when solving balance beam tasks, these tasks are similar to others that have been found to rely on mental simulation. Across five experiments, participants’ predictions cannot be explained with simulation or rules alone, but we find evidence that individuals rely on both capacities. The mixture of strategies that people use to solve these stability problems is consistent with a resource-rational trade-off that accounts for the costs and benefits of using those strategies. Finally, we find that participants can rationally adapt this mixture of strategies to perform more efficiently given the distribution of task instances they encounter, demonstrating the flexible and online nature of the computational trade-offs in intuitive physics.

Keywords: physical reasoning; resource rationality; simulation; rules and heuristics

Integrating heuristic and simulation-based reasoning in intuitive physics

1 Introduction

Humans have remarkable capabilities for understanding, making predictions about, and acting on the physical world. We easily toss and catch balls, stack up plates or blocks, and pour water from a pitcher into a glass – all tasks that seemingly require an understanding of what will happen next, or how our actions will affect the world. This capability for predicting the outcome of physical events is even thought to underlie the quintessentially human tool use capabilities that give rise to our complex culture (Allen, Smith, & Tenenbaum, 2020; Osiurak & Reynaud, 2020). Yet despite the importance of this capability, there remains active debate about how the mind accomplishes physical prediction, with theories often falling into two broad camps: either suggesting that we use simulatable mental models of the world (Battaglia, Hamrick, & Tenenbaum, 2013; Smith & Vul, 2013; Ullman, Spelke, Battaglia, & Tenenbaum, 2017; Hegarty & Just, 1993), or that we rely on a system of rules and heuristics (Siegler, 1976; Davis, Marcus, & Frazier-Logue, 2017; Vasta & Liben, 1996).

The debate about whether physical reasoning is predicated on simulation or rules has a long history, with both sides claiming evidence that supports their theory and proposing models of these theories that explain human behavior. However, there has been a growing appreciation that physical reasoning might not be a monolithic entity, but instead could be accomplished by a set of different cognitive systems (Hegarty, 2004; Smith, Battaglia, & Vul, 2018; Zago & Lacquaniti, 2005). But if people use multiple systems for physical reasoning, then they must decide which system to use in any given scenario. While some research has proposed "rules of thumb" for choosing between simulation and rules – for instance, that we are more likely to use simulation with more realistic stimuli (Schwartz, 1995) – these are not universally true, nor do they explain *why* one system should be favored over another, and cannot provide precise predictions about when and how cognitive systems should trade off with one another.

In this paper, we propose a computational framework for understanding the trade-off between different cognitive strategies for physical reasoning, called the *Integration of Simulation*

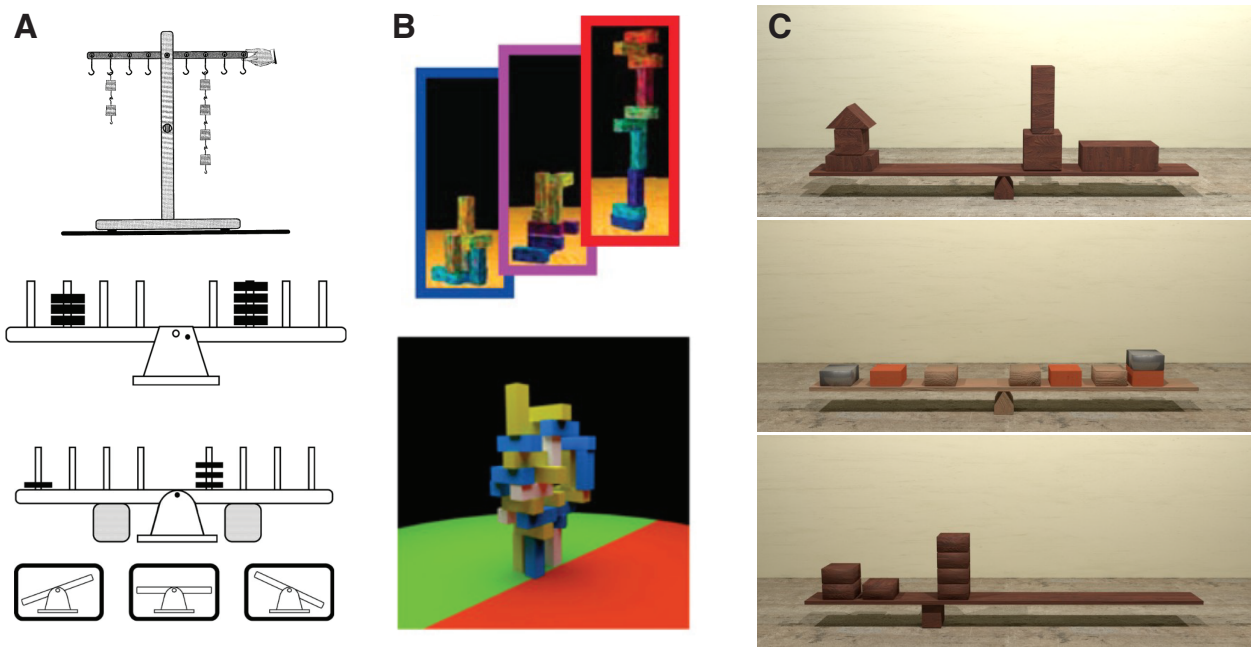

Figure 1. Stimuli to test judgments of "which direction will it fall?" *A:* Traditional balance beam stimuli that are used to argue for rule-based physical reasoning are typically presented as simple diagrams (from top to bottom: Boom, Hoijtink, & Kunnen, 2001, Jansen, Raijmakers, & Visser, 2007, van der Maas & Jansen, 2003). *B:* Block-tower stimuli that are used to argue for simulation-based physical reasoning typically use complex configurations where simple rules are difficult to apply (top: Battaglia et al., 2013, bottom: Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016). *C:* Our stimuli are designed to recreate problems that test stability judgments, but introduce more realistic visuals and physical variations, requiring incorporating estimating mass from shape or material, or incorporating the weight of the balance beam.

and Rules (ISR) framework. We formulate this as a resource-rational trade-off between systems, where the mind's goal is to find a 'good enough' solution as efficiently as possible (Gershman, Horvitz, & Tenenbaum, 2015; Gigerenzer & Goldstein, 1996; Griffiths, Lieder, & Goodman, 2015). Thus people will use simple rules, even biased ones, if the expected loss in accuracy is offset by an efficiency gain from forgoing more cognitively costly simulation. We test this framework in the domain of stability judgments, which have a long history of competing theories to explain human behavior (Marcus & Davis, 2013): there are many theories and models that suggest that judging whether and how objects will cause a balance beam to tip over relies on rules or decision trees

(Figure 1A; Siegler, 1976; Rijn, Someren, & Maas, 2003), and another set that suggest mental simulation is the basis of judgments of whether a tower of blocks will fall (Figure 1B; Battaglia et al., 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Zhou, Smith, Tenenbaum, & Gerstenberg, 2022). Using a judgment task that combines features of experiments from both theoretical camps (Figure 1C), we demonstrate the human stability judgments do in fact use a combination of simulation and rules, and provide a model instantiating the ISR framework that quantitatively explains people's judgments across a variety of balance beam scenarios.

The rest of this paper is structured as follows. In Section 2 we review the theoretical foundations of simulation-based and rule-based physical reasoning, as well as proposing the ISR framework for combining the two as a resource-rational trade-off. In Section 3 we review the literature on how people judge stability in the case of balance beams, and show that, although there is evidence that in many cases people use rules to form their predictions, there remain cases that are difficult to explain through rules alone. We then introduce a set of experiments in Section 4 that test peoples' predictions about balance beams, and demonstrate that rules alone are not sufficient to explain the pattern of responses. Next, in Section 5 we describe how the Integration of Simulation and Rules model can be applied to stability judgments, and demonstrate that it explains the empirical behavior well. In Section 6 we demonstrate that this model generalizes well to explain human predictions about more complex balance beam configurations, and in Section 7 show that the Integration of Simulation and Rules model naturally explains a finding from the balance beam literature that has been challenging to explain with rules alone (Ferretti & Butterfield, 1986). In Section 8 we show that, as expected under resource-rational trade-offs, changing the distribution of balance beams that people expect to see will also change the mixture of strategies they use for reasoning about stability. Finally, in the discussion we review implications of this framework for understanding physical reasoning and cognitive strategy selection more broadly.

2 Systems for physical reasoning

2.1 Physical reasoning using simulation

The ability to predict how the world will unfold is a crucial feature of cognition. Since Craik (1943) suggested that we have a "small-scale model of external reality" that we can use to run forward and understand the impact our actions will have on the world, many people have suggested that this capability underlies spatial reasoning (Kosslyn & Ball, 1978; Shepard & Metzler, 1971), language comprehension (Bergen, 2012), and social reasoning (Gallese & Goldman, 1998). More recently, Battaglia et al. (2013) extended this theory to physical inference by suggesting we all have a "game engine" in our minds that we can use to simulate how physical events will unfold over time.

There are two key features of this simulation engine that we focus on here. First, it runs simulations on mental models of the world using approximately accurate physical principles. Second, our mental representations of the world are noisy due to uncertainty about the world, and therefore simulation necessarily provides us with probabilistic representations over future world states.

The proposal that we 'simulate' physics makes a specific claim about the nature of this process: there is a direct mapping between the simulation process and the process that occurs in the world (Fisher, 2006; Moulton & Kosslyn, 2009). This does not mean that we perfectly represent physical laws such as Newton's laws of motion, but rather that our simulation runs forwards our mental models in a stepwise way that is similar to how the world itself might evolve. Thus the cognitive process that performs this simulation might be an approximation of the laws of physics (Battaglia et al., 2013; Ullman et al., 2017), just as computer physics engines do not explicitly perform calculations derived from Newtonian mechanics but instead use collision rules that approximate these laws of motion (Millington, 2007).

Nonetheless, just as there are a limited set of principles of Newtonian mechanics that can be applied across a wide range of objects and situations, our mental simulations use a limited set of principles to form generalizable predictions. In this way our physical predictions can naturally

extend to objects and situations we have never observed before: because we know about fluid dynamics and how it is affected by viscosity, we can not only reason about how liquids that we are familiar with will pour (Bates, Yildirim, Tenenbaum, & Battaglia, 2019), but also rapidly infer the viscosity of an unknown fluid from a single observation and use that information to predict how that fluid will pour in other situations (Kubricht et al., 2016). Thus a key strength of this mental model is that with just a limited set of principles that define how objects with different shapes and properties interact, it can produce predictions for an almost infinite set of scenarios that we might encounter in our day-to-day lives.

Even if physical simulation is based on relatively accurate principles, its predictions will not always accurately track the future. The mental model of the world that is simulated is not perfectly accurate, since perception is noisy and cannot perfectly localize objects. Furthermore, there are latent physical properties such as density, stiffness, or elasticity that we can only noisily infer from object materials (Fleming, 2014; Paulun, Schmidt, Assen, & Fleming, 2017; Yildirim, Smith, Belledonne, Wu, & Tenenbaum, 2018) or how they interact with other objects (Sanborn, Mansinghka, & Griffiths, 2013; Hamrick et al., 2016; Hauf, Paulus, & Baillargeon, 2012; Neupärtl, Tatai, & Rothkopf, 2020; Yildirim et al., 2018).

Because we have uncertainty in the localization and properties of objects, predicting how those objects will act in the future will compound this uncertainty, not just leading to variability (Battaglia et al., 2013; Smith & Vul, 2013), but also systematic biases in judgments of physical outcomes. These biases may arise either because the perception of object properties are themselves biased, or because physical outcomes are non-symmetric. For instance, because perception of object velocities is biased towards slower motion (Stocker & Simoncelli, 2006), people who are asked to judge the relative masses of two object from a collision – which requires comparing both the pre- and post-collision velocities – will produce characteristic biases (Sanborn et al., 2013). And people are more likely to judge a stable tower of blocks to be unstable than to judge an unstable one to be stable, not because we are inherently biased to call things unstable, but because it is more likely that incorrectly judging the position of a block in a stable tower will

cause it to become unstable than vice versa (Zhang, Wu, Zhang, Freeman, & Tenenbaum, 2016).

This uncertainty also implies that simulation will not provide just a single answer, but rather a range of possible futures. This distribution over futures can help us by allowing us to calibrate when we should be more or less sure about our decision making (Smith & Vul, 2015), but because we sample only a limited set of outcomes (Hamrick, Smith, Griffiths, & Vul, 2015) individual predictions can be erroneous even if physical simulation provides an unbiased estimate. Thus even if physical inference relies on generally unbiased simulation in a wide set of situations, it is only under nearly noiseless conditions with simple dynamics that it will provide certainty in its predictions.

2.2 Physical reasoning using rules

In contrast to theories of physical reasoning that propose we use simulatable mental models, others have pointed out that simulation cannot explain the full range of ways that people reason about the world (Ludwin-Peery, Bramley, Davis, & Gureckis, 2020, 2021; Davis & Marcus, 2015). Instead, many claim that physical knowledge can be formalized as a set of axioms and rules that can be combined to produce logically consistent statements about the world (Hayes, 1979). This suggests that physical reasoning could be performed using first order logic (Davis et al., 2017) or decision trees (Siegler, 1976) applied to propositional statements about the scene (e.g., “Object A is surrounded by object B on all sides” or “Object A is heavier than object B”). These propositions are thought to be lifted from our perception and formed into “qualitative” representations (Forbus, 1983).

Although rules are typically defined as being specific to a certain physical situation, they allow flexibility along other dimensions. While simulations require that all relevant objects have at least roughly known (or assumed) positions, motions, and properties, rules can be used even with incomplete information: for instance, we know that objects in a grocery bag will remain in the bag to be carried home without needing to know specifics about *what* those items are (Davis et al., 2017). The more limited, qualitative representations underlying these rules can also help with abstraction, since it can be easier to find similarities in relations between objects across scenes

than to find how continuous representations relate to each other (Forbus & Gentner, 1986; Bassok & Holyoak, 1989). And finally, on computing hardware, physical simulation is significantly more computationally expensive than implementing symbolic rules, and this same cost asymmetry between simulation and rules has been proposed to exist in the mind as well (Davis & Marcus, 2015).

However, the downside of these rules is that they can be based on erroneous logic and produce biased predictions. Indeed, much of the literature on intuitive physics has focused on the reasoning errors that people make, such as misconceptions about how objects exiting from curved tubes will travel (McCloskey, Caramazza, & Green, 1980), how mass distribution affects the way a wheel will roll (Proffitt, Kaiser, & Whelan, 1990), or how balance beams will tip (Siegler, 1976). Furthermore, rules are defined to provide predictions over discrete outcomes, and can therefore be ill equipped to handle physical predictions with continuous outcomes, such as predicting exactly where a ball bouncing around a scene will come to rest (Forbus, Nielsen, & Faltings, 1990). Thus while rules have the benefit of providing discrete, abstractable predictions, these predictions can also be too coarse for certain judgments, and can produce non-physical errors and biases.

2.3 Multiple systems for physical reasoning

While prior theories have suggested that we have more than one system for conceptualizing physics, it is not clear how these systems might coexist to produce a unified physical understanding. Often, these systems have been thought to be separately activated by either the nature of the stimuli or the task used to probe physical knowledge (Hegarty, 2004; Kubricht, Holyoak, & Lu, 2017; Schwartz & Black, 1996b; Zago & Lacquaniti, 2005). For instance, Kozhevnikov and Hegarty (2001) argue that simulation is used for automatic, immediate judgments, but that these judgments can be later overridden with more explicit, logical reasoning. Schwartz (1995) suggests that simulation is used to reason about objects that look realistic, while analytic strategies are used for diagramatic stimuli. Under this theory, more natural stimuli enable simulation by leading people to represent a picture as the object itself, or by allowing for

observation of objects' velocities and other kinematic properties that can be used to extrapolate motion (Kaiser, Proffitt, Whelan, & Hecht, 1992; DeLucia & Liddell, 1998). But it is also possible that these stimuli do not *require* simulation, but rather *facilitate* it: it is easier to form mental models of the world with natural or kinematic information. Indeed, when presented with a diagram of a pendulum and asked an explicit question about its motion, people will naturally provide biased, rule-like answers; if, on the other hand, they absorb the cost of setting up the mental model by imagining the pendulum in motion first, their responses are more likely to follow accurate physical relationships (Frick, Huber, Reips, & Krist, 2005). Similarly, judgments about how water will pour from cups are typically subject to a number of biases (Vasta & Liben, 1996), but if people are asked to first imagine tilting a cup of water, their judgments become close to veridical (Schwartz & Black, 1999).

This suggests another explanation for why we see such separation of cognitive processes across tasks: some scenarios or queries might be considerably better suited for one cognitive system over another, and so that system is preferentially chosen for solving those problems even if it is in theory possible to use a different system. For instance, animation might make simulation less costly (Kaiser et al., 1992), familiar problems can encourage retrieval of prior instances from memory (Kaiser, Jonides, & Alexander, 1986), and simulations that produce time-varying predictions about objects' locations are more likely to be used in visuo-motor tasks that require precise localization (Smith et al., 2018).

2.4 Adjudicating between systems for physical reasoning

In order to understand how people might decide which cognitive system to apply in any instance of physical reasoning, we turn to the framework of *resource rational* strategy selection. This framework asks how a rational reasoner might select strategies – perhaps consciously, perhaps implicitly – to solve their problems under resource constraints (e.g., time pressure or cognitive limitations; Gigerenzer & Goldstein, 1996; Lieder & Griffiths, 2020; Simon, 1955). Under this framework for strategy selection, an agent's goal is not always to choose the strategy that is

most likely to solve its problem, but to choose the strategy that is most efficient, given the costs and utilities inherent in the problem structure. This efficiency is defined as the *Value of Computation* (VoC): the utility (U) from applying the strategy minus the costs of performing the strategy (e.g., metabolic costs of processing, or opportunity costs of thinking rather than doing something else; C). However, because strategies can be non-deterministic, they may produce different outcomes (and hence different utilities) or require more or less effort even applied to the same problem; therefore the VoC is defined as the expected utility over all possible ways that a strategy might resolve for a problem. Thus, for any given strategy (S) and problem (P_i), the VoC can be calculated as (Lieder & Griffiths, 2017; Russell & Wefald, 1991):

$$VoC(S,P_i) = E[U(S,P_i) - C(S,P_i)] \tag{1}$$

The goal is therefore to produce a strategy that can be applied to a set of problems ($\bar{P}$) that maximizes this expected Value of Computation:

$$S^* = \arg\max_S \sum_{p_i \in \bar{P}} VoC(S,P_i) \tag{2}$$

Here we define a strategy as a set of actions or cognitive operations ($S = \{s_1, s_2, ..., s_n\}$) that are assembled sequentially or as a decision tree. For instance, when encountering a multiplication problem $x * y$, a strategy might be to attempt retrieval of the answer from memory (s_1) and then, if retrieval fails, to manually solve the problem by adding x together y times (s_2; Siegler, 1988). This definition allows flexibility and backups in cognitive plans, such that one rapid plan can be prioritized while other slower but more accurate plans can be used if the first fails (Siegler & McGilly, 1989). There can also be choice points within each of the primitive strategies, e.g., how much time to spend retrieving an item from memory, or how much effort to dedicate to a cognitive operation (Gershman et al., 2015).

It is important to note that analytically determining the best strategy for a set of problems is computationally intractable: for any given situation knowing for sure how well a strategy will perform or how much effort it will take requires actually using that strategy. Thus analytically

determining the optimal strategy requires implementing every strategy, which defeats the point of reasoning about which strategy to use. Instead, there have been a number of heuristics suggested that can learn approximate solutions to this selection problem, including reinforcement learning-like updates where strategies that are successful become more preferred (Erev & Barron, 2005; Siegler, 1999) or learning approximations to the VoC function (Lieder & Griffiths, 2017). These learning heuristics suggest that strategy selection is not necessarily a conscious choice driven by explicit comparisons between strategies, but rather is typically an implicit process that operates automatically (Siegler, 1988). Regardless of the heuristic used, any good strategy selection process should allow people to approach problems flexibly based on the relative costs and benefits of each strategy within a given environment. Thus it may be globally inefficient to use a strategy that is more likely to give the correct answer if a more erroneous but less costly strategy provides more value; but conversely, if there is a large enough difference in accuracy or small enough difference in cognitive costs, the more accurate strategy should be preferred.

This framework thus provides a window into understanding how different cognitive systems might trade off in physical reasoning: when simulation provides better information or is less costly to set up, it should be more likely to be used; conversely, when simple rules will suffice to solve our problems, we should rely on those instead.

2.5 The Integration of Simulation and Rules framework

We propose the Integration of Simulation and Rules (ISR) framework to instantiate the way people might select between different systems for physical reasoning. This framework (pictured in Fig. 2) assumes that to reason about a physical problem, people must hold or create a set of possible ways to solve that problem. People then select a strategy from this set (perhaps implicitly) that is expected to maximize the next Value of Computation, according to the principles of resource rationality. Finally, people apply the selected strategy to their mental representation of a particular scene in order to produce an answer to the problem under consideration. We note that this framework suggests an ordering for the process that the mind must perform to solve arbitrary

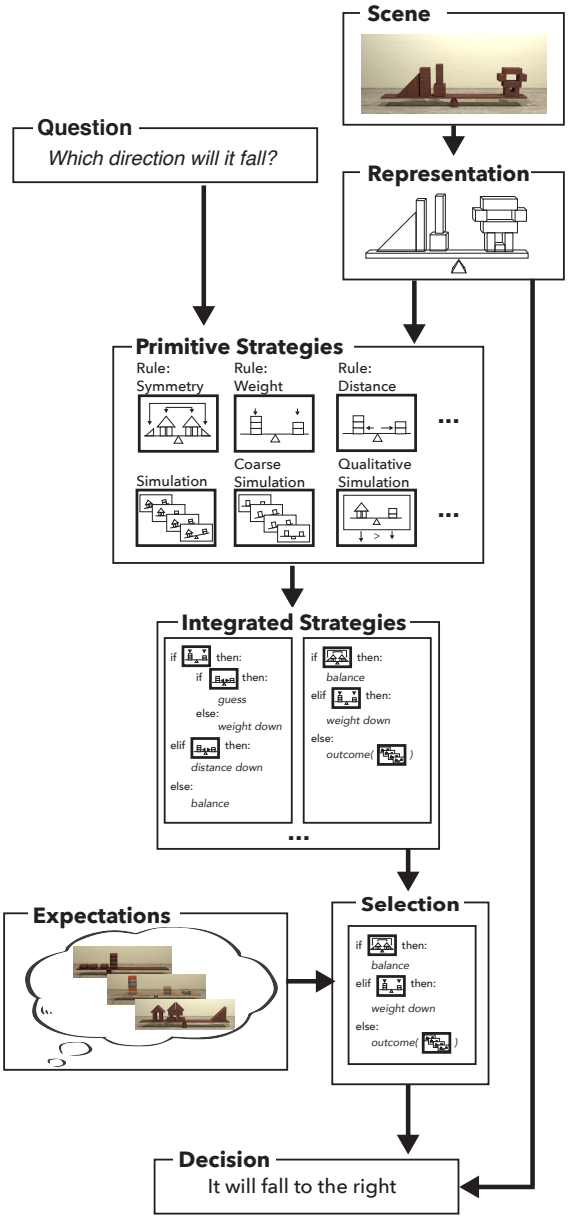

Figure 2. The structure of the Integration of Simulation and Rules (ISR) framework to capture peoples' strategy selection and use. When confronted with a problem under consideration (the *question* that applies to a *scene* that is parsed into a mental *representation*), people rely on a set of *primitive strategies* that could possibly provide an answer. These primitive strategies can be chained together into *integrated strategies* that form simple programs, using the primitive strategies as choice points or outputs. Next, one integrated strategy must be chosen out of all of the theoretically possible ones; this selection is done to approximately maximize Value of Computation of using that strategy given expectations over the problem set to be encountered. Finally the integrated strategy is applied to the scene representation to produce an answer to the question under consideration.

physical problems, but does not make strong commitments to the particular ways each of these steps are performed (e.g., whether strategies have been formed in the past and are retrieved from memory, or whether they are computed on the fly; see Discussion Section 9.1).

We propose that people have a set of "primitive strategies" – rules, heuristics, or ways of simulating the world – that could provide an answer to the question under consideration (e.g., Dehaene, 1992; Siegler, 1987). However, these primitive strategies are not guaranteed to give an answer (e.g., if a rule depends on which object has greater weight but the weights are equal, or if simulation provides a flat distribution over possible outcomes). Thus we suggest that people chain these primitive strategies together into "integrated strategies:" small programs that use the primitive strategies either as choice points (e.g., if one object has a greater weight, do A, otherwise, do B), or as outputs (e.g., if one object has greater weight, choose that object).[1] This integrated strategy can be applied to the scene under consideration, and will provide an answer to the required question.

However, even with a limited number of primitive strategies, there is a combinatorial explosion of the number of possible integrated strategies that can be constructed; thus people must decide which of these strategies to consider. Here we assume, consistent with the resource rationality theory, that people will select an integrated strategy that they expect will (approximately) maximize the Value of Computation, given an expectation about the problems that they believe they will encounter. This selection is not necessarily a conscious decision, but is a cognitive process that flexibly activates a strategy based on the expected utilities and costs. But because expectations about the range of problems is uncertain, the VoC will be approximate, and so people might pick multiple integrated strategies and select between them. Thus when people expect to encounter situations where simple rules will mostly provide accurate answers, we should observe people using integrated strategies that rely on early use of those rules, but when in a domain where those rules are not expected to be helpful, people should use integrated strategies that do not include those rules as a primitive strategy.

---

[1] For a discussion of how primitive strategies might be performed in parallel, see the Discussion, Section 9.1.2.

Finally, while people select integrated strategies for a set of problems, people must apply the selected strategy to a single problem. This requires first perceiving and forming a mental representation of the scene (which can be subject to perceptual noise; Battaglia et al., 2013; Smith & Vul, 2013), then running the selected strategy on that representation to produce an answer to the question under consideration.

In order to test the ISR framework, we require a domain where we expect to observe people using a mixture between rules and simulation. However, in many previously studied situations, it is impossible to tease apart whether selecting rules versus simulation is a choice or is fixed based on the problem: either the rules or heuristics that come to mind are significantly less helpful than simulation and are never chosen, or because the rules are *too* helpful and so are always chosen. Instead, we turn to a task that can be explained by both simulation and logical rules: judging if and how a balance beam will tip. While performance on this task has historically been described as based on a set of decision tree based rules (Siegler, 1976; Normandeau, Larivée, Roulin, & Longeot, 1989; Jansen & van der Maas, 1997), there are some suggestions that human predictions on this task are not entirely rule-based (Ferretti & Butterfield, 1986; Schapiro & McClelland, 2009). This task is also in many ways similar to stability judgments that are explained by simulation but not rules (Marcus & Davis, 2013): judging how and whether a set of blocks formed into a tower might fall (Battaglia et al., 2013; Hamrick et al., 2016; Zhou et al., 2022). In order to understand how rules and simulation might combine in stability judgment tasks, we first review historical evidence for and against the use of rules on classical balance beam judgments, then turn to experiments that are expected to require a combination of simulation and rules.

3 Predictions of stability for balance beams

In classical balance beam tasks, people (often children) are presented with a diagram of a balance beam on which a stack of blocks sits on either side of the center pivot, and are asked to judge whether the beam will stay balanced, tip to the left, or tip to the right (Figure 1A). The way that the beam actually falls depends on the net torque, which can be calculated for each side by a

multiplicative combination of the weight on the beam and the distance that weight sits from the center of the beam – whichever side has more torque is the side to which the beam falls. However, people often do not use this torque calculation and instead provide inaccurate judgments of the direction the beam will fall.

3.1 Rules for balance

Siegler (1976) proposed that balance judgments were derived from a decision tree made of binary judgments about the balance beam. He further suggested that these decision trees developed through childhood, starting from a simple focus on a single dimension of the scene, and growing to understand the combination of weight and distance. Thus younger children would display more characteristic errors because they used simpler sets of rules, while older children and adults would have more developed (though still mostly imperfect) rule sets and would therefore be more accurate. Siegler (1976) proposed four stages representing different sets of "rules" through which people would develop (see Figure 3):

- **Rule 1:** A singular focus on weight. If there is more weight on one side than the other, the beam would be predicted to fall that direction, but would balance otherwise.
- **Rule 2:** Separately considering weight and distance. This builds on Rule 1, such that if the weights are equal on both sides of the beam but the weight on one side is further from the center, the side with the greater distance should fall rather than balance. However, when in conflict, weight takes precedence over distance.
- **Rule 3:** Confusion about integration. With this rule, children are able to recognize that greater weight but less distance can compensate for lesser weight but greater distance. However, they are unsure of *how* to integrate the two dimensions, so when they come into conflict, children 'muddle through' and simply guess randomly.
- **Rule 4:** Mature integration. Children who use this rule understand how weight and distance combine in a multiplicative fashion and therefore can determine how any beam will fall

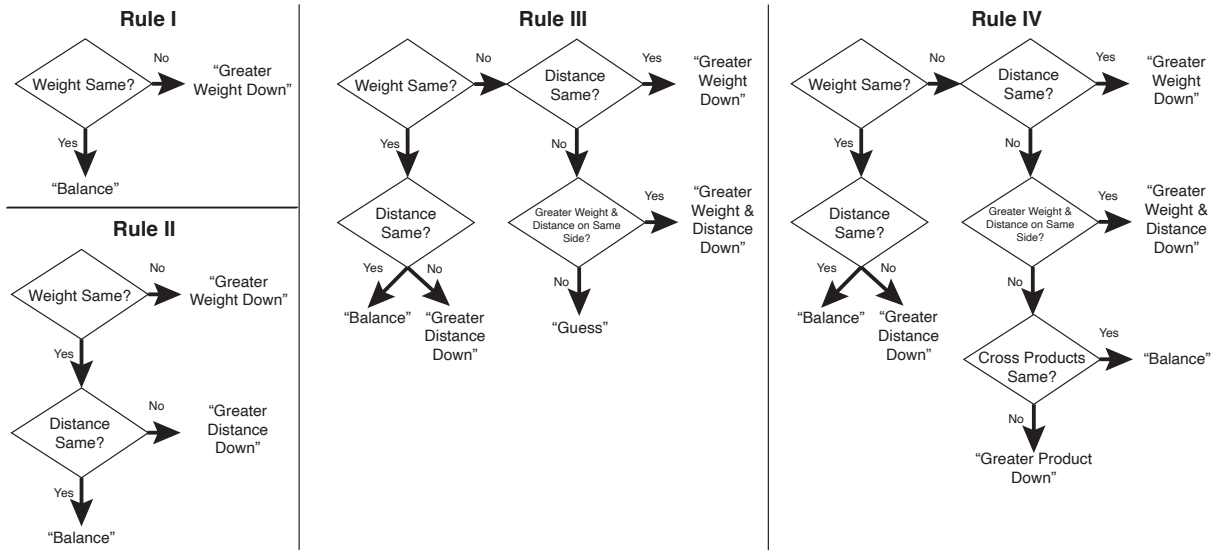

Figure 3. The rule-based models of Siegler (1976). Children begin by considering only a single dimension (typically the weight of the blocks; Rule 1), then separately consider weight and distance (Rule 2), and then recognize that weight and distance combine but are unsure how (Rule 3), before finally integrating weight and distance appropriately (Rule 4).

correctly.

To test for the use of these rules, Siegler (1976) proposed a classification of six beam configurations that would elicit different predictions depending on which rule people are using (see Figure 4). One configuration was 'balanced' where the same size stack of blocks was positioned the same distance away on each side, providing a symmetrical system with no net torque. The 'weight' configuration had stacks on each side that were equally distant from the center, but one side had more blocks (and therefore more weight). Conversely, the 'distance' configuration trials had the same number of blocks on each side, but the stacks on one side were positioned further from the center and therefore that was the side that would fall. The remaining three configuration were 'conflict' trials, where one side of the beam had more weight but the stack was closer to the center, while the other side had less weight that was positioned further out. In 'conflict-balance' trials, the blocks were perfectly positioned to net out to no torque. In the 'conflict-weight' trials, the side with more weight but less distance would fall, while in the 'conflict-distance' trials the side that

was further but with less weight would fall. According to the rule classification that Siegler proposed, every rule should treat all beams within the same classification identically, but the application across classifications will differ according to the rule used (see Figure 4). Therefore the rule that a person is using to make their predictions can be inferred based on their pattern of responses across different beam classifications.

This framework has formed the backbone of most research into balance beam judgments: nearly all studies use the same six beam classifications, and most assume that people use a set of rule-based decision trees to form their judgments. However, even within this framework there have been debates about the exact set of rules that people use. For instance, rather than using Rule 3 – guessing on the conflict trials – some researchers have suggested that people use systematic yet biased rules on those trials: e.g., that they add rather than multiply the weight and distance (Normandeau et al., 1989; Wilkening & Anderson, 1982), or assume that any conflict should cause the beam to balance (Normandeau et al., 1989).

3.2 Rules, or rule-like behavior?

In contrast to theories that suggest people use rules to make balance judgments, other researchers have debated whether people are truly using rules to form their judgments, or whether their judgments are based on more continuous processes that appear to be rule-like due to the classification process (Quinlan, van der Maas, Jansen, Booij, & Rendell, 2007). The Rule Assessment Methodology (Siegler, 1976; Siegler, Strauss, & Levin, 1981) classifies participants by how well their behavior matches a hypothetical rule-user, but does not allow for the use of multiple rules or other processes that might happen to produce the same output (Jansen & van der Maas, 2002; Kerkman & Wright, 1988; Wilkening & Anderson, 1982). Proponents of continuous processing have therefore suggested that “rule-like” judgments do not necessarily imply the use of rules, and, to prove this point, have developed connectionist models that are classified by the Rule Assessment Methodology as using rules that develop in the same stages as expected for people (McClelland, 1988, 1995; Schapiro & McClelland, 2009; Shultz, Schmidt, Buckingham, &

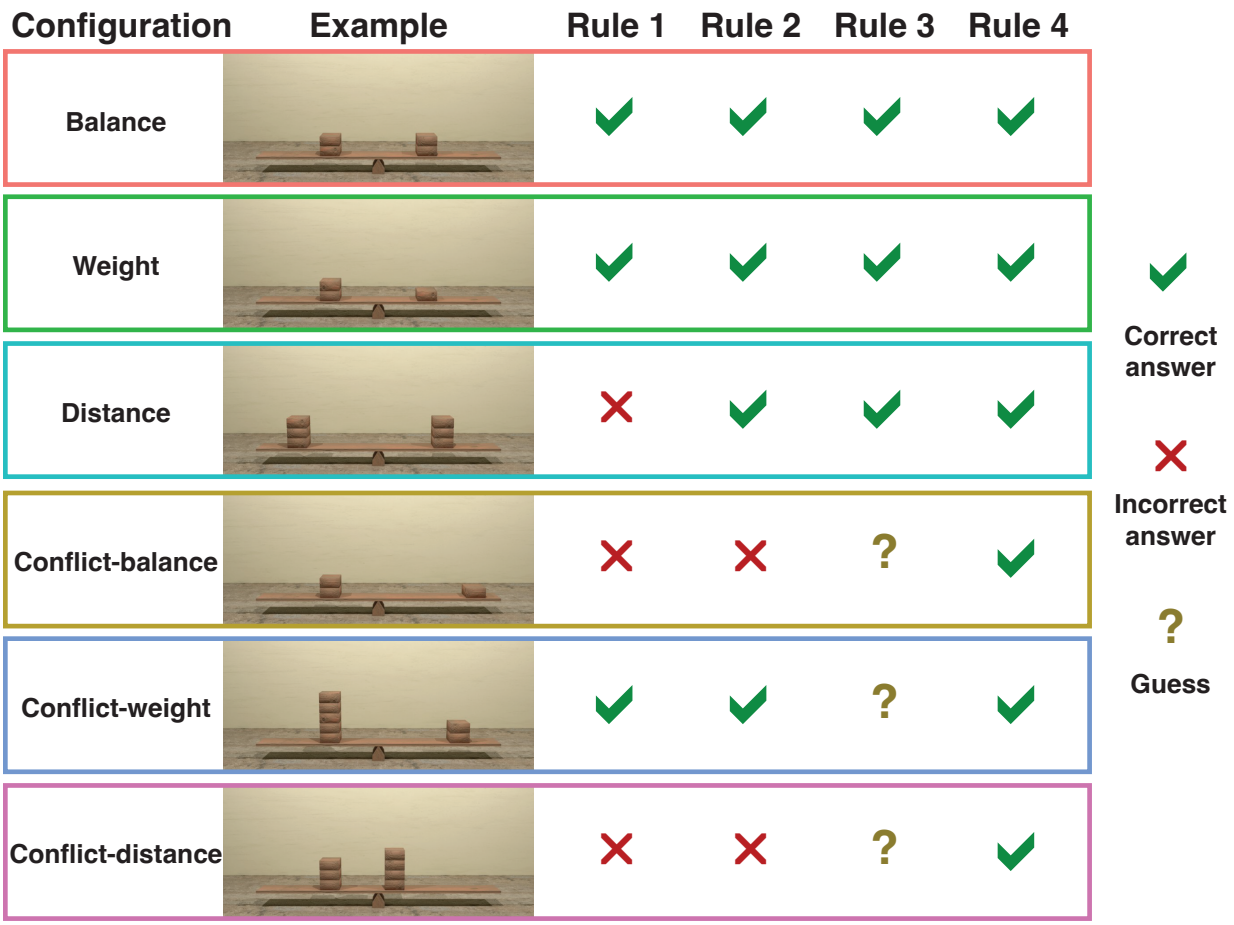

Figure 4. Examples of each of the beam configurations and associated accuracy of predictions from each rule. People using Rule 1 should only get the 'balance,' 'weight,' and 'conflict-weight' trials correct, incorrectly predicting the 'distance' trials will balance and that all of the conflict trials will fall to the side with the greatest weight. Rule 2 would produce the same predictions except in the 'distance' configurations, which would be predicted correctly. Using Rule 3, children should predict all of the simple beams correctly but 'muddle through' and guess on all conflict trials. Finally, children using Rule 4 should make predictions about all configurations accurately. Surrounding colors are indicators of the beam configuration that are common to all figures in this paper.

Mareschal, 1995).

Therefore, more advanced criteria were developed to determine whether people are using rules or rely on continuous processes. If rules are driving behavior, then behavior should be consistent and invariant across all classes of problems that the rules treat as the same (Jansen et al., 2007; Quinlan et al., 2007).[2] Thus a necessary but not sufficient test for the use of rules is 'bimodality' (Jansen & Van der Maas, 2001): if each individual is using a single rule, then their predictions for all balance beams within a single classification should either be all correct or all incorrect (excepting response noise), and so the distribution of accuracies across participants should be bimodal, clustering around 0% and 100%. Indeed, Jansen and Van der Maas (2001) searched for evidence of bimodality on distance problems, and found that the vast majority of children tested received either perfect accuracy or all incorrect marks, providing evidence for the use of rules.

However, Ferretti and Butterfield (1986) found evidence against the invariance of rules across all instances of a classification: when children are presented with balance beams for which the difference in torques between the two sides is particularly large, they often behave as if they are using a more advanced and more accurate rule than would be expected based on their responses to less extreme problems. This change in performance, called the "torque difference effect," is inconsistent with a system of rules that strictly compares weight and distance, since if people are using fixed rules then all beams of a given configuration type should be treated identically. Jansen and van der Maas (1997) argues that this classification difference is only statistically significant in the most extreme level and therefore children use consistent rules for most beam configurations, leading others to argue that rules are used for 'difficult' problems while 'easy' problems are solved by visual heuristics (Zimmerman & Pretz, 2012). However, there is a numerical increase in accuracy across all levels of torque difference (especially for the conflict-type beams) which cannot be explained by the strict use of rules (Schapiro & McClelland, 2009). The torque difference effect has therefore often been highlighted as a signature of more continuous

[2] However, c.f. Siegler (1996) for a discussion of how children might use multiple rules at a transition point.

underlying process, since many connectionist models can capture this effect (McClelland, 1995; Shultz et al., 1995), although a rule-based ACT-R model of balance beam judgments also displays a torque difference effect, albeit only when the model is transitioning between rules (Rijn et al., 2003).

3.3 The presentation and choice of balance beam stimuli

Human predictions about how balance beams will tip have been studied extensively in the past five decades, across a variety of age ranges and using a number of different modeling techniques. Much of this research has pointed towards the theory that people do use rules to make these predictions, but other evidence suggests that rules cannot explain the full range of human behavior on these tasks. Because there is a large set of empirical, developmental, and computational data on this task, and because this task is one for which rules can explain some but not all of the empirical data, we consider these stability judgments to be a good test-bed for capturing the trade-off between rules and simulation in physical judgments. However, because so much focus has been placed on the use of rules to support balance beam judgments, we consider why this is so different than the towers task of Battaglia et al. (2013) and how we might design stimuli that bridge the gap between the two.

Prior research into balance beams has typically used stimuli that consist of diagrams that are marked at the distances from the center that weights can sit, use identical and discrete blocks to represent the weight, and often only allow a single stack of blocks on each side of the beam (see Fig. 1A). In this way, the two dimensions of weight and distance can be quantified with almost no uncertainty, and the comparison between sides is easier because it does not require integrating the effects of different weights at different distances on the same side. This is in contrast to real-world judgments of stability, where there can be multiple objects of different, uncertain weights at multiple distances from the point of balance – for instance, imagine a waiter balancing a full tray of different dishes on one hand. Furthermore, we expect that showing people diagrams of balance beams rather than images with more realistic features should make people more likely to use

analytic reasoning as opposed to simulatable mental models (Schwartz, 1995). Thus people may have been using mostly rules in prior studies of balance beams because of the artificial nature of the stimuli rather than because reasoning about stability is only predicated on rules.

We therefore consider how people might make predictions about balance beams using more realistic and uncertain stimuli. We use realistic, 3-D images of balance beams, remove any distance markers along the beam to avoid certainty about the distance dimension, allow objects at multiple distances per side, and across different experiments we add uncertainty to either the estimation of weights (using non-uniform blocks and stacks or blocks of different materials) or the position of the pivot point or size on which the beam balances.

Even with this uncertainty, simple rules could apply: people can noisily estimate the weight on each side of the beam and apply the simple heuristic that the side with the greater estimated weight will fall regardless of where those objects are placed. On the other hand, these changes make the balance beams more similar to the towers of Battaglia et al. (2013) where people's judgments are well explained by simulation. We therefore test whether people's predictions with these more realistic, uncertain stimuli rely solely on rules, solely on simulation, or whether they can only be explained as a selection of strategies that combine rules and simulations.

4 Experiments 1–3: Stability judgments under uncertainty

We tested judgments about stability across three experiments that varied in the way balance beams were modified to introduce complexity (see Fig. 5). In Experiment 1, the stacks of blocks could be replaced by stacks of non-uniform shapes to introduce uncertainty in the size of the objects, and therefore the weight comprising each stack, in order to test whether rules might be used when there is less certainty about the relevant physical quantities. In Experiment 2, the blocks could be made of different materials – wood, brick, or iron – so that participants were required to account for physical density when making their judgments. In Experiment 3, the size and location of the pivot that supported the beam was varied so that participants would need to take the weight of the beam into account to judge balance. Other than stimulus differences, the

task was identical across all three experiments: participants would view a computer-generated image of a balance beam and were asked to indicate whether they believed it would fall left, fall right, or balance.

4.1 Procedure

Participants were recruited online from Amazon's Mechanical Turk using psiTurk (Gureckis et al., 2016). Participants were limited to those with IP addresses in the United States. All participants were compensated for their time depending on the length of the experiment ($1.20 for Experiment 1 and Experiment 3, which took ~8-12 minutes, or $1.50 for Experiment 2, which took ~10-15 minutes). There were 25 participants each in Experiments 1 and 2, and 48 participants in Experiment 3. Sample sizes were estimated and found to provide reliable results for Experiment 1, then set to provide approximately the same number of data points for each stimulus for all other experiments. These and all other experiments were approved by MIT's Committee on the Use of Humans as Experimental Subjects, approval number 08120030.

There was a common cover story across all experiments: a friend with a poor sense of balance was building sculptures using a computer program and participants were asked to help him decide whether they would fall, and if so, which way. Participants were always introduced to the task and the different types of stimuli they would encounter in each experiment. They were then asked to make judgments about an introductory set of balance beams that they were told had already been built, and therefore would receive feedback after their judgment by observing a movie of how the beam would fall. The introductory stimuli were counterbalanced to provide equal number of beams that balanced, fell to the left, or fell to the right, and were designed to not have conflicts between weight and distance.

In the experiment, participants saw static balance beams and were asked to make judgments about whether the beam would "tip left", "stay balanced", or "tip right" by clicking on one of three buttons with their mouse. Both the response made and the time since the start of the trial

were recorded.[3] Participants were given no feedback on these trials.

At the end of the experiment, participants were asked to answer two open-ended questions: “Did you use any particular strategies to decide when a sculpture would balance or tip?” and “Did you notice anything about this task you would like to tell us?”

4.2 Materials

Balance beams for all experiments were constructed to conform to one of the six classification types from Siegler (1976): balance, weight, distance, conflict-balance, conflict-weight, or conflict-distance. To avoid any directional bias in judgments, beams were mirrored and participants were equally likely to see either mirrored version of the beam; however, for ease of reporting we normalize all responses to the version of the beam that falls to the left (or, in the case of the conflict-balance beams, the side with more weight would be on the left). For instance, if the version of the balance beam that a participant saw would actually fall to the right but they judged it to fall left, their response would be recorded as ‘right’ for all analyses.

All stimuli were created in the Cycles rendering engine of the Blender 3D modeling software (Community, 2018). Introductory movies were made in Blender using its built-in physics engine to simulate the motion of individual blocks; how the beam tipped or balanced always matched with the expected behavior from the beam classification.

4.2.1 Experiment 1: Shapes. Experiment 1 was designed to reduce the certainty in the weight stacked on each side of the beam by allowing objects of non-uniform shape to rest on one or both of the sides. To create the trials, we first developed base configurations that consisted of a set of weight and distance “stacks” for each side of the beam to fit into the six Siegler classifications. These base configurations were then transformed into four separate trials. In the ‘blocks’ version, for each stack a number of blocks equal to the weight was placed at the

[3] Because responses were indicated by clicking on a button, reaction times could be contaminated by autocorrelation effects: if participants indicated the same response twice in a row, they would not need to move their mouse and therefore would respond more quickly. We therefore used these times mainly to ensure that participants were paying attention and not simply “clicking through” the experiment.

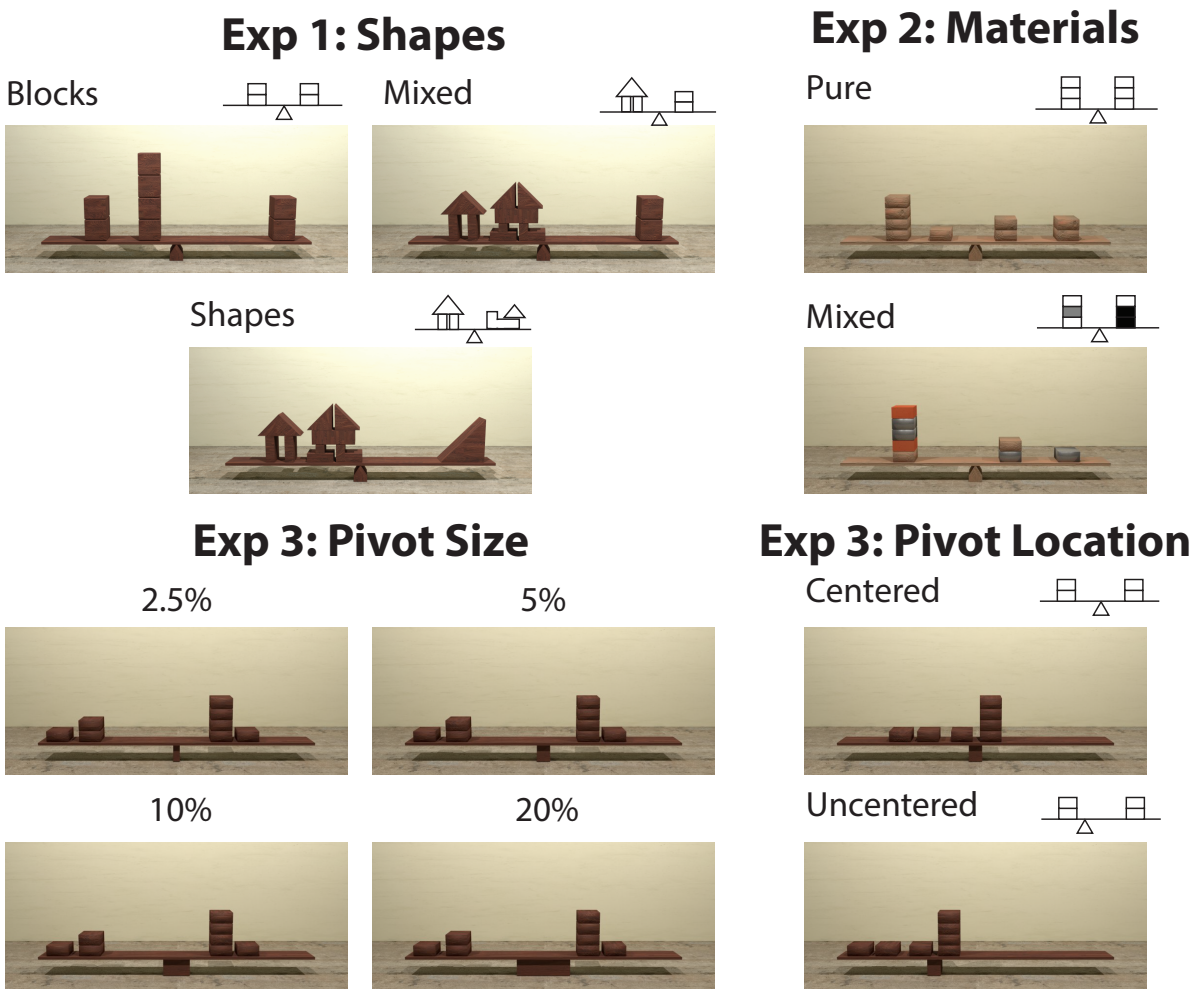

Figure 5. Example stimuli from each of the three experiments. In Experiment 1, the stimuli could be comprised of pure blocks, pure shapes, or shapes on one side and blocks on the other. In Experiment 2, the blocks could all be the same material, or could be a mixture of wood, brick, or iron. In Experiment 3, the pivot size could be resized from 2.5% of the length of the beam up to 20% of the length of the beam, or could be positioned away from the center of the beam.

appropriate distance from the center of the beam. In the 'shapes' version, all of these stacks were replaced with non-block shapes, either individually or stacked on one another. These shape groups were created by picking from one of 29 stable, canonical shape stacks, rescaling that stack so that it was the same volume of material as the blocks,[4] then placing that shape stack so that the center

[4] All blocks and shapes had the same, constant depth, so this was equivalent to equalizing the area of the closest face.

of mass is positioned at the appropriate distance from the center of the beam. Each canonical shape stack was arbitrarily chosen, but created such that the stacks always had some visible distance from their nearest neighbors. Finally, there were two 'mixed' beams in which one side was comprised of blocks and the other side was comprised of shapes. See Figure 5 for examples.

We created five base configurations each of the basic trials, and seven of each of the conflict trials, for a total of 36 base configurations and 144 trials. All participants saw all trials, but whether the beam was mirrored or not was counterbalanced across participants. Trials were balanced so that there were an equal number of balance beams that fell left, right, or balanced.

4.2.2 Experiment 2: Materials. Experiment 2 was designed to test whether people can account for different material densities when making balance judgments, rather than simply counting the number of blocks. In this experiment, blocks could be made of materials with a wide range of densities: either wood, brick, or iron. To construct these stimuli, we first gathered material densities online (*The Engineering Toolbox: Densities of Common Materials*, 2010), assuming the wood was a heavy wood such as elm or mahogany at $0.8\ g/cm^3$, that brick was $2.0\ g/cm^3$, and that iron was $7.2\ g/cm^3$. This produced densities in a ratio of 1 : 2.4 : 9.

Half of the stimuli were created so that the blocks were always made of the same material (the 'pure' trials), and the other half were created so that the blocks were made of different materials (the 'mixed' trials); however, the balance beam was always made of wood. Stimuli were randomly generated to conform to the six Siegler classification types, using the actual weight of the blocks as the 'weight' measure, rather than just the number of blocks.

For each of the pure and mixed trial types, there were eight randomly generated trials for each of the three basic trial types, and sixteen trials for each of the three complex types, for a total of 144 trials. All participants saw all trials, but mirroring of trials was counterbalanced across participants. Trials were balanced so that there were an equal number of balance beams that fell left, right, or balanced.

To acquaint participants with the different materials, in the introduction participants were acquainted with the materials and shown movies to demonstrate that one brick was heavier than

two wood blocks but lighter than three, that one iron block was heavier than three bricks but lighter than four bricks, and that one iron block balanced with 9 wood blocks but was lighter than 10 wood blocks. After the introductory trials, this information was again summarized to remind participants that brick is 2-3 times heavier than wood, iron is 3-4 times heavier than brick, and iron is about 9 times as heavy as wood. Finally, during the experiment, labeled pictures of a single block of each of the three materials was shown beneath the trial to prevent confusion; however, the material densities were not shown again.

4.2.3 Experiment 3: Pivot Size and Location. Experiment 3 was designed to determine whether and how people comprehend balance beams when the pivot is not a central point, but instead is a table on which the beam can rest, or is positioned off-center. Trials were created separately to test the effect of pivot size and pivot location, but were combined in the experiment to ensure participants observed a nearly equal split between trials that balanced, fell left, or fell right.

The trials to test for pivot size were formed based on the weight, distance, conflict-weight, and conflict-distance configurations, so that if the pivot were a point, the beam would fall. However, rather than being a point, the pivot was a box with a width that was either 2.5%, 5%, 10%, or 20% of the width of the beam. There were 10 trial bases for each of the four beam configurations, half of which used a centered pivot, and half in which the pivot was placed off-center. Because the beam no longer rests on a point, configurations of blocks that would fall in this idealized situation will not necessarily fall with a larger object to balance on. We therefore constructed these stimuli so that two of the configurations (one with a centered pivot, one off-centered) would stay balanced on all four beam widths, another two would tip on the 2.5% pivot but balance on the rest, another two would balance on the 10% and 20% pivot but not the others, another two would only balance on the 20% pivot, and the final two trials would always tip even on the largest pivot. Because the configuration of the blocks was the same across the four different pivot sizes, participants only saw two of the four possible pivot sizes for a given configuration – one standard beam and one mirrored beam, for a total of 80 pivot size trials.

The pivot location trials tested whether people would account for the weight of the beam

itself when that would differentiate how the whole beam would fall. These trials were based on either conflict-weight or conflict-distance configurations, so that it would fall according to that classification if the pivot were small and centered. However, each trial also had a 'shifted' version where the blocks and pivot were moved together so that the beam would fall in the opposite direction from the basic trial (e.g., if the basic trial were a conflict-distance trial, it would fall to the side with more distance when the pivot were centered, but would fall to the side with more block weight in the shifted version, and vice versa). There were twelve base trials each of the conflict-distance and conflict-weight versions, with half using a pivot that was 2.5% of the beam width and the other half using one that was 5% of the beam width. Participants all observed both the normal and shifted version of each trial, but one of the pair was always a mirrored version, for a total of 48 pivot location trials.

These trials could not be perfectly counterbalanced between the three outcomes (balance, fall left, or fall right), but were designed to be as close as possible. Because half of the pivot size trials balance but the pivot location trials always fell, 40 of the 128 trials (31.25%) were balanced, and because of mirroring an equal number of the remaining trials fell left or right.

4.3 Transparency and Openness

We report how we determined our sample sizes, all data exclusions, and all manipulations and measures in these studies. All data, analysis code, and research materials are available at `https://github.com/kasmith/balance_beams`. The study's design and analyses were not pre-registered.

4.4 Behavioral Results

Participants' average predictions across all experiments and conditions can be observed in Figure 6. Interpreting these result is challenging without a comparison to a normative model of how people *should* behave, and so most of our analysis is performed in the following section after our model is described. However, here we apply standard tests of rule use to the most basic balance beams and show that they do not capture participants' behavior nearly as well as would

be expected from prior studies. Additionally, we report further evidence against the use of rules in the Appendix (Section A1).

Because the non-standard beams we used might be more likely to rely on simulation, we limit our standard tests for rule use to the most basic stimuli in our experiments. Basic balance beams were defined as those that did not vary from the simple beam types used in prior experiments: the blocks-only stimuli of Experiment 1 (36 trials), the beams with only a single material of Experiment 2 (72 trials), and the trials from Experiment 3 with the smallest, centered pivot (20 trials).

First, we applied the Rule Assessment Methodology of Siegler (1976) to classify participants by the closest matching rule. To assign a participant as Rule 1, 2, or 4, that participant's predictions must agree with the predictions of the rule at least 86.7% of the time (matching with 26 of 30 trials used by Siegler). To assign a participant as Rule 3, that participant had to achieve an accuracy of greater than 83% on the non-conflict trials (including at least 75% accuracy on the distance trials) and had to deviate from weight cues in at least 22% of the conflict trials.[5] All other participants were deemed 'unclassifiable.'

As can be seen in Table 1, many participants did not appear to be using a consistent rule, with almost half of the participants in the Shapes and Pivot experiments being unclassifiable. This is in contrast to prior studies, where typically fewer than 10% of participants could not be classified according to the Rule Assessment Methodology (Siegler et al., 1981). Furthermore, the majority of participants who could be classified were assessed to be using Rule 3 (72%), which is the rule that is often thought of not as a single rule but a set of heterogeneous strategies for "muddling through" (Normandeau et al., 1989).

In addition, a necessary but not sufficient test for the use of rules is 'bimodality' (Jansen & Van der Maas, 2001): beam classifications were designed so that these rules should treat every instance within that class identically, and therefore if people are using a rule that accurately judges

[5] We did not include the 'addition' rule because the stimuli were not designed to be able to disentangle this rule from either Rule 3 or 4. However, see Section 5.2.2 for a model-based rule assessment using the addition rule.

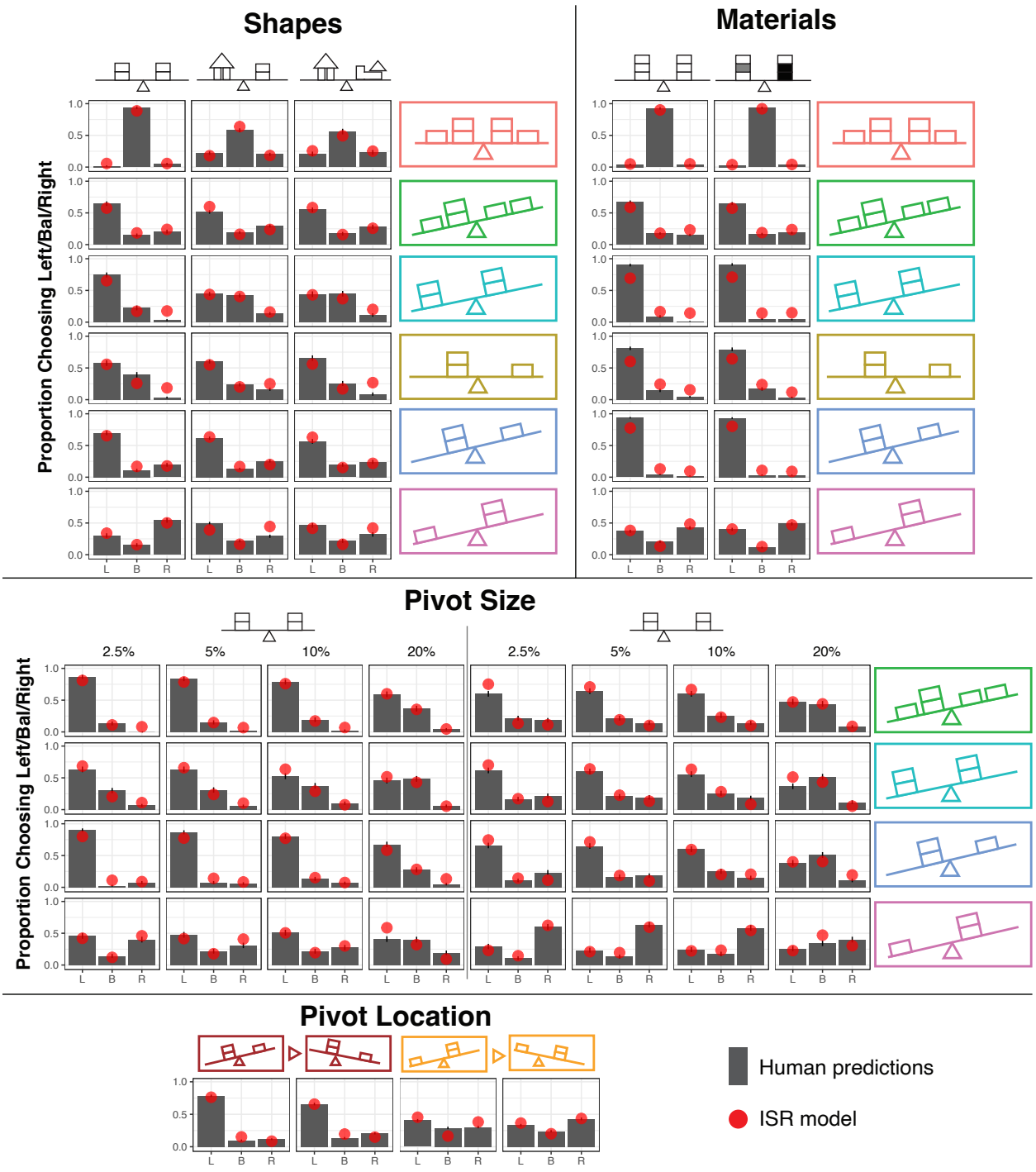

Figure 6. Distribution of predictions by participants (bars) and the Integration of Simulation and Rules model (red dots) across all experiments and balance beam variants. Error bars represent 95% confidence intervals on expected participant responses for that trial category. In general, the ISR model explains both the correct and incorrect predictions that participants make very well, including how those predictions are expected to vary across different trial types.

Table 1

Assignment of rule types to participants of each experiment. Numbers indicate counts of participants with each assignment. Rule classifications were based on Siegler (1976), using only the trials most similar to those in prior experiments.

	Rule 1	Rule 2	Rule 3	Rule 4	Unclassifiable
Exp 1: Shapes	4	3	6	0	12
Exp 2: Materials	0	5	14	0	6
Exp 3: Pivot	1	1	21	2	23

that classification, they should have perfect accuracy across all instances within a class (excepting response noise), and conversely with an incorrect rule should have zero accuracy across all instances. Crucially, this implies that there should be very few participants who provide an accurate response for some but not all of the instances. Thus if participants' accuracies for a single instance were plotted as a histogram, this would appear as a bimodal distribution with clusters at 0% and 100% accuracy, and very few participants would appear in between.

Following the example of Jansen and Van der Maas (2001), we look for bimodality on the distance problems, and further limit the configurations considered to the most basic balance beams, as described above. However, unlike Jansen and Van der Maas (2001), we found no clear evidence of bimodality in any of the three experiments: zero or perfect accuracy was found for only 36% of participants in Experiment 1, 32% of participants in Experiment 2, and 65% of participants in Experiment 3 (though there were only two basic distance trials in the third experiment, perhaps inflating this number; see Figure 7).

If participants were using rules exclusively for any subset of balance beams, much less for the entire experiment, we would expect these rules to be found at least on the simplest trials most similar to balance beams from previous research. However, because we cannot easily classify most participants on even the basic trials, and because we do not find evidence of bimodal response patterns as a sign of consistent rule use, it does not appear that our participants were in

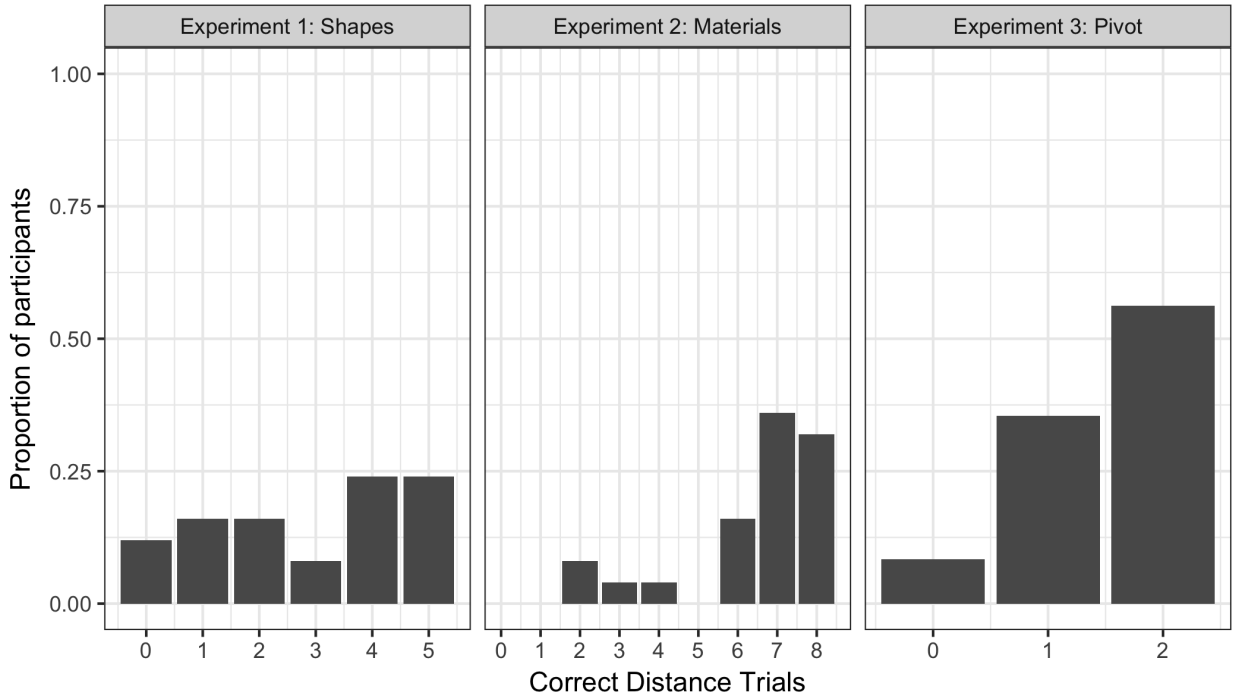

Figure 7. Histogram of participants with different solution rates on the basic distance trials, split by experiment. A large number of participants with zero or perfect evidence would be evidence suggestive of rule-use; however, we do not find evidence of that bimodality even on the simplest distance trials.

general using rules alone to solve this task (and see Section 5.4 for further analysis of individual uses of rules). We therefore consider a model that builds strategies out of both rules and simulation to make predictions about how balance beams will fall.

5 The Integration of Simulation and Rules model

5.1 Model structure

We suggest that instead of relying solely on rules or solely on simulation, people select from both options to make judgments about balance beams. Crucially, we assume that rules and simulation comprise the building blocks that people compose into integrated strategies for solving their problems (Siegler, 1988).

As described in Section 2.5, the Integration of Simulation and Rules (ISR) framework was

designed to capture this theory. Here we instantiate it into a concrete model that is designed to solve balance beam problems like people do (Fig. 8). Because this model considers a single question – whether and how the beam will fall – the model assumes that all strategies will be constructed to answer this question.

The ISR model constructs integrated strategies by linking together different primitive strategies – rules or simulations – similar to the decision trees proposed by Siegler (1976). These primitive strategies can either provide a prediction about what will happen to the balance beam, or are *undecidable* and pass the decision through to the next primitive strategy (e.g., the 'symmetry' rule can suggest that the beam will balance if the structures are symmetric, but does not make any predictions if they are not; see Fig. 8A). For tractability, we only consider "pass-through" programs for integrated strategies, in which primitive strategies can either provide a prediction, or pass the decision along to another primitive strategy for assessment.[6]

We consider four primitive strategies suggested by the literature, and detailed below: (1) *symmetry judgments* that predict the beam will balance if it is symmetric, (2) the *weight rule* that predicts the side with more weight will fall, (3) the *distance rule* that predicts the side with objects further from the pivot will fall, and (4) forming predictions based on the outcome of *physical simulation*.

In theory, an integrated strategy could be formed from any ordering of the primitive strategies (or the empty set, corresponding to random guessing). In practice, however, we do not expect that people will consider every possible strategy formed in this way, but instead will select from a small set in order to avoid the effort required to search through a large number of strategies (Milli, Lieder, & Griffiths, 2021). Empirically, we find this to be true – we only found evidence that people that people use two integrated strategies other than guessing: symmetry -> weight rule -> physical simulation (SWP), and symmetry -> physical simulation (SP). See Section 5.2 and Appendix

[6] We make this simplifying assumption because (a) without it the space of potential integrated strategies would be infinite, which would be intractable to search through, and (b) this assumption would be sufficient to reproduce any of the rules proposed by (Siegler, 1976).

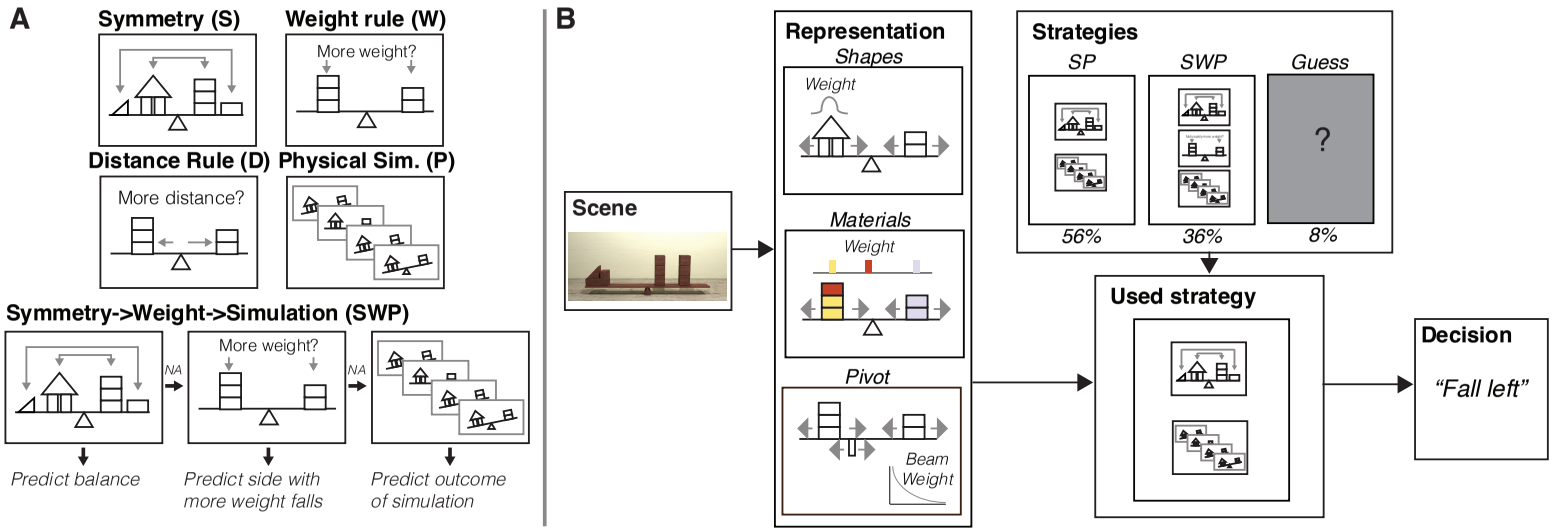

Figure 8. Diagram of the Integration of Simulation and Rules model instantiated for balance beam judgments. *A.* Integrated strategies can be constructed out of four primitive strategies (*top*), chained together into a decision tree. For instance, in the symmetry->weight->simulation (SWP) strategy (*bottom*), the scene will first be checked for symmetry across both sides of the balance beam. If the beam is symmetric, it outputs a 'balance' judgment, but otherwise passes the decision onto the weight rule. If the weight rule notices that one side has enough more mass than the other, it outputs a prediction that side will fall, but otherwise it is *undecidable* and passes the decision to the physical simulation module. The simulation module then outputs the outcome predicted by a noisy simulation model. *B.* The overall construction of the Integration of Simulation and Rules model. People first noisily perceive the scene, with different uncertainty depending on the beam structure as defined by the experiment. Given this perception, they then probabilistically select an integrated strategy to apply to this representation, and use that to make their decision. The strategies displayed here (SP, SWP, and guessing) were the only strategies we found participants to use.

Section A2 for further details.

Next the ISR model must apply these integrated strategies to a scene (Fig. 8B). It first encodes the image of the scene into an object-based representation (Battaglia et al., 2013; Yildirim, Wu, Kanwisher, & Tenenbaum, 2019). This perception module will act differently for each experiment due to differences in the uncertainty about object locations and properties induced by the different scenarios (e.g., there is more uncertainty about the size/mass of non-block shapes), but will form a common representation that can be used for all rules and simulations.

However, for each scene that people encounter, they must pick one of the integrated strategies that they will use to form a prediction. We assume that this is a probabilistic selection from the available strategies, chosen concurrently with but not dependent on perception. Similar to how we must discover empirically the set of integrated strategies that people consider, the probability of selecting between strategies was estimated by fitting to human data.

Nonetheless, even though both the set of integrated strategies and their relative weightings were fit to empirical data, we consider whether those choices are in fact aligned with a resource-rational framework, in which the choice of strategies roughly maximizes the expected rewards and cognitive costs. Although calculating this utility requires assumptions about cognitive costs, we find in Section 5.5 that under a broad set of assumptions, resource rationality does hold.

5.2 Empirical use of strategies

In order to determine which integrated strategies people do in fact use, we used forward model building: starting from pure guessing, we added individual integrated strategies that explained participants' predictions across all three experiments better until there was no strategy that improved explanatory power. Using this technique, we found that only two strategies are required to explain participants' behavior: symmetry -> weight rule -> physical simulation (SWP), and symmetry -> physical simulation (SP). To validate this finding, we compared this SP/SWP model to a number of alternatives (Fig. 9), including the possibility of using *any* integrated strategy (*All Strategies*), using only integrated strategies that combined symmetry, weight, and physical

simulation (in any ordering; *S/W/P Strategies*), including integrated strategies that used the distance primitive strategy similar to how the main model used the weight primitive strategy (SDP, SDWP, or SWDP; *Plus Dist*) or instead of integrated strategies that included the weight rule (*Instead Dist*), using any integrated strategy not including each of the three observed primitive strategies (*No Symmetry/Weight/Simulation*), using only physical simulation (*Simulation Only*), and adding any individual strategy on top of SP/SWP (see Appendix Section A3.1).

Because all of these model variants have different numbers of parameters (to fit the mixtures of different numbers of strategies, or because some perceptual parameters are irrelevant without corresponding strategies), we used crossvalidation techniques in order to compare them on equal footing. We randomly split the trials in half – with an equal split for each of the three experiments – fit all models to half of the data, and compared model likelihoods on the other half. We then took the difference between the baseline (SP/SWP) and comparison model log-likelihoods such that positive values of this metric would indicate the comparison model outperformed the baseline. Finally, we repeated this process 50 times to test the range of performance differences under various trial splits.

We find that any change to the model that adds integrated strategies added at best marginal explanatory power, whereas any model variants that change or remove strategies have reliably worse performance (see Fig. 9 and A3). There was no model that reliably performed better than the baseline – allowing all integrated strategies provided the largest numerical improvement but still failed to do so reliably ($\Delta LLH_{CV} = 15.4$, $90\%\ CI = [-1.5, 30.6]$), and there was no additional individual strategy that outperformed the baseline (Appendix Section A3.1).

5.2.1 Using only physical simulation. The above analysis contains a comparison to a standard model of physical simulation, noted as 'simulation only' in Fig. 9. This model assumes that people have perceptually uncertainty and use noisy unbiased simulation, similar to many other cognitive models of intuitive physics (Battaglia et al., 2013; Smith & Vul, 2013; Hamrick et al., 2016; Sanborn et al., 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Zhou et al., 2022), but does not allow for the use of any rule-like primitive strategies. Surprisingly, this model

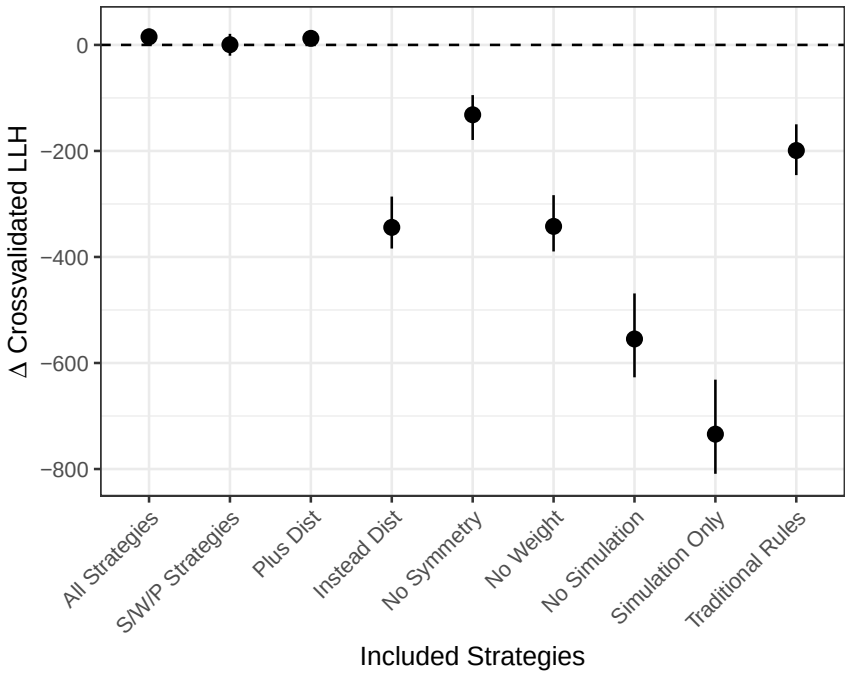

Figure 9. Differences in cross-validated log-likelihood for model variants versus the baseline (SP/SWP) model. Points indicate mean difference in log-likelihood over 50 samples, bars indicate the 10th to 90th quantiles. The dotted line indicates parity with the baseline model, so points below the line indicate worse predictive power for that model variant. Models that add more possible strategies vs. the baseline model add very little explanatory power, whereas all models that remove or change strategies (including using traditional balance beam rules) fit participants' judgments much worse.

performs *worst* out of all model variants ($\Delta LLH_{CV} = -734$, $90\%\ CI = [-809, -631]$), suggesting that noisy simulation alone cannot explain behavior on this task.

5.2.2 Using only rules. While we showed that people cannot be using stimulus-based rules to make balance judgments (see Sec. 4.4), this does not preclude the possibility that people are still using the deterministic rules noted in prior literature, but have more perceptual uncertainty on this task than when judging more diagramatic balance beams in earlier work. We therefore test whether we can explain participants' predictions using a model that includes the same noisy perception as the Integration of Simulation and Rules model, but assumes that participants are using either the rules proposed by Siegler (1976) or the addition rule (Normandeau et al., 1989) to

form their judgments.[7]

To place the "traditional rules" model on the same footing as the ISR model, we assume that people start with the same types of perceptual uncertainty by filtering stimuli through the noisy perception module to produce judgments of the weights and positions of each stack of objects. This representation is then subject to each of the five proposed systems of rules to determine how a person using each of those rules would respond. The only difference between the rules used in this model and those proposed in prior literature is that these rules assume approximate matching, as strict equality cannot applied with noisy perception. For instance, weights are judged to be different only if the difference in the representation exceeds a threshold. Similarly, the rules model includes a parameter for distance matching as well as matching torques for Rule 4. This process is repeated 500 times to form a distribution of predictions from each rule for each trial. Finally, for aggregate model fitting, we assume that the participant group is comprised of people who each use one of the five rules, and thus fit the proportion of users of each rule in addition to allowing some responses to be 'guesses' similarly to the ISR model.

However, as can be seen in Fig. 9, the 'traditional rules' model does not explain peoples' predictions as well as the ISR model ($BIC_{ISR} = 22,845$, $BIC_{rules} = 23,722$; $\Delta LLH_{CV} = -199$, $90\%\ CI = [-246, -150]$), suggesting that people in aggregate do not only use rules on this task.

5.3 Performance across experiments and trials

Across all experiments, conditions, and trials, the ISR model using SP and SWP integrated strategies was approximately as accurate as people were (model: 50.6% accurate, participants: 52.8%). Furthermore, this model explained the differences in accuracy by trial very well ($r = 0.89$). This explanatory power transfers well across all of the experimental manipulations (see Figure 10 and Table 2).

[7] We do not include the "buggy rule" (van Maanen, Been, & Sijtsma, 1989) because it makes the same predictions about judgments as the addition rule, only differing in predictions of reaction time. We also do not include the qualitative proportionality rule (Normandeau et al., 1989), which predicts that people should always judge conflict items to balance, because no participant judged the beam to balance in more than 50% of conflict problems.

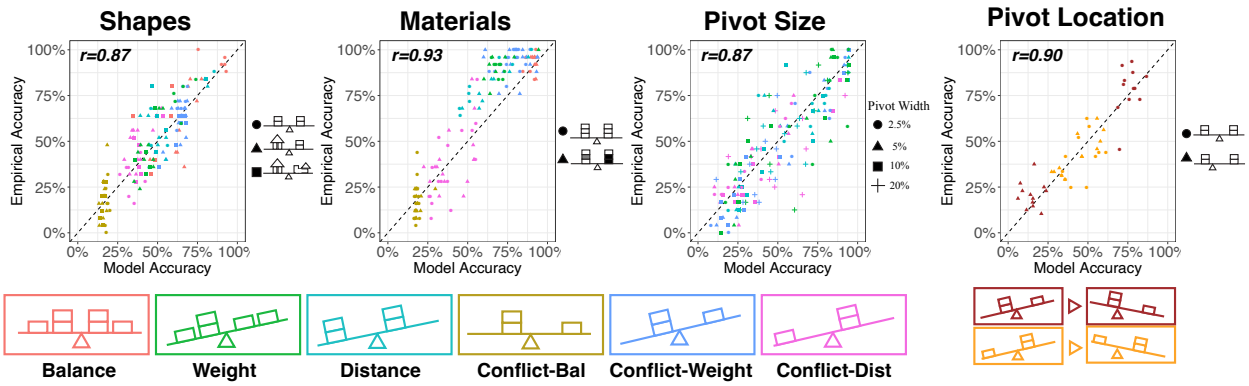

Figure 10. Comparison of model vs empirical accuracy across experiments. Each point is a single trial, comparing the accuracy of the Integration of Simulation and Rules model (x-axis) with the accuracy of participants (y-axis) on that same trial. Colors represent the Siegler beam classifications for trials, and the point shapes represent the variant along the dimension being tested in each experiment. Accuracy correlations were high (> 0.87) across all experiment types.

Table 2

Model accuracies and correlation by experimental manipulation. Correlation is calculated by trial, comparing the model accuracy with participants' accuracy for each trial. There is high correlation and low bias in accuracies across all experiments.

Experiment	Acc. Correlation	Human Acc.	Model Acc.
Exp 1: Shapes	0.87	49%	47%
Exp 2: Materials	0.93	62%	55%
Exp 3: Pivot Size	0.87	50%	52%
Exp 3: Pivot Location	0.90	46%	45%

This model predicts by-trial accuracy approximately as well as theoretically possible. We can measure how well we can predict the trial accuracies of half of the participants by using either (a) the other half of the participants, or (b) the ISR model. As can be seen in Table 3, model correlations are approximately at the same level as the split-half correlations, suggesting that it is fit up to the noise ceiling.

Table 3

Comparison of split-half correlations of trial accuracies by experiment versus model correlations using half of the participants. Participants were split in half 500 times, and the average accuracy by trial was correlated between the two halves (split-half) or one half accuracy was correlated with model accuracies.

Experiment	Split-Half Correlation	Half-Data Model Correlation
Exp 1: Shapes	0.74	0.80
Exp 2: Materials	0.92	0.91
Exp 3: Pivot Size	0.84	0.83
Exp 3: Pivot Location	0.88	0.87

In addition to testing how the model explains *whether* people are inaccurate, we also consider *how* their errors occur. We can do so by investigating how well the model captures the distribution of the three different responses: 'left', 'balance', and 'right'. As can be seen in Figure 6, across all trial variants within all experiments, the ISR model captures the pattern of responses, both correct and incorrect.

Together, this suggests that the model does a good overall job of explaining people's aggregate predictions about the balance beams in all three experiments. We next test whether the ISR model explains participants' predictions for each individual better than alternatives.

5.4 Individual use of just simulation or rules

While we find that participants' aggregate predictions can only be explained my a mixture of simulation and rules, it is possible that individual participants might be using only physical simulation or only rules. To test for this, we fit model parameters individually for each participant using the crossvalidation methodology above, and compared these fits to two different models that (1) assumed people used only physical simulation, and (2) assumed people were using a single rule type.

We fit the individual rules-only model in the same way as the aggregate model, but rather than assuming that responses were drawn from a mixture of rule-users, we assigned each

participant a single rule: the one that maximized the likelihood of that participant's data. To compare with the individual use of rules, we also fit the ISR model for each participant, by allowing the proportion of each strategy used (SP vs. SWP vs. guessing) to vary by individual.[8] Because these models had differing numbers of parameters, we compared these models using cross-validated likelihoods: splitting the trials in half, fitting the parameters to one half, then comparing the likelihood of the cross-validated trials, and repeating 50 times.

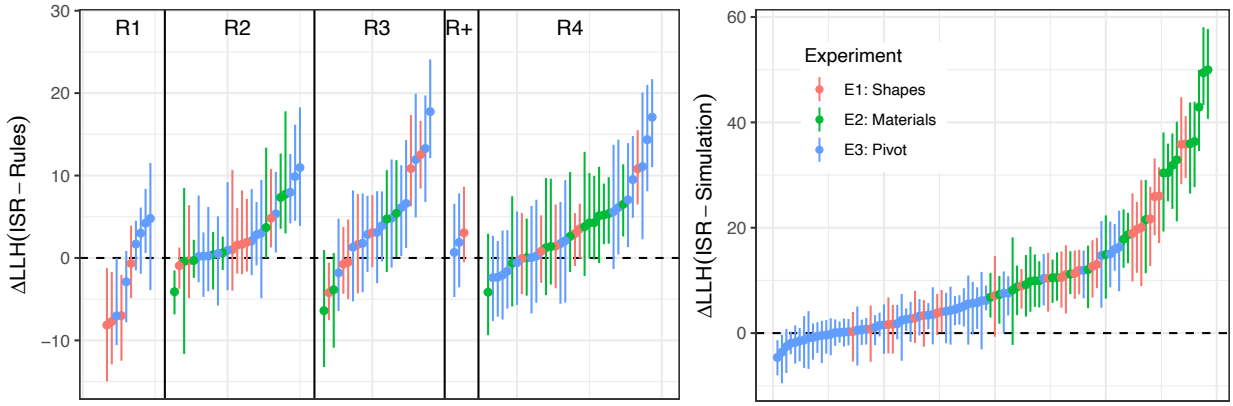

Figure 11. Comparison of Integration of Simulation and Rules vs. rules-only model (*left*) or simulation-only model (*right*) by individual, demonstrating that most participants' predictions are better explained by the Integration of Simulation and Rules model. The y-axis represents the difference in cross-validated log-likelihood between the two models, where higher values indicate the ISR model has better explanatory power. Each dot represents a different participant, color-coded by experiment and split by best fitting rule, and the lines are 5th to 95th percentiles of differences based on 50 cross-validated fits.

As can be seen in Figure 11 (left), the majority of participants made predictions that were better described by the Integration of Simulation and Rules model than by a rules model with probabilistic perception (73 out of 98, $p = 1.28 * 10^{-6}$). Furthermore, while the difference in model fits reliably favored the Integration of Simulation and Rules model at the 90% confidence level for 26 participants, only 5 participants were better described by the rules only model at that level. We can then check how the Rule Assessment Methodology (described in Section 4.4) would classify

[8] See Appendix Section A3.2 for discussion of individual differences in strategy usage.

these participants, and find that two of the five are unclassifiable according to this methodology, while the other three are classified the same using the rules model fits and the Rule Assessment Methodology (one using rule 1, one using rule 2, and one using rule 3).

Similarly, Figure 11 (right) shows that the Integration of Simulation and Rules model explains the majority of participants better than a model that uses simulation alone (85 out of 98; $p = 4 * 10^{-14}$). The Integration of Simulation and Rules model fit 42 participants reliably better at the 90% confidence level, whereas simulation alone outperformed the ISR model with only 2 participants.

Thus there may be some participants who use rule-like strategies alone or only rely on simulation, but even if so, this would be a small subset of the participants tested.

5.5 Resource rationality of the strategy choice

Although the prior analyses demonstrate that participants were using a mixture of strategies to make their predictions, if this is due to a resource rational choice, then we should expect that the mixture of strategies used overall will (approximately) maximize value of computation across the balance beam problems that participants encountered.

In the analysis below, we only consider strategies comprised of the symmetry, weight, and simulation primitives. We find that people do not use the distance rule in our experiments (Section 5.2), perhaps because they do not notice it as a possibility (see Discussion Section 9.2.3 for further detail), and thus seek to study whether people are making rational use of the primitives that they do recognize. An analysis including the distance primitive is included in Appendix Section A3.3.

We define value of computation of a single integrated strategy (S) on single problem as the benefit gained from getting that problem correct (R) multiplied the probability of being correct using that strategy, minus the cognitive cost of applying that strategy for that problem ($C_{S,i}$):

$$VoC_{S,i} = R * P(correct)_{S,i} - C_{S,i} \tag{3}$$

However, estimating these benefits and costs is challenging. While we did not directly incentivize participants for correct answers, we assume that there is intrinsic motivation to do well which can act as a reward; however, we cannot measure the magnitude of this reward. Similarly, the cost of using a strategy can be decomposed into the cost of using any of the individual primitive rules or simulation, but we cannot directly estimate the cognitive costs of those primitive strategies. We therefore fix the benefit of getting a correct prediction to 1 and define the costs of the various primitives as a proportion of this reward.

While we still cannot directly estimate the primitive costs, we can make some assumptions about their ordering. Because symmetry is often computed quickly, and perhaps pre-attentionally (Wolfe & Friedman-Hill, 1992), we assume that this component incurs the lowest cost. It is relatively easy for participants to compare two quantities, and so it is theorized that simple operations like the weight rule will be less costly than simulation (Davis & Marcus, 2015). Thus we can define the inequality for the symmetry judgment costs (c_s), weight rule costs, (c_w), and physical simulation costs (c_p) as $c_s \leq c_w \leq c_p$. Given the cost of the primitives, we can calculate the cost of using an integrated strategy on problem i as this cost times the probability that the primitive rule or simulation is reached in the strategy chain, and thus is used. This will differ by problem even for a fixed strategy (e.g., for the symmetry -> simulation strategy, if the balance beam is clearly symmetric, then the symmetry rule will almost always trigger and thus the probability of using physical simulation will be near zero, but conversely a very non-symmetric beam will almost always trigger both the symmetry rule check, which will fail, and then simulation). We calculate the chance of using each primitive strategy by running the Integration of Simulation and Rules model with a fixed strategy 100 times for each balance beam, then tallying the proportion of the time each primitive strategy is activated for that integrated strategy and beam. We therefore define the cost as:

$$C_{S,i} = c_{sym} * P(sym)_{S,i} + c_{wght} * P(wght)_{S,i} + c_{phys} * P(phys)_{S,i} \tag{4}$$

And so the value of computation for a given strategy and balance beam is defined as the

probability that the integrated strategy will provide the correct answer minus the associated cost. For an expected *set* of problems, the value of computation can be defined as the total value across all of those problems:

$$VoC_S = \sum_{i \in Problems} P(correct)_{S,i} - (c_{sym} * P(sym)_{S,i} + c_{wght} * P(wght)_{S,i} + c_{phys} * P(phys)_{S,i}) \quad (5)$$

If we make assumptions to fix the various primitive costs, we can use these equations to calculate the value of computation for each integrated strategy across the trials encountered in all three experiments, and thus determine the most efficient strategy to use to make predictions about those particular balance beams. Thus we can investigate how the most efficient strategy changes under different reasonable cost assumptions.

However, participants will not know precisely which balance beams they will encounter, and so we consider which strategies might be most efficient for *similar* sets of balance beams. To do so, we resample trials with replacement – keeping an equal number of trials per experiment – as a proxy for how experiments with similar trials might have been constructed. We repeat this process 100 times, and so for each cost setting, can calculate the proportion of the time any integrated strategy is most efficient across these proxy balance beam sets.

We can then investigate which integrated strategies are reasonably efficient across a large range of cost settings. We consider a grid of cost parameters, where c_{phys} ranges from 0.03 to 0.27 of a correct answer, and settings of c_{sym} and c_{wght} that ranged from 10% to 100% of the value of c_{phys} (see Fig. 12). We also consider what might be 'reasonable' strategies to use, and define these as integrated strategies that are the most efficient in at least 10% of the resampled trial mixtures.

Figure 12A shows the way that these reasonable strategies change under different cost assumptions. Each panel represents a different value of c_{phys}, and each of the boxes within that panel refer to the values of c_{sym} and c_{wght}. This box can have different colored sub-squares representing different sets of strategies: blue represents the SP strategy, red represents the SWP

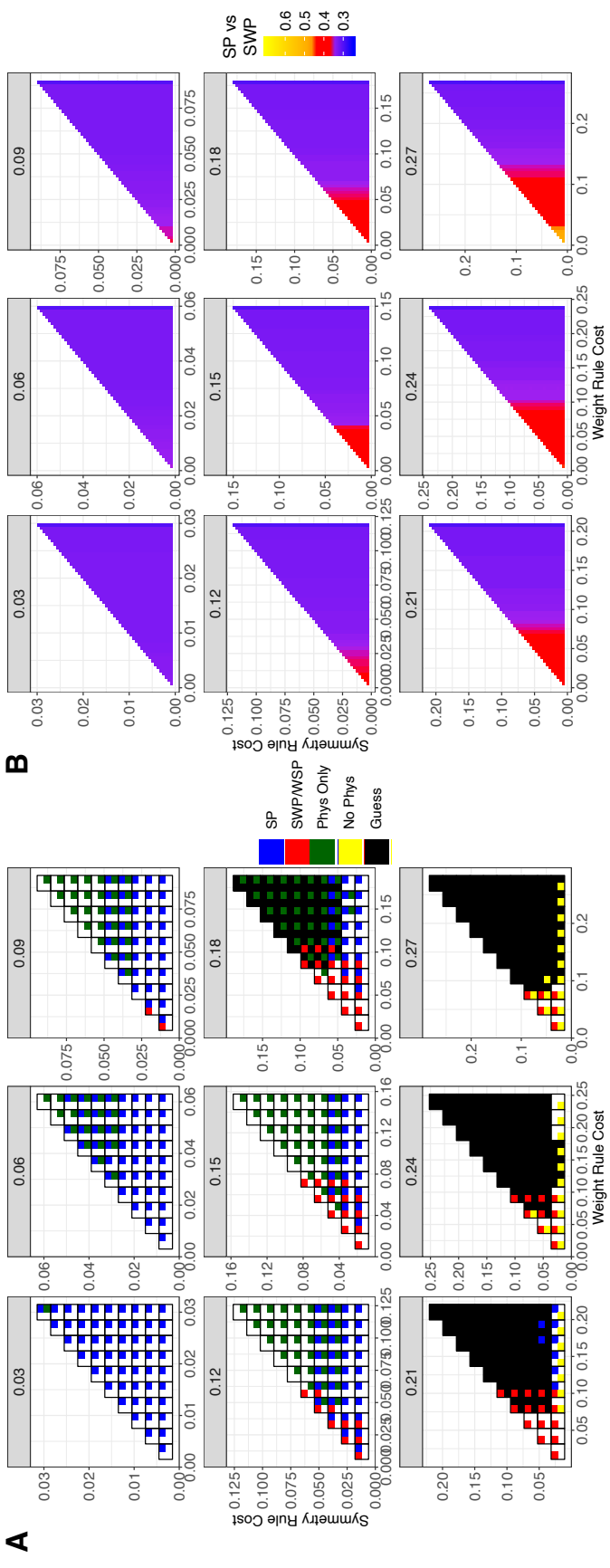

Figure 12. Plots of reasonable strategies (*A*) or optimal mixture between SP and SWP integrated strategies (*B*). Panels represent simulation costs, while plot axes represent weight rule (x) and symmetry rule (y) costs. Under the assumption of moderate simulation costs (middle row) and relatively low weight rule costs, both of the empirically observed strategies (SP and SWP) are the best strategy choices; similarly, in this area the optimal mixture between these strategies is close to the observed mixture of 39% (red area in panel B). Across a wide range of intermediate primitive strategy costs, the observed choice of integrated strategies and mixtures between them are consistent with a resource-rational strategy choice.

and WSP strategies,[9] green represents a strategy of just using physical simulation, and yellow represents all strategies that do not include physical simulation (S, W, SW, and WS). If the square has a black background, then simply guessing is also a reasonable strategy. For further results see also Appendix Figure A5.

Qualitatively, Figure 12A shows different regimes of cost settings. The top row, where the costs of the primitive strategies are low, approximates a regime without costs. Thus the weight rule is almost never used since, if costs are not an issue, it simply introduces a bias that reduces the overall number of correct judgments. Conversely, the bottom row is dominated by guessing, as it represents a regime where all primitive strategies are costly and often do not provide enough information to justify their expense. However, in the middle row with a moderate cost for simulation, there is a good mixture of the SP and SWP strategies, especially when the symmetry cost is very low and the weight rule cost is moderately low (the bottom-left quadrants of the panels). Thus there exists a moderately large swath of cognitive cost values for which the most efficient strategies are those that we observe in our participants.

Given that the choice of observed strategies can be explained in a resource-rational framework, we next ask whether the observed proportion of SP and SWP strategies is also consistent with a resource-rational trade-off. Using the same resampling scheme, we assume that only the SP and SWP strategies are available, and calculate what the optimal mixture of these strategies would be to produce the highest average value of computation across all of the proxy balance beam sets. We compare this to the empirically observed mixture of 39% using the SWP strategy (excluding guessing). Figure 12B shows the optimal mixture of these strategies, where the red regions represent $\pm 5\%$ from the empirical value, more yellow values represent higher use of the SWP strategy, and purple/blue regions represent higher use of the SP strategy. In the regions where both the SP and SWP strategies are found to be optimal (the bottom-left quadrants

[9] We included these two together because the weight rule and symmetry judgments are mutually exclusive – if a balance beam has enough of a weight difference to trigger the weight rule, it cannot be symmetric. Thus the optimal ordering of these strategies is difficult to untangle.

of the middle row), we find that the optimal mixture between these two integrated strategies is similar to the mixture we observe across our participants.

Thus, even though we cannot reliably estimate the relative cognitive costs of using these various strategies, we do find a large regime in which people's choice of a mixture between integrated strategies is consistent with a resource-rational selection of strategies.

6 Experiment 4: Generalizing to more complex stimuli

Although the Integration of Simulation and Rules model can capture human stability judgments on simple beams as well as variants in the shape or material of blocks and the size or position of the pivot, to validate and extend this model, we test whether we can explain balance judgments on beams that vary in shape, material, and pivot at the same time, using a model that was fit only on singular variants.

6.1 Experiment

The generalization experiment procedure was nearly identical to the first three experiments, only changing the stimuli used, and presenting extended instructions that described all three balance beam variants. Twenty seven participants were recruited from Mechanical Turk in exchange for $2.00.

The trials consisted of 192 different balance beams, created using the same tools as before. These beams were split evenly between the six beam configuration (balance, weight, distance, and the conflict counterparts). Within each configuration, trials were created such that:

- Half of the trials were made only of blocks, and half were made of a mixture of blocks and shapes.
- Half of the trials were made of a single material, and half were made of items of mixed materials.
- Half of the pivots were at the center of the beam, and half were off-center.

- Half of the pivots were the smallest made in the pivot experiment (2.5% of the beam width) and half were the second largest (10% of the beam width).

Trials were counterbalanced such that there were 12 trials with each of these 16 possible variations, split evenly across configuration types. As in the previous experiments, each participant observed an equal number of trials that fell left, fell right, or balanced; however, trials were normalized for reporting such that all beams that fall should fall left, and the side with more weight in the conflict-balance trials was on the left.

6.2 Results

To demonstrate generalization of the Integration of Simulation and Rules model, we applied the model from before to the predict how the balance beams in this experiment would fall. Because all of the trials here were made of an amalgam of trial types from the prior experiments, no parameter fitting was required, making these out-of-sample predictions.

Overall, we could predict accuracy across trials well ($r = 0.79$, see Figure 13, left), suggesting good generalization. Furthermore, the Integration of Simulation and Rules model generalizes to this out-of-sample experiment better than the traditional rules model ($\Delta LLH = 99.6$).

The ISR model also generalizes well to novel stimuli. We can categorize the trials by how much they deviate from the "standard" beam: a balance beam made of blocks of a single material on a small, centered pivot. In this way, the model has been fit based on balance beams that look like the standard beam, or with any one of the four deviations above, but has never seen any combination of those changes (with the exception of only pivot centering and size). Yet as can been seen in the right panel of Figure 13, the ISR model generalizes to these novel scenarios with 2-4 changes well (albeit slightly worse than the trials it was fit on).

7 Experiment 5: Explaining the torque-difference effect

According to rule-based explanations of balance predictions, one of the most inexplicable findings in the literature is the torque-difference effect: that people can more accurately predict

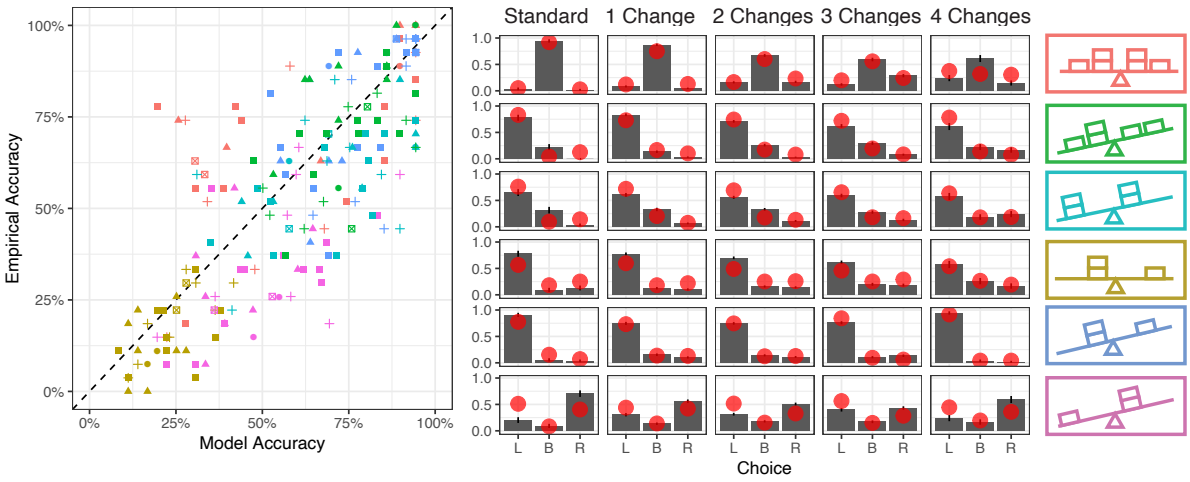

Figure 13. *Left:* Human accuracy accuracy on Experiment 4 versus zero-parameter model predictions. Each point represents a separate trial. Colors indicate the beam configuration based on the legend to the right, while the shapes indicate how much the trial deviates from a standard trial (circle: standard; triangle: 1 change; square: 2 changes; cross: 3 changes; boxes: 4 changes). *Right:* Participants' choices (bars) versus Integration of Simulation and Rules model choices (red dots), split by beam configuration and deviation from standard trial. The Integration of Simulation and Rules model can predict the responses of new participants on novel combinations of trials well.

how a beam will fall when the difference in torque between the two sides is larger, even when those beams would be treated identically within a set of rules (Ferretti & Butterfield, 1986). Some have explained this finding by appealing to "visual heuristics" (Zimmerman & Pretz, 2012) that are activated only when there is a large difference between the sides of the beam (Jansen & van der Maas, 1997); however, it has not been well described how these visual heuristics work or when they should be activated instead of rules.

On the other hand, if people are using a combination of rules and simulation, we would expect that performance should increase *any* time the torque difference increases: even if the use of rules does not change, noisy simulation will be more likely to produce the correct answer when the difference between the sides is larger. To test whether we can explain the torque-difference effect with the ISR model, we replicate Ferretti and Butterfield (1986).

7.1 Experiment

We recruited 21 participants from Mechanical Turk, who were compensated $1.20 for their time. The experimental procedure was identical to that of previous experiments; only the materials were different.

To produce the balance beams for this experiment, we replicated the methodology of Ferretti and Butterfield (1986), although we used twice the number of stimuli. All balance beams were standard beams, comprised of only a single stack of up to six identical blocks on each side of a point pivot. To further replicate Ferretti and Butterfield (1986), the stack of blocks could only be placed at one of six equally spaced distances away from the pivot.

There were 144 stimuli used in this experiment. Of these, eight each were balance or conflict-balance configurations. The remainder of the stimuli were split equally into 32 trials of weight, distance, conflict-weight, and conflict-distance trials. Each of these groups were further split into four torque-difference levels. In all cases, a torque value was calculated for each side of the beam by multiplying the distance position (1–6) by the number of blocks in the stack. For the simple weight and distance trials, the there was only a difference of 1 unit of torque between the sides for torque-difference level 1, a difference of 3 for level 2, 12 for level 3, and between 24 and 30 for level 4. For the conflict-weight and conflict-distance trials, the torque difference for level 1 and 2 were the same – 1 and 3 respectively – but there was a difference of 5 for level 3 and 18-24 for level 4.

As in previous experiments, stimuli were mirrored so that half of the beams that fell would fall to the right and half to the left (however, because there were fewer balance or conflict-balance trials than the rest, there was not an equal split between all three options).

7.2 Results

Because Ferretti and Butterfield (1986) investigated the torque-difference effect in children, we first check that the torque-distance effect can be found in adults. As can be seen in the left panel of Figure 14, there is a clear difference in accuracy by beam classification

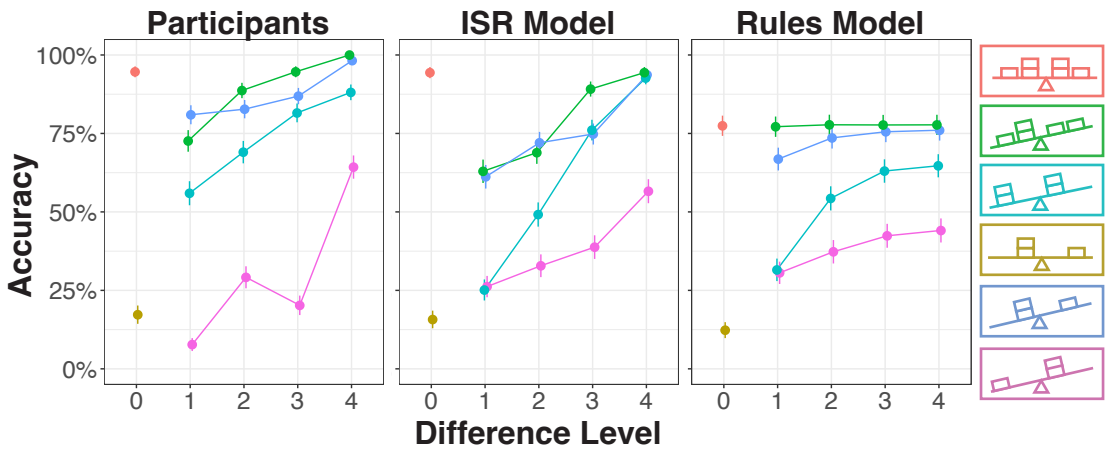

Figure 14. Accuracy on Experiment 5 for participants (*left*), the Integration of Simulation and Rules model (*center*), and the traditional rules model (*right*), grouped by beam classification and torque-difference level (balance and conflict-balance configurations could not have a torque difference so were set at level 0). Both participants and the Integration of Simulation and Rules model demonstrate an increase in accuracy across all beam types as the torque-difference increases, while the rules-only model only does so for the distance trials.

($F(3,112) = 337,\ p \approx 0$), and by the difference level ($F(3,112) = 85,\ p \approx 0$), as well as an interaction between the two factors ($F(9,112) = 8.95,\ p = 4.4 * 10^{-10}$). Although Jansen and van der Maas (1997) suggests that the torque-difference effect is driven by only the most extreme differences, even excluding beams of difference level 4 we find evidence for a difference in accuracy across classification ($F(3,84) = 282,\ p \approx 0$), difference level ($F(2,84) = 30,\ p = 1.6 * 10^{-10}$), as well as an interaction ($F(6,84) = 3.75,\ p = 0.0023$). Thus we find evidence for a torque-difference effect across all difference levels, as expected by the ISR model.

To directly test how well the ISR model explains the torque-difference effect, we use the same model fit on the data from Experiments 1-3 to predict accuracy on this data, and find that it correlates well across trials with empirical accuracy ($r = 0.87$). As can be seen in the middle panel of Figure 14, the model's accuracy generally follows the same trend as human accuracy, increasing with greater torque difference levels. However, the largest deviation between human

and model accuracy is in the simple distance trials with a small torque difference, perhaps because people are paying more attention to the difference in distances when all else is equal.

The traditional rules model with noisy perception, on the other hand, cannot capture the torque difference effect nearly as well (it produces only slight accuracy increases across the torque levels; Figure 14, right) or capture human predictions as well as the ISR model ($\Delta LLH = 80.6$, see Figure 15). While it does expect accuracy to rise with difference level in the simple distance trials because the distances become more perceptually distinct, it underpredicts accuracy at small torque differences just as the Integration of Simulation and Rules model does. Furthermore, it expects little to no increase in accuracy across any of the other beam configurations. Thus we can naturally explain the torque difference effect as a combination of rules and physical simulation, but cannot explain it by rules alone, even for small torque differences.

8 Experiment 6: Shifting the use of rules

A core claim of resource rationality is that people will in general use strategies that provide better expected value on the problems they expect to encounter. Thus, if people are in a situation where they expect the weight rule to be useful, we should expect them to use the weight rule more often; conversely, if people expect to encounter more balance beams where the weight rule will provide the wrong answer, then people should use that rule less. In this experiment, we provide participants with "training" stimuli that either consist of many trials where the weight rule is accurate or where it is inaccurate, and investigate whether this training impacts the use of the weight rule in future balance judgments.

8.1 Experiment

We recruited 48 participants from Mechanical Turk, who were compensated $\$2.00$ for their time. Participants were randomly assigned one of two conditions: the *weight rule accurate* or *weight rule inaccurate* conditions, for 24 participants in each condition. One participant from the inaccurate condition was excluded from analysis because their median time to respond to the test

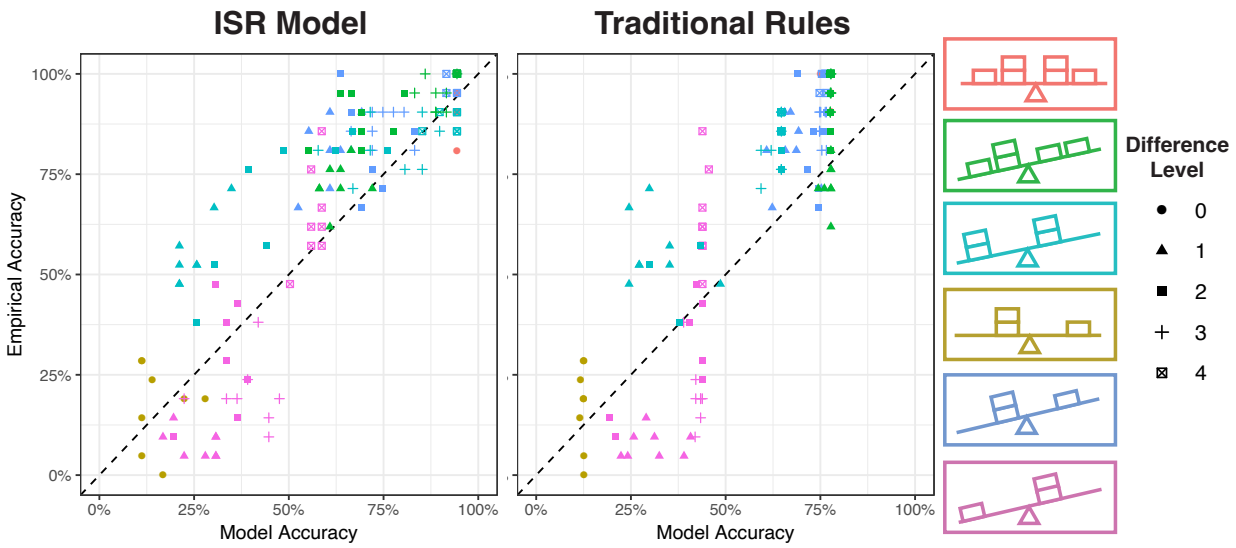

Figure 15. Correlation of empirical accuracy (y-axis) with Integration of Simulation and Rules model accuracy (*left*) and traditional rules model accuracy (*right*). Each point represents a single trial, with the color representing the beam configuration and the shape representing the torque-difference level. Circles indicate the zero-level difference (balanced beams), triangles are a one-level torque difference, squares a two-level, crosses a three-level, and hollow boxes a four-level. The traditional rules model mostly treats all instances of a given beam classification the same, regardless of the torque-difference level, and so cannot capture the behavior of participants, who do not.

stimuli was $220ms$, which was over two standard deviations from the average of all other participants (mean: $2,134ms$, sd: $851ms$), and indicative of "clicking through" the experiment.

The experiment proceeded in two phases: the 'training' phase and the 'test' phase. The training phase consisted of 30 trials that were similar to prior experiments, except that after participants indicated their prediction for how the beam would fall, a movie would play showing the motion of the balance beam and blocks. Each balance beam was constructed only from simple blocks of a single material, and could be one of three types: conflict-weight, conflict-distance, or asymmetric-balance. Conflict-weight and conflict-distance trials are as defined before, but asymmetric-balance trials were created so that there were equal numbers of blocks on each side in configurations that produced equal torques, but were not symmetric; thus the weight rule could

not be applied to these configurations at all.

Participants were all given the same 10 asymmetric-balance trials, but the mixture of the other 20 trials differed by condition: participants in the 'weight accurate' condition were asked to judge 16 conflict-weight trials and 4 conflict-distance trials (so that the weight rule would be accurate 80% of the time), while participants in the 'weight inaccurate' condition judged 4 conflict-weight and 16 conflict-distance trials (so that the weight rule would only be accurate 20% of the time).

The 'test' phase was identical across all participants. Participants were given 150 trials with simple, single-material blocks: 50 conflict-weight, 50 conflict-distance, and 50 conflict-balance. The movies no longer played after participants made their choice to avoid shifting strategies with further feedback.

8.2 Results

In the 'test' phase, we do not find any evidence that overall accuracy differed by training type (weight accurate: 49%, weight inaccurate: 46%; $\chi^2(1) = 1.88,\ p = 0.17$), but do find an interaction between training type and balance beam class ($\chi^2(2) = 204,\ p \approx 0$), suggesting that training affected *which* trials participants got wrong or right.

Because all of the test trials were conflict trials, the weight rule will *always* suggest the side with more weight should fall down, while simulation will produce a distribution of responses (and the symmetry rule should never trigger). Thus we can study whether participants in the 'weight accurate' condition use the weight rule more by testing whether they make more predictions that the weight-side will fall than participants in the 'weight inaccurate' condition. Indeed, we do find that the participants in the 'weight accurate' condition were more likely to choose the side with more weight across all trials (50% vs. 36%; $\chi^2(1) = 5.88,\ p = 0.015$), and for each of the different beam classes (CW: 72% vs. 50%, CD: 31% vs. 18%, CB: 47% vs. 38%; $\chi^2(2) = 11.3,\ p = 0.0035$; Fig. 16). Thus we find evidence that exposing participants to sets of balance beams where the weight rule is more helpful causes them to use the weight rule more

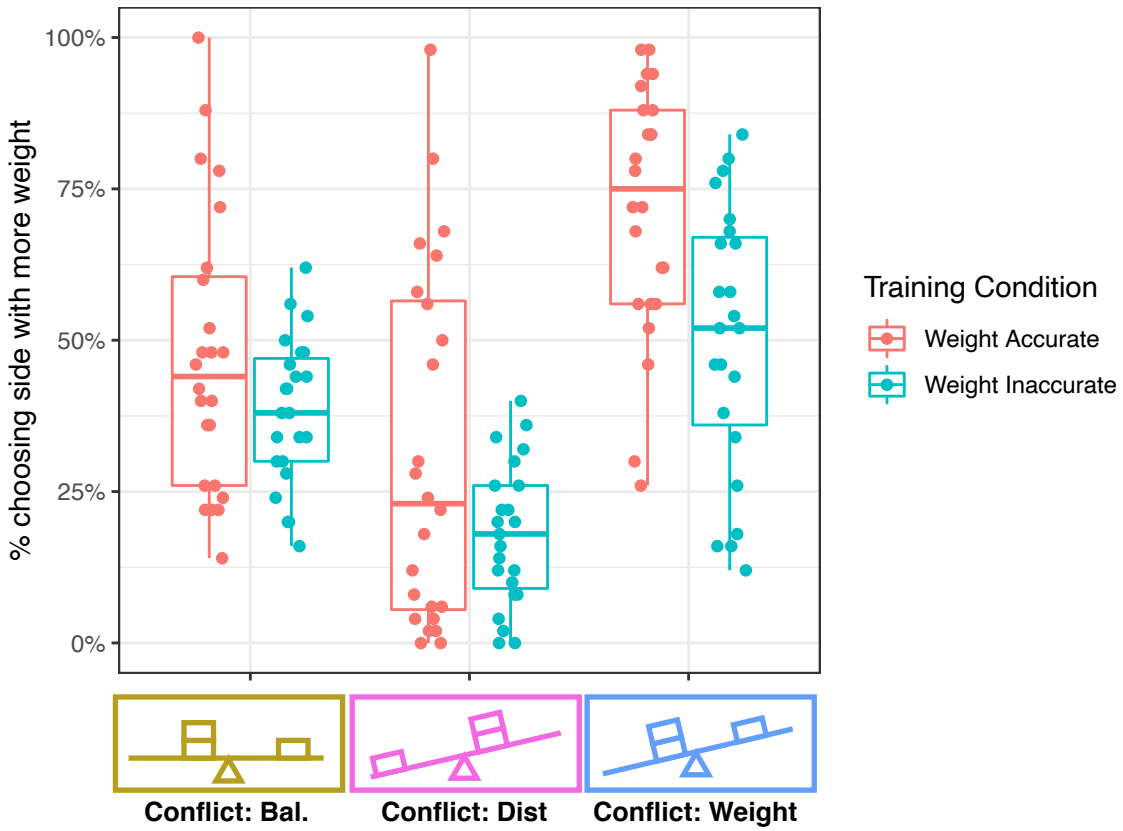

Figure 16. Proportion of trials predicted to fall to the side with more weight, split by beam type (x-axis) and training condition. Each point represents a participant's predictions in that condition, with box plots representing the median participant and interquartile range. Across all beam types, participants in the "weight accurate" training condition were more likely to predict that the side with more weight would fall, suggesting that they are relying more on the weight rule to make their predictions.

often on subsequent problems, as would be expected under a resource-rational framework.

9 Discussion

Here we argue that human physical reasoning is not solely based on a system of rules, nor solely on mental simulations of physics. Instead, people bring both rules and simulation to bear when reasoning about the physical world, and combine them in a way that trades off between accuracy and efficiency.

We studied this trade-off in the domain of judgments of stability – a domain that has been

explained both as based on rules as well as using simulation. Using scenes of balance beams that are more realistic and more varied than the diagrammatic stimuli typically used for these tasks, we found that participants' predictions were not well explained by the system of rules that has historically been used to explain balance beam judgments, nor by pure simulation. Instead, they can be explained by a system that combines both rules and physical simulation as a set of integrated strategies that are selected as a resource-rational trade-off. This framework can naturally explain previous findings like the torque-difference effect (Ferretti & Butterfield, 1986) that have been difficult to reconcile with pure use of rules, as well as why people shift their use of rules in response to different expectations about the problems they are expected to solve.

This provides further evidence that we use different cognitive systems for different types of physical reasoning (Schwartz & Black, 1996b; Kozhevnikov & Hegarty, 2001; Zago & Lacquaniti, 2005; Smith et al., 2018), but extends the previous work with a framework for understanding how these cognitive systems are chosen by treating the selection of these systems as a resource-rational trade-off (Griffiths et al., 2015; Lieder & Griffiths, 2020). Nonetheless, while the ISR framework describes the structure of cognitive systems used to trade off between simulation and rules to solve physical problems, the way that these systems are implemented in the mind requires further study. In the remainder of this discussion, we first consider the commitments made by the ISR framework and different ways the framework might be instantiated, then discuss the structure of the individual simulation and rule-based systems including their relation to more general dual-process theories.

9.1 Implications of the ISR framework

The Integration of Simulation and Rules framework describes the process that people use to combine simulation and rule-based systems for solving physical problems, suggesting that we (1) have access to a set of primitive strategies that we combine into integrated strategies, (2) choose integrated strategies to use in a way consistent with resource rationality, and (3) we apply those integrated strategies to models of the world filtered through the uncertainty of perception (see

Section 2.5 and Fig. 2). This framework commits to the general ordering of information processing and types of representations and systems available in order to implement strategy selection and usage.

However, within this general framework, there are many ways in which the mind could instantiate the various component cognitive systems. Here we discuss three particular considerations that will require further research. First, are the various strategies constructed beforehand and retrieved during the problem solving process, or are they constructed on the fly? Next, within an integrated strategy, are the primitive strategies executed sequentially or in parallel? Finally, how does the mind determine the relevant utilities and costs in order to perform strategy selection?

9.1.1 The timing of strategy formation. The ISR framework contains the assumption that people have access to a set of primitive strategies that they form into integrated strategies, but is agnostic to *when* the strategies are constructed. Here we discuss whether both the primitive and integrated strategies exist prior to their use in the ISR, or whether they are constructed in response to a physical problem.

Primitive strategies are heterogeneous, consisting of (at least) simulation and rule-based approaches, and thus the prior availability might differ depending on the primitive strategy. Simulation, for instance, is considered a general purpose system for physical cognition (Battaglia et al., 2013) that is thought to underlie even infants' earliest reasoning about the physical world (Ullman et al., 2017; Smith et al., 2019; Ullman & Tenenbaum, 2020). Thus we might expect that simulation is a generally available primitive strategy that can be used across a range of physical problems. However, the particular way in which the simulator is used might depend on the problem context. For instance, how far into the future should simulations look to decide if a balance beam will fall? Too long would be wasteful if blocks are predicted to be just sitting on the ground, but too short might cause one to inappropriately decide that a teetering beam will balance. More nuanced uses are possible (for instance, "simulate until either the beam has touched the ground or there has been no motion for X seconds"), but these require using the simulator in a way that is more

tailored to a particular physical problem. So the particular *use* of physical simulation might need to be determined in response to a particular problem; for certain common events (e.g., determining if an object is “falling”) we might have readily available ways of using the simulator, but this will not always be the case.

On the other hand, it is less clear whether the rules that people use have been previously formed, or whether they are constructed on the fly (diSessa, 2014). Historically, knowledge of physics was considered to be derived from flawed “intuitive theories” (McCloskey, 1983). This would suggest that our intuitions about what causes a set of objects to be stable or fall are derived from preexisting knowledge applied to a particular scenario. But others have proposed a theory of “knowledge in pieces” that suggests that we construct explanations from a loose collection of conceptions about the world (some erroneous; diSessa, 1993). Under this theory heuristics like the weight rule would not be prespecified, but instead created on the fly using more primitive bits of knowledge like “heavier things push down harder”. Finally, “framework theories” propose a combination of both of the prior theories, suggesting that we have loose conceptual frameworks for understanding the world that can be chosen from or combined for any particular causal explanation (Vosniadou, 2019); thus some rules might be readily accessible while others might be constructed as needed.

The formation and selection of integrated strategies will generally occur after primitive strategies are available, though again it is unclear whether these are pre-computed or constructed on the fly. For commonplace problem types, it is likely that people have already learned the most appropriate integrated strategy to deploy (e.g., to add two numbers children understand early on that retrieval from memory should be the first primitive strategy to use, followed by a strategy that performs the addition from scratch; Siegler & Shipley, 1995). However, the process of choosing an integrated strategy might be intertwined with primitive strategy formation: if existing strategies cannot provide a reasonable answer, a new primitive strategy would need to be formed (Shrager & Siegler, 1998).

Nonetheless, the precise timing of strategy performing is not crucial within the

ISR framework; it only matters that both the primitive and integrated strategies are available as needed. However, determining exactly how the mind implements this framework will require further study of how and when strategies are formed or recalled.

9.1.2 Sequential choice or parallel systems? In this paper we propose that the selection between cognitive strategies for physical reasoning is itself a *choice*, even if this choice is implicit. Here we consider an alternative framework for explaining these results: that all relevant primitive strategies are automatically activated simultaneously and accumulate evidence until they reach a confidence threshold. Under this framework, the primitive strategy that reaches this threshold first is the one that produces a decision for the particular scenario. This framework is similar to the evidence accumulation models (Ratcliff, Smith, Brown, & McKoon, 2016; N. J. Evans & Wagenmakers, 2019) that have been proposed for simple decision systems where evidence accumulates for separate choices from a single system – for instance, deciding whether to push a button in the (possible) presence of a stop signal (Matzke, Love, & Heathcote, 2017) or determining which of two letters was briefly observed (Ratcliff & Rouder, 2000).

While this framework suggests that all rules and simulation should be activated simultaneously in order to accumulate evidence, there is a choice inherent in the framework: these models often have a free "drift rate" parameter that controls how quickly evidence is accumulated (Ratcliff et al., 2016).[10] If people are making judgments based on the evidence accumulation framework, this rate parameter must be flexibly set; without this flexibility, it would be impossible to explain how people shift their use of the weight rule based on prior experiences (Section 8). And so while the evidence accumulation framework proposes a different process for choosing the cognitive systems to use to make physical judgments, it serves a similar purpose of selecting systems based on their expected utility via the drift rate parameter. The current experiments suggest only that this choice is necessary and that it is made in a way that is equivalent to a resource-rational trade-off, but further work is required to explain precisely *how* the relevant

[10] Alternately, this parameter can be recast as the evidence threshold where faster accumulation is equivalent to a lower threshold, but these formulations are typically indistinguishable (N. J. Evans & Wagenmakers, 2019).

cognitive systems are selected and activated.

Nonetheless, it is possible that the integrated strategies are a combination between automatically activated and selected systems. For instance, the symmetry rule is found to be the first primitive strategy used in both integrated strategies that people use, and symmetry may be perceived pre-attentively (Wolfe & Friedman-Hill, 1992). And so it could be that symmetry is noticed automatically, and thus can be applied with effectively no effort while the decision to use (or not use) the weight rule prior to simulation is a choice.

9.1.3 Deciding between cognitive systems. This work expands upon past research that has suggested that human physical reasoning is based on multiple cognitive systems that often posits "rules of thumb" for choosing between those systems (Kaiser et al., 1992; Schwartz, 1995). We propose that this choice between different systems for physical reasoning can be understood as selecting strategies that are expected to maximize cognitive efficiency: being accurate enough while expending as little cognitive effort as possible. While only tested here in the case of balance beams, this provides a generalizable framework for understanding human physical reasoning, and shares principles that are thought to underlie how people decide how far ahead to plan (Callaway et al., 2018; Ho et al., 2021) or select general cognitive strategies (Lieder & Griffiths, 2017).

This framework unifies the "rules of thumb" that have previously been used to explain the dichotomy between simulation and rules in physical reasoning: all of the manipulations that are expected to induce simulation are those that would make simulation more informative or less costly, thus increasing its utility. For instance, a perceptual-motor task that involves predicting where a ball will land requires fine-grained information about object trajectories that can be extracted from simulation but not from simple rules about ballistic motion, and so will be much more likely to rely on simulation (Smith et al., 2018). Conversely, additional details in a diagram (Schwartz, 1995) or motion information (Kaiser et al., 1992) provide information that might be irrelevant for rules, but is an important part of the scene representation that underlies simulation, and so simulating scenes with some features unknown might be either less informative or harder to do.

This strategy choice framework also comes with an inherent challenge: it requires holding calibrated expectations for how well each strategy will perform in a given scenario, as well as the associated cognitive costs. But it is impossible to precisely know the costs and utilities of using a strategy to solve a problem before actually using that strategy (Russell & Wefald, 1991), which defeats the point of selecting a strategy in the first place. Instead, deciding how to approach a problem has been framed in computer science as selecting heuristics that best approximate these values (Russell & Wefald, 1991; Gershman et al., 2015). Cognitive scientists have studied this value approximation as a learning problem, suggesting that people up-weight strategies that have been successful in the past (Siegler, 1988; Siegler & McGilly, 1989), or directly learn the approximate values of using individual strategies (Lieder & Griffiths, 2017).

Yet it is precisely the challenge of determining the relevant costs and benefits that might explain individual differences in the use of rules versus simulation. While across all participants, the trade-off between strategies that included or did not include the weight rule was in line with what would be expected under a resource-rational trade-off, there was a large amount of heterogeneity in how individual participants decided between strategies: some effectively never used strategies that included the weight rule, while others almost always used it where applicable (see Appendix Section A3.2). It is possible that this is because forming simulations or applying rules is more or less costly for some people, but this could also be due to differences in individuals' estimates of the utility of each strategy. Determining the stability of a balance beam from a static image is not a task that most people perform regularly, and thus individual strategy use choices might reflect differences in value estimates driven by prior experiences. Nonetheless, the fact that strategy choice is sensitive to the statistics of the environment suggests that with further experience, individual value estimates can be refined to better match the true costs and benefits.

Calling this selection process a "choice" also does not imply that this is a conscious decision that people are making. Rather, the claim is that there is some cognitive system that forms and applies the integrated strategy, and that this system performs the selection of the integrated strategy in a way that accounts for the relevant expected utilities and costs of applying that

strategy, as well as the range of problems expected to be encountered. While this choice might be performed implicitly, however, it is subject to conscious control: if people are asked to imagine a physical process before answering a question, they are less likely to make judgments that are thought to result from biased rules (Frick et al., 2005; Schwartz & Black, 1999), suggesting that the use of simulation can be consciously imposed.

9.2 Systems for physical reasoning

9.2.1 Comparisons to dual process theories. Psychologists have long theorized that people use multiple cognitive systems for reasoning about many different domains, from numeric cognition (Feigenson, Dehaene, & Spelke, 2004) to decision making (Kahneman, 2011). In physical reasoning in particular, multiple systems have been proposed due to both behavioral evidence (Schwartz & Black, 1996b; Kozhevnikov & Hegarty, 2001; Smith et al., 2018), as well as evidence from cognitive neuroscience: people recruit different brain areas for making predictions (Fischer et al., 2016) or inferences about mass or stability (Schwettmann, Tenenbaum, & Kanwisher, 2019; Pramod, Cohen, Tenenbaum, & Kanwisher, 2021) than they do for solving word problems that require physical knowledge (Jack et al., 2013; Mason & Just, 2016). Here we consider how these systems fit within broader theories of human cognitive architecture, and why having multiple systems to rely upon for physical reasoning might be desirable.

At first blush, the cognitive systems studied in this paper appear to map onto a common dichotomy found in the psychological literature, that we have two systems for reasoning about the world: a fast, effortless, intuitive system (System 1) and a slower, deliberative system (System 2; Kahneman, 2011; J. S. B. T. Evans, 2008). The physical simulation can be thought of as a ‘System 1’ process, as it is a domain-specific system that operates automatically, even in the absence of a task that would rely on this system (Fischer et al., 2016). Conversely, rules for physical understanding are logical, sequential, and use a style of decision making – decision trees – that have been proposed as the basis of rule-based reasoning outside the domain of physics (e.g., Gigerenzer & Goldstein, 1996).

However, there are key differences between the dual systems proposed here and the typical description of Systems 1 and 2. Often System 1 processes are considered to be 'heuristics' that work quickly based on incomplete information, and the 'analytic' System 2 can override those heuristic judgments with slower but unbiased reasoning (J. S. B. T. Evans, 2006). In the case of physical reasoning, on the other hand, simulation is considered to be relatively well calibrated to real-world physics (Battaglia et al., 2013; Sanborn et al., 2013; Smith et al., 2018), but the weight rule considered here is a heuristic that throws away relevant information about the location of objects on the balance beam. And while System 1 processes are often thought to be rapid (or automatic) and less costly than engaging System 2 processes, in this case it is the analytic rules that seem to accrue lower cognitive costs than simulation.

These differences re-raises the question of why we have multiple systems for physical reasoning. Many theories of System 1 vs. System 2 processing explain this as a difference in cost-benefit trade-offs: System 1 processes provide us with rapid, cognitively cheap, and typically accurate information, but System 2 processes can provide us with the correct answer in situations where System 1 processes fail, albeit at a greater cognitive cost (J. S. B. T. Evans, 2008). Yet if physical rules are incomplete and potentially erroneous while our intuitive simulation system is unbiased, why should we use those rules?

One reason that biased rules might sometimes be preferable over simulation is that our simulations, despite being unbiased, are also noisy and uncertain. Simulation must be robust to variability in the real world, including uncertainty in the perception of objects (Battaglia et al., 2013) as well as variability in the dynamics of object motions as they interact with each other (Smith & Vul, 2013). Thus even a system calibrated to real-world physics will produce variability in responses (Smith & Vul, 2015) as well as potentially biased judgments (Sanborn et al., 2013). One key characteristic of rules, however, is that they are deterministic and invariant to small differences between scenarios that do not cross rule boundaries (Jansen et al., 2007). Thus rules can be helpful to provide us with certainty in cases where simulation can be uncertain, even if they provide the wrong answer in some situations. This benefit can been observed in how participants'

accuracy changed in the 'shapes' experiment from the pure blocks to the beams with non-standard blocks: when the beams were stacked with standard blocks only (making the weight rule easier to apply) participants were significantly more accurate where the weight rule would help (the 'weight' and 'conflict-weight' configurations) because simulation provides less certainty, but this comes at the cost of being *less* accurate in situations where the weight rule would provide a clearly incorrect answer (the 'conflict-distance' configurations; see Figure A1). Thus while simulation can provide us with predictions of how the world might unfold across a wide range of physical scenarios, rules can provide us with additional certainty in specific situations where that certainty is helpful.

9.2.2 The accuracy of physical simulation. Much of the prior research that has found that people use simulation for physical reasoning has suggested that this simulation is based on approximately correct physical principles (Battaglia et al., 2013; Smith et al., 2018; Sanborn et al., 2013; Gerstenberg et al., 2021; Warren, Kim, & Husney, 1987; Deeb, Cesanek, & Domini, 2021), but still may include 'simplifications' to Newtonian physics that can give rise to erroneous predictions, including around balance judgments (Ullman et al., 2017). While an unbiased simulator cannot reasonably produce the biases we observe in people's judgments of balance beams (Marcus & Davis, 2013), a system that inappropriately integrates weight and distance to calculate torque could be engineered to produce similar weight-biases to human judgments.

Yet we argue that assuming a biased simulator cannot explain the full set of data here. Although a biased simulator could reproduce any single set of biased judgments, it cannot explain why people shift their judgments based on previously encountered balance beams (Section 8) – this would require a theory that suggests that people are overwriting the physical knowledge they had a lifetime to learn to become more biased, despite only watching videos with accurate Newtonian physics. And so while characterizing the precise simplifications that the mind makes to perform efficient physical reasoning is an area of outstanding research (Ullman et al., 2017; Li et al., 2022; Bass, Smith, Bonawitz, & Ullman, 2021), it is unlikely that these simplifications can explain people's judgments of how balance beams fall.

9.2.3 The reduced use of rules compared with prior studies. The dominant theory about how people solve balance beam problems for the past half century has been that they use a system of rules, and for good reason: the Rule Assessment Methodology has typically been able classify over 90% of people's judgments (Siegler, 1976), and clustering algorithms group most people together in sets that are expected to make near-deterministic judgments that map on to the expected judgments from those rule sets (Jansen & van der Maas, 1997). Yet despite those successes, the same systems of rules cannot explain how people solved the tasks described in this paper – rules alone cannot explain judgments on the simplest beams that are identical in principle to problems used in prior work (Section 4.4), and the combination of simple rules and simulation that explains human judgments best does not include anything beyond the simplest weight and symmetry rules (e.g., people do not use a 'distance' rule).

These differences could be driven by variations in how stimuli were constructed. In prior work, participants were typically presented with pen-and-paper or computerized diagrams of a balance beam (Ferretti & Butterfield, 1986; Jansen & van der Maas, 2002; van der Maas & Jansen, 2003; Boom et al., 2001), though Siegler (1976) did allow children to play with real balance beams. While we also used computerized scenes, our images of balance beams were developed in a 3-D rendering engine to be as realistic as possible. Less realistic, more diagrammatic displays are more likely to elicit analytic solutions from people (Schwartz, 1995; Schwartz & Black, 1996a). Thus the realism of our scenes might cause people to be less reliant on rules, so those rules alone cannot explain the more complex judgments of more naturalistic stimuli.

Another difference between the stimuli we used and prior work is that previous diagrams of balance beams typically include posts marked on the beam on which the blocks could rest, while the blocks in our experiments were stacked on the beam without these supports. These posts provide more certain information about how far a stack of blocks rests from the center pivot – e.g., a stack on the third post is three times as far as a stack on the first post – whereas this value must be estimated based on the perceptual distance when the struts are not available. This could affect people's judgments in two ways.

First, if the perception of distance is not noisy, then simulation will provide more accurate judgments on the pure ‘distance’ configurations, and if the certainty increases enough, these judgments will be indistinguishable from the prototypical distance rule. This would imply that the “distance rule” is not in fact a rule but instead an epiphenomenon of physical simulation. However, this explanation is unlikely because use of the distance rule has been found to be stable in the past (Jansen & van der Maas, 2002), and children self-report that they are using an explicit rule based on the distance (Siegler, 1976).

Alternatively, the posts themselves could make the distance dimension more salient. Because the posts are spaced at regular intervals and constrain where the blocks can be placed on the beam, people might infer that the dimension that defines those posts – the distance – is important and should be incorporated into any system of rules that they use. This would assume that people do not have a consistent set of rules that are universally applied to all judgments of balance beams, but rather that rules are constructed from more primitive pieces of knowledge about the world (diSessa, 1993; Kloos & Van Orden, 2009). But this would be a simple extension of the reason that young children do not use a distance rule: they do not encode the distance properly (Siegler, 1976), but if given training where the distance dimension is salient, they begin to use this rule (Jansen et al., 2007). Thus we may not observe the distance rule in these experiments because people do not explicitly notice that distance is an important dimension to consider.

A final difference between the stimuli used in our experiments versus prior experiments is that many of our experiments contained balance beam variations for which basic rules are less easy to apply – for instance, if the pivot is off center, people might recognize that the weight of the beam needs to be accounted for but do not have a rule that captures this intuition. This might have led participants to quickly learn that attempting to apply rules would provide less utility in general, since they are more likely to provide no information about the outcome of the scene. As people can adapt their use of various strategies based on learned utilities (Section 8), this could cause a shift away from strategies purely based on rules.

There are therefore a number of differences between the experiments described in this paper that could explain why people are less likely to use rules to make their judgments about the stability of a balance beam. However, each of these differences make the stimuli here more similar to real-world scenarios. Imagine, for instance, a waiter picking up a tray filled with different dishes in one hand and judging whether it will be stable. This is a realistic scene, there are no markings indicating how far from the center each dish is, the dishes might have different sizes or weights, and the waiter might pick up the tray in a point that is not directly in the middle. Thus the attributes of the diagrams that make people more likely to rely on rules are less prevalent in real-world judgments of stability, and so we argue that most day-to-day physical reasoning will be based on a combination of rules and simulation, rather than on rules alone.

9.2.4 Developing biased rules. In this work, and in a large portion of prior research into people's balance beam judgments, it is assumed that from an early age, people have access to rules that they can use to determine how beams will fall, and that the weight rule is most easily accessible (Siegler, 1976; Wilkening & Anderson, 1982; Jansen & van der Maas, 2002; though see also Section 9.1.1). Yet it remains an open question how people develop a rule that is biased despite ample evidence of cases where it does not work, and how this development is so consistent even in young children.

One theory of physical understanding suggests that people hold a large set of 'axioms' or 'proto-knowledge' that can be flexibly combined to create explanations for arbitrary physical systems (Hayes, 1979; diSessa, 1993; Kloos & Van Orden, 2009; Rule, Tenenbaum, & Piantadosi, 2020). This suggests that we do not carry around specific rules for every situation, but instead construct those rules on the fly when we encounter a situation in which they might be needed. Thus a weight rule that applies to (rarely encountered) judgments about balance beams might be developed as a response to being asked to make that judgment.

But then why is it the weight rule in particular that people consistently consider, from children to adults? Some prior research has suggested that young children can only encode one feature at a time, and that weight readily comes to mind as an explanation for tipping things over, given prior

experience (Siegler, 1976); this could explain why 5-6-year-olds appear to rely on the weight rule alone. Yet misconceptions and errors persist even up to adulthood, when people should be able to encode and combine multiple features: even teenagers still do not appropriately integrate weight and distance, mostly being classified as users of Rule III (Siegler, 1976).

Failure to discover the correct rule could be viewed as a generalization of the efficiency trade-off in strategy construction that is studied in this work. If discovering complex rules that appropriately integrate weight and distance is costly or time-consuming, people might stop when they discover 'good enough' rules that explain a majority but not all of their prior experiences or imagined scenarios. Nevertheless, characterizing both the knowledge base and systems for constructing these rules for arbitrary scenes remains an outstanding challenge for future research.

9.3 Conclusion

For decades, research into human physical reasoning has claimed both that it is based on simulation and that it is based on logical rules. Here we argue that this is not a question of which type of reasoning we use, but instead how these different types of reasoning combine to help us understand and interact with the physical world. By viewing this problem as a selection between cognitive strategies that maximizes efficiency, we can explain not just how people make judgments about stability, but perhaps how people construct strategies for understanding the world in general.

10 Acknowledgments

The authors would like to thank Kelsey Allen, Ernest Davis, and Tomer Ullman for their helpful comments on this manuscript.

KAS and JBT were supported by National Science Foundation Science Technology Center Award CCF-1231216, NSF grant 2121009, Office of Naval Research Multidisciplinary University Research Initiative (ONR MURI) N00014-13-1-0333, the DARPA Machine Common Sense program, and a research grant from Mitsubishi Electric.

11 References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310. Doi: 10.1073/pnas.1912341117

Bass, I., Smith, K., Bonawitz, E., & Ullman, T. (2021). *Partial Mental Simulation Explains Fallacies in Physical Reasoning* (Tech. Rep.). PsyArXiv. Doi: 10.31234/osf.io/y4a8x

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(1), 153–166. Doi: http://dx.doi.org/10.1037/0278-7393.15.1.153

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLOS Computational Biology*, *15*(7), e1007210. Doi: 10.1371/journal.pcbi.1007210

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. Doi: 10.1073/pnas.1306572110

Bergen, B. K. (2012). *Louder Than Words: The New Science of How the Mind Makes Meaning*. Basic Books.

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance Classes, strategies, or rules for the Balance Scale Task? *Cognitive Development*, 19.

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. L. (2018). A resource-rational analysis of human planning. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (p. 6).

Community, B. O. (2018). *Blender - a 3D modelling and rendering package.* Stichting Blender Foundation, Amsterdam: Blender Foundation. Retrieved from `http://www.blender.org`

Craik, K. J. W. (1943). *The Nature of Explanation*. CUP Archive.

Davis, E., & Marcus, G. (2015). The Scope and Limits of Simulation in Cognitive Models. *arXiv preprint arXiv: 1506.04956*, 27.

Davis, E., Marcus, G., & Frazier-Logue, N. (2017). Commonsense reasoning about containers using radically incomplete information. *Artificial Intelligence*, *248*, 46–84. Doi: 10.1016/j.artint.2017.03.004

Deeb, A.-R., Cesanek, E., & Domini, F. (2021). Newtonian Predictions Are Integrated With Sensory Information in 3D Motion Perception. *Psychological Science*, *32*(2), 280–291. Doi: 10.1177/0956797620966785

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1), 1–42. Doi: 10.1016/0010-0277(92)90049-N

DeLucia, P. R., & Liddell, G. W. (1998). Cognitive motion extrapolation and cognitive clocking in prediction motion tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 901–914. Doi: http://dx.doi.org/10.1037/0096-1523.24.3.901

diSessa, A. A. (1993). Toward an Epistemology of Physics. *Cognition and Instruction*, *10*(2/3), 105–225. Doi: doi.org/10.1080/07370008.1985.9649008

diSessa, A. A. (2014). A History of Conceptual Change Research: Threads and Fault Lines. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 88–108). Cambridge University Press. Doi: 10.1017/CBO9781139519526.007

The engineering toolbox: Densities of common materials. (2010). `http://www.engineeringtoolbox.com/density-materials-d_1652.html`.

Erev, I., & Barron, G. (2005). On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. *Psychological Review*, *112*(4), 912–931. Doi: http://dx.doi.org/10.1037/0033-295X.112.4.912

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395. Doi: 10.3758/BF03193858

Evans, J. S. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. Doi: 10.1146/annurev.psych.59.103006.093629

Evans, N. J., & Wagenmakers, E.-J. (2019). *Evidence Accumulation Models: Current Limitations*

and Future Directions (Tech. Rep.). PsyArXiv. Doi: 10.31234/osf.io/74df9

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. Doi: 10.1016/j.tics.2004.05.002

Ferretti, R. P., & Butterfield, E. C. (1986). Are Children's Rule-Assessment Classifications Invariant across Instances of Problem Types? *Child Development*, *57*(6), 1419–1428. Doi: 10.2307/1130420

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, *113*(34), E5072–E5081. Doi: 10.1073/pnas.1610344113

Fisher, J. C. (2006). Does Simulation Theory Really Involve Simulation? *Philosophical Psychology*, *19*(4), 417–432. Doi: 10.1080/09515080600726377

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research*, *94*, 62–75. Doi: 10.1016/j.visres.2013.11.004

Forbus, K. D. (1983). Spatial and Qualitative Aspects of Reasoning about Motion. In *AAAI* (Vol. 80, pp. 170–173).

Forbus, K. D., & Gentner, D. (1986). Learning physical domains: {T}owards a theoretical framework. In *Machine learning: An artificial intelligence approach* (Vol. 2, p. 311).

Forbus, K. D., Nielsen, P., & Faltings, B. (1990). Qualitative Kinematics: A framework. In *Readings in Qualitative Reasoning About Physical Systems* (pp. 562–567). Elsevier. Doi: 10.1016/B978-1-4832-1447-4.50057-2

Frick, A., Huber, S., Reips, U.-D., & Krist, H. (2005). Task-Specific Knowledge of the Law of Pendulum Motion in Children and Adults. *Swiss Journal of Psychology*, *64*(2), 103–114. Doi: 10.1024/1421-0185.64.2.103

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501. Doi: 10.1016/S1364-6613(98)01262-5

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245),

273–278. Doi: 10.1126/science.aac6076

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*. Doi: https://doi.org/10.1037/rev0000281

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. Doi: http://dx.doi.org/10.1037/0033-295X.103.4.650

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. Doi: 10.1111/tops.12142

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. Doi: 10.3758/s13428-015-0642-8

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76. Doi: 10.1016/j.cognition.2016.08.012

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.*

Hauf, P., Paulus, M., & Baillargeon, R. (2012). Infants Use Compression Information to Infer Objects' Weights: Examining Cognition, Exploration, and Prospective Action in a Preferential-Reaching Task. *Child Development*, *83*(6), 1978–1995. Doi: 10.1111/j.1467-8624.2012.01824.x

Hayes, P. J. (1979). The Naive Physics Manifesto. In *Expert systems in the micro electronic age* (pp. 242–270). Edinburgh University Press.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285. Doi: 10.1016/j.tics.2004.04.001

Hegarty, M., & Just, M. A. (1993). Constructing Mental Models of Machines from Text and Diagrams. *Journal of Memory and Language*, *32*(6), 717–742. Doi: 10.1006/jmla.1993.1036

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2021). Control of mental representations in human planning. *arXiv:2105.06948 [cs]*. Retrieved from [2021-07-08]http://arxiv.org/abs/2105.06948

Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccia, A. H., & Snyder, A. Z. (2013). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, *66*, 385–401. Doi: 10.1016/j.neuroimage.2012.10.061

Jansen, B. R. J., Raijmakers, M. E. J., & Visser, I. (2007). Rule transition on the balance scale task: a case study in belief change. *Synthese*, *155*(2), 211–236. Doi: 10.1007/s11229-006-9142-9

Jansen, B. R. J., & van der Maas, H. L. J. (1997). Statistical Test of the Rule Assessment Methodology by Latent Class Analysis. *Developmental Review*, *17*(3), 321–357. Doi: 10.1006/drev.1997.0437

Jansen, B. R. J., & Van der Maas, H. L. J. (2001). Evidence for the Phase Transition from Rule I to Rule II on the Balance Scale Task. *Developmental Review*, *21*(4), 450–494. Doi: 10.1006/drev.2001.0530

Jansen, B. R. J., & van der Maas, H. L. J. (2002). The Development of Children's Rule Use on the Balance Scale Task. *Journal of Experimental Child Psychology*, *81*(4), 383–416. Doi: 10.1006/jecp.2002.2664

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, *14*(4), 308–312. Doi: 10.3758/BF03202508

Kaiser, M. K., Proffitt, D. R., Whelan, S. M., & Hecht, H. (1992). Influence of animation on dynamical judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 669–689. Doi: http://dx.doi.org/10.1037/0096-1523.18.3.669

Kerkman, D. D., & Wright, J. C. (1988). An exegesis of two theories of compensation development: Sequential decision theory and information integration theory. *Developmental Review*, *8*(4), 323–360. Doi: 10.1016/0273-2297(88)90013-5

Kloos, H., & Van Orden, G. C. (2009). Soft-Assembled Mechanisms for the Unified Theory. In J. Spencer (Ed.), *Toward a Unified Theory of Development Connectionism and Dynamic System Theory Re-Consider* (pp. 253–267). Oxford University Press. Doi: 10.1093/acprof:oso/9780195300598.003.0012

Kosslyn, S. M., & Ball, T. M. (1978). Visual Images Preserve Metric Spatial Information: Evidence from Studies of Image Scanning. *Journal of Experimental Psychology: Human Perception and Performance*, 14.

Kozhevnikov, M., & Hegarty, M. (2001). Impetus beliefs as default heuristics: Dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review*, *8*(3), 439–453. Doi: 10.3758/BF03196179

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive Physics: Current Research and Controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. Doi: 10.1016/j.tics.2017.06.002

Kubricht, J. R., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). Probabilistic Simulation Predicts Human Performance on Viscous Fluid-Pouring Problem. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society.*

Li, Y., Wang, Y., Boger, T., Smith, K., Gershman, S. J., & Ullman, T. (2022). *An Approximate Representation of Objects Underlies Physical Reasoning* (Tech. Rep.). PsyArXiv. Doi: 10.31234/osf.io/vebu5

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762–794. Doi: http://dx.doi.org/10.1037/rev0000075

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. Doi: 10.1017/S0140525X1900061X

Ludwin-Peery, E., Bramley, N., Davis, E., & Gureckis, T. (2021). Limits on Simulation Approaches in Intuitive Physics. *Cognitive Psychology*, *127*. Doi: 10.31234/osf.io/xhzuc

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken Physics: A Conjunction-Fallacy Effect in Intuitive Physical Reasoning. *Psychological Science*, *31*(12), 1602–1611. Doi: 10.1177/0956797620957610

Marcus, G. F., & Davis, E. (2013). How Robust Are Probabilistic Models of Higher-Level Cognition? *Psychological Science*, *24*(12), 2351–2360. Doi: 10.1177/0956797613495418

Mason, R. A., & Just, M. A. (2016). Neural Representations of Physics Concepts. *Psychological Science*, *27*(6), 904–913. Doi: 10.1177/0956797616641941

Matzke, D., Love, J., & Heathcote, A. (2017). A bayesian approach for estimating the probability of trigger failures in the stop-signal paradigm. *Behavior research methods*, *49*(1), 267–281.

McClelland, J. L. (1988). *Parallel Distributed Processing: Implications for Cognition and Development* (Tech. Rep.). DTIC Document.

McClelland, J. L. (1995). A connectionist perspective on knowledge and development. In *Developing cognitive competence: New approaches to process modeling* (pp. 157–204). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

McCloskey, M. (1983). Intuitive Physics. *Scientific American*, *248*(4), 122–131. Retrieved from [2019-05-31]http://www.jstor.org/stable/24968881

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear Motion in the Absence of External Forces: Naïve Beliefs About the Motion of Objects. *Science*, *210*(4474), 1139–1141. Doi: 10.1126/science.210.4474.1139

Milli, S., Lieder, F., & Griffiths, T. L. (2021). A rational reinterpretation of dual-process theories. *Cognition*, *217*, 104881. Retrieved from [2021-12-07]https://www.researchgate.net/profile/Falk-Lieder/publication/354608618_A_r

Millington, I. (2007). *Game Physics Engine Development*. CRC Press.

Moulton, S. T., & Kosslyn, S. M. (2009). Imagining predictions: mental imagery as mental

emulation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1273–1280. Doi: 10.1098/rstb.2008.0314

Neupärtl, N., Tatai, F., & Rothkopf, C. A. (2020). *Intuitive physical reasoning about objects' masses transfers to a visuomotor decision task consistent with Newtonian physics* (preprint). Animal Behavior and Cognition. Doi: 10.1101/2020.02.14.949164

Normandeau, S., Larivée, S., Roulin, J.-L., & Longeot, F. (1989). The Balance-Scale Dilemma: Either the Subject or the Experimenter Muddles Through. *The Journal of Genetic Psychology*, *150*(3), 237–250. Doi: 10.1080/00221325.1989.9914594

Osiurak, F., & Reynaud, E. (2020). The elephant in the room: What matters cognitively in cumulative technological culture. *Behavioral and Brain Sciences*, *43*, e156. Doi: 10.1017/S0140525X19003236

Paulun, V. C., Schmidt, F., Assen, J. J. R. v., & Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of Vision*, *17*(1), 20–20. Doi: 10.1167/17.1.20

Pramod, R., Cohen, M., Tenenbaum, J., & Kanwisher, N. (2021). *Invariant representation of physical stability in the human brain* (preprint). Neuroscience. Doi: 10.1101/2021.03.19.385641

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990). Understanding wheel dynamics. *Cognitive Psychology*, *22*(3), 342–373. Doi: 10.1016/0010-0285(90)90007-Q

Quinlan, P. T., van der Maas, H. L. J., Jansen, B. R. J., Booij, O., & Rendell, M. (2007). Re-thinking stages of cognitive development: An appraisal of connectionist models of the balance scale task. *Cognition*, *103*(3), 413–459. Doi: 10.1016/j.cognition.2006.02.004

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human perception and performance*, *26*(1), 127.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in cognitive sciences*, *20*(4), 260–281. Doi: 10.1016/j.tics.2016.01.007

Rijn, H. v., Someren, M. v., & Maas, H. v. d. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science*, *27*(2), 227–257. Doi: 10.1207/s15516709cog2702_4

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The Child as Hacker. *Trends in Cognitive Sciences*, *24*(11), 900–915. Doi: 10.1016/j.tics.2020.07.005

Russell, S., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence*, *49*(1-3), 361–395. Doi: 10.1016/0004-3702(91)90015-C

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411–437. Doi: http://dx.doi.org/10.1037/a0031912

Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*, *110*(3), 395–411. Doi: 10.1016/j.cognition.2008.11.017

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. Doi: 10.3758/BF03200924

Schwartz, D. L., & Black, J. B. (1996a). Analog Imagery in Mental Model Reasoning: Depictive Models. *Cognitive Psychology*, *30*(2), 154–219. Doi: 10.1006/cogp.1996.0006

Schwartz, D. L., & Black, J. B. (1996b). Shuttling Between Depictive Models and Abstract Rules: Induction and Fallback. *Cognitive Science*, *20*(4), 457–497. Doi: 10.1207/s15516709cog2004_1

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.

Schwettmann, S., Tenenbaum, J. B., & Kanwisher, N. (2019). Invariant representations of mass in the human brain. *eLife*, *8*, e46619. Doi: 10.7554/eLife.46619

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*,

171(3972), 701–703. Retrieved from [2019-07-09]http://www.jstor.org/stable/1731476

Shrager, J., & Siegler, R. S. (1998). SCADS: A Model of Children's Strategy Choices and Strategy Discoveries. *Psychological Science*, *9*(5), 405–410. Doi: 10.1111/1467-9280.00076

Shultz, T. R., Schmidt, W. C., Buckingham, D., & Mareschal, D. (1995). Modeling cognitive development with a generative connectionist algorithm. In *Developing cognitive competence: New approaches to process modeling* (pp. 205–261). Retrieved from [2020-03-24]http://ego.psych.mcgill.ca/perpg/fac/shultz/personal/Recent_Publications_f

Siegler, R. S. (1976). Three Aspects of Cognitive Development. *Cognitive Psychology*, *8*, 481–520.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, *116*(3), 250–264. Doi: 10.1037/0096-3445.116.3.250

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*(3), 258–275. Doi: http://dx.doi.org/10.1037/0096-3445.117.3.258

Siegler, R. S. (1996). Unidimensional thinking, multidimensional thinking, and characteristic tendencies of thought. In *The five to seven year shift: The age of reason and responsibility* (pp. 63–84). Chicago: University of Chicago Press.

Siegler, R. S. (1999). Strategic development. *Trends in Cognitive Sciences*, *3*(11), 430–435. Doi: 10.1016/S1364-6613(99)01372-8

Siegler, R. S., & McGilly, K. (1989). Strategy choices in children's time-telling. In *Time and human cognition: A life-span perspective* (pp. 185–218). Oxford, England: North-Holland. Doi: 10.1016/S0166-4115(08)61042-0

Siegler, R. S., & Shipley, C. (1995). Variation, Selection, and Cognitive Change. In *Developing cognitive competence: New approaches to process modeling* (pp. 31–76). Retrieved from

[2023-06-30]https://www.tc.columbia.edu/faculty/siegler/sieglershipley95.pdf

Siegler, R. S., Strauss, S., & Levin, I. (1981). Developmental Sequences within and between Concepts. *Monographs of the Society for Research in Child Development*, *46*(2), 1. Doi: 10.2307/1165995

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99. Doi: 10.2307/1884852

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different Physical Intuitions Exist Between Tasks, Not Domains. *Computational Brain & Behavior*, *1*(2), 101–118. Doi: 10.1007/s42113-018-0007-3

Smith, K. A., Mei, L., Yao, S., Wu, J., Spelke, E. S., Tenenbaum, J. B., & Ullman, T. D. (2019). Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Object Representations. In *33rd Conference on Neural Information Processing Systems.* Vancouver, Canada.

Smith, K. A., & Vul, E. (2013). Sources of Uncertainty in Intuitive Physics. *Topics in Cognitive Science*, *5*(1), 185–199. Doi: 10.1111/tops.12009

Smith, K. A., & Vul, E. (2015). Prospective uncertainty: The range of possible futures in physical predictions. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.*

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, *37*(3), 332–341. Retrieved from [2020-03-24]https://www.jhuapl.edu/spsa/pdf-spsa/spall_tac92.pdf

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585. Doi: 10.1038/nn1669

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. Doi: 10.1016/j.tics.2017.05.012

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian Models of Conceptual Development: Learning

as Building Models of the World. *Annual Review of Developmental Psychology*, *2*(1), 533–558. Doi: 10.1146/annurev-devpsych-121318-084833

van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*(2), 141–177. Doi: 10.1016/S0022-0965(03)00058-4

van Maanen, L., Been, P., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In *Mathematical psychology in progress* (pp. 267–287). Springer.

Vasta, R., & Liben, L. S. (1996). The Water-Level Task: An Intriguing Puzzle. *Current Directions in Psychological Science*, *5*(6), 171–177. Doi: 10.1111/1467-8721.ep11512379

Vosniadou, S. (2019). The Development of Students' Understanding of Science. *Frontiers in Education*, *4*. Retrieved from [2022-11-04]https://www.frontiersin.org/articles/10.3389/feduc.2019.00032

Warren, W. H., Kim, E. E., & Husney, R. (1987). The Way the Ball Bounces: Visual and Auditory Perception of Elasticity and Control of the Bounce Pass. *Perception*, *16*(3), 309–336. Doi: 10.1068/p160309

Wilkening, F., & Anderson, N. H. (1982). Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structure. *Psychological Bulletin*, *92*(1), 215–237. Doi: http://dx.doi.org/10.1037/0033-2909.92.1.215

Wolfe, J. M., & Friedman-Hill, S. R. (1992). On the Role of Symmetry in Visual Search. *Psychological Science*, *3*(3), 194–198. Doi: 10.1111/j.1467-9280.1992.tb00026.x

Yildirim, I., Smith, K. A., Belledonne, M., Wu, J., & Tenenbaum, J. B. (2018). Neurocomputational Modeling of Human Physical Scene Understanding. In *2018 Conference on Cognitive Computational Neuroscience.* Philadelphia, Pennsylvania, USA: Cognitive Computational Neuroscience. Doi: 10.32470/CCN.2018.1091-0

Yildirim, I., Wu, J., Kanwisher, N., & Tenenbaum, J. (2019). An integrative computational architecture for object-driven cortex. *Current Opinion in Neurobiology*, *55*, 73–81. Doi: 10.1016/j.conb.2019.01.010

Zago, M., & Lacquaniti, F. (2005). Cognitive, perceptual and action-oriented representations of falling objects. *Neuropsychologia*, *43*(2), 178–188. Doi: 10.1016/j.neuropsychologia.2004.11.005

Zhang, R., Wu, J., Zhang, C., Freeman, W. T., & Tenenbaum, J. B. (2016). A Comparative Evaluation of Approximate Probabilistic Simulation and Deep Neural Networks as Accounts of Human Physical Scene Understanding. *arXiv:1605.01138 [cs, q-bio]*. Retrieved from [2019-05-21]http://arxiv.org/abs/1605.01138

Zhou, L., Smith, K., Tenenbaum, J., & Gerstenberg, T. (2022). Mental Jenga: A counterfactual simulation model of physical support. *Journal of Experimental Psychology: General*. Doi: 10.31234/osf.io/4a5uh

Zimmerman, C., & Pretz, J. (2012). The interaction of implicit versus explicit processing and problem difficulty in a scientific discovery task. In *Psychology of Science: Implicit and Explicit Processes.* Oxford University Press.

Appendix

A1 Additional behavioral results

Here we report additional details on the behavioral results from the first three experiments. While these results were mainly designed to be analyzed based on models of human predictions, analysis of participants' patterns of accuracy and specific choices can provide further evidence that they are not using a set of rules that are invariant across all instances under which they are expected to be invariant.

A1.1 Experiment 1: Shapes

In the first experiment, we tested how judgments of balance would be affected if the balanced objects were made up from irregular shapes that would make weight judgments less certain. If participants were using rules alone to make judgments about these balance beams, then as estimates of weights become more uncertain we would expect that participants would find it more difficult to apply a rule that relied on weights, and therefore judgments should become more noisy, but should always become closer to chance (33% accuracy).

We find that the block vs. arbitrary shape distinction does affect participants' predictions (see Figure A1, Left). Accuracy differed according to the beam classification ($F(5,126) = 48.38, p \approx 0$) and the use of shapes ($F(2,126) = 3.27, p = 0.041$). Most importantly though, the change in accuracy across shape types was modulated by the beam classification type ($F(10,126) = 5.60, p = 6.9 * 10^{-7}$), suggesting that making weight more difficult to judge has a different impact across the various balance beam types.

As the blocks become more irregular shapes and it became harder to judge their weight, there is a significant drop in accuracy in the balance trials (red), the weight trials (green), and a lesser drop in the accuracy of the conflict-weight trials (dark blue). These trials are ones in which simple rules that account for symmetry or weight comparisons provide the correct answer – and so noise in the weights makes judging symmetry or comparing weights more difficult, leading to a decrease in accuracy. However, there is a drastic increase in the accuracy in the conflict-distance

trials (pink), where weight comparison rules lead people astray. This cannot be explained by the use of rules alone.

Thus as rules become more difficult to apply, people appear to rely less on those rules, which hinders performance where those rules are beneficial, but conversely helps performance in cases where heuristic rules are incorrect.

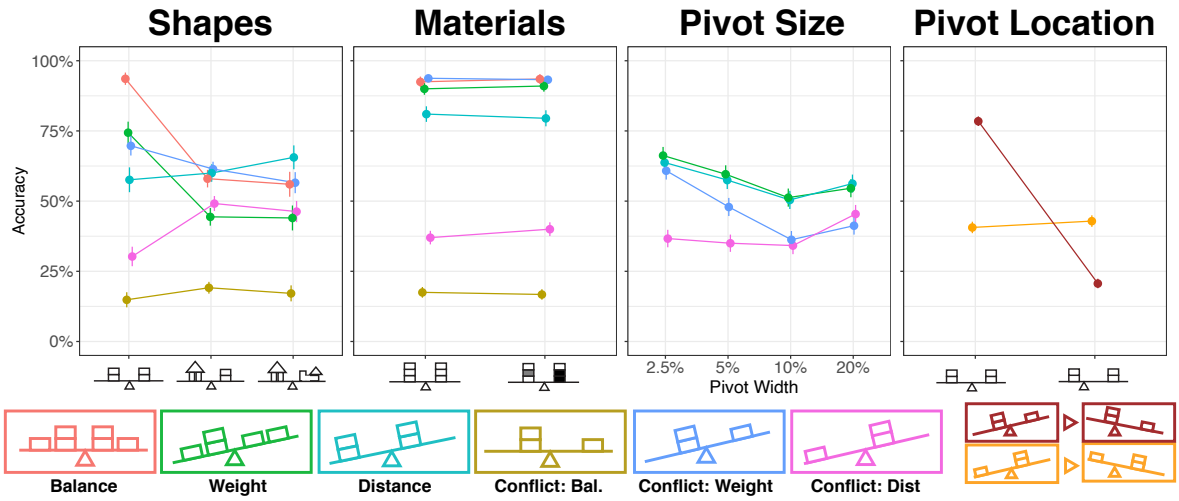

Figure A1. Empirical accuracy across experiments by trial type. Each panel represents a different classification of beam types from the three experiments, splitting Experiment 3 trials into the two subtypes. Colors represent the Sielger beam classifications for trials – from left to right they represent balance, conflict-balance, weight, distance, conflict-weight, and conflict-distance trials. The pivot location trials had a separate classification, with dark red representing trials that were conflict-weight when the pivot was centered and conflict-distance when uncentered, and vice versa for orange. The x-axis splits trials by the relevant dimension being tested in each experiment. The y-axis is the average accuracy across all trials in that classification. Bars represent 95% confidence intervals on the estimated mean accuracy.

A1.2 Experiment 2: Materials

In the second experiment, we tested how a density affected participants' judgments of balance beams. It is possible that rule-based judgments about balance beams could either account for the different weights of blocks of different materials, or could work solely on numeric

properties of the scene (e.g., counting blocks). However, physical simulation does use the different densities and masses of objects to model the world (Hamrick et al., 2016; Yildirim et al., 2018; Schwettmann et al., 2019), so if people do not account for the different materials it would suggest that they are using a mental process other than the intuitive physics engine.[1]

While there was a large difference in accuracy across the beam configuration types ($F(5, 132) = 224, p \approx 0$), there was no evidence of a difference depending on whether the blocks were all the same or multiple materials ($F(1, 132) = 0.05, p = 0.82$), nor was there an interaction between the two ($F(5, 132) = 0.13, p = 0.98$; see Figure A1, Middle-Left). This suggests that participants were accounting for differences in weight between the blocks, and that their estimation of the relative densities was calibrated to the actual densities used in this experiment (likely because this information was provided in the instructions).

A1.3 Experiment 3: Pivots

The pivot experiment was designed to test whether people appropriately account for the way the pivot supports the balance beam. Because changes in the size or location of the pivot will sometimes have effects on the way the beam actually falls, there is not an obvious way that we should expect these changes will affect participants' accuracy (see Figure A1, middle-right and right). Instead, we can investigate the particular ways that participants react to these changes.

If participants understood that wider pivots provided more support and therefore were more likely to balance, we should expect them to respond 'balance' more often with wider pivots, and indeed we find this to be true ($F(3, 152) = 17.36, p = 9.6 * 10^{-10}$, Fig. A2A). However, if participants did not think the beam would balance, we would not expect the size of the pivot to change whether participants believed the beam would fall left or right, and participants did not ($F(3, 152) = 0.11, p = 0.95$, Fig. A2B).

We also tested whether participants incorporated the weight of the balance beam into their judgments. If the position of the pivot is shifted along with the blocks configuration, the only thing

[1] Also see Appendix Section A4 for an additional experiment that directly tested material versus pure shape judgments.

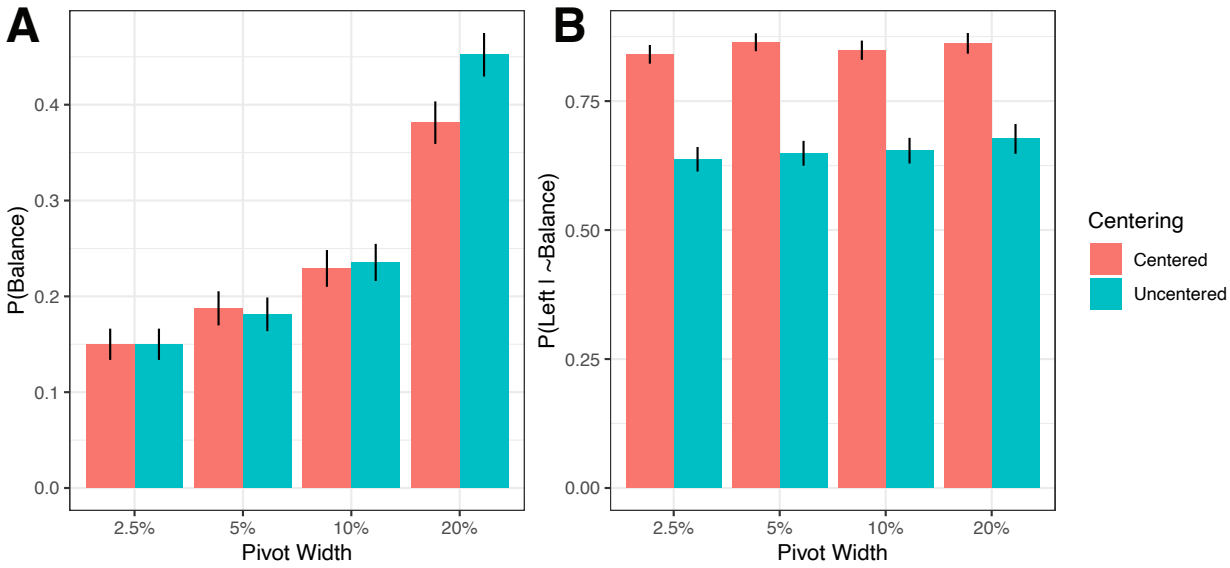

Figure A2. How balance and fall judgments change over pivot sizes in Experiment 3. *A:* As the pivot becomes wider, participants are more likely to believe that a beam will balance. *B:* However, if participants do not believe the beam will balance, their judgment of which direction the beam will fall is unaffected by pivot size.

that will change is that the beam itself will contribute part of its weight to destabilize the balance in the opposite direction of the shift. We find that our participants do act in this manner: their predictions that a beam will fall opposite to the direction of the pivot location increased by 10.1% as compared to the unshifted trials ($t(23) = 4.54, p = 0.00015$), suggesting they do incorporate the beam into their balance judgments.

A2 Model details

Here we provide further details on the structure of the Integration of Simulation and Rules model, including both how the perceptual system gives rise to a scene representation, and how the individual primitive strategies function.

A2.1 Noisy Perception

The noisy perception module was designed to capture how people might take the visual image of a balance beam in each of the experiments and translate it into a mental representation useful for making predictions. While we believe that the internal representation that people hold is a 3-D model from which we can simulate the future (Battaglia et al., 2013) or extract information about the weight of individual objects (Hamrick et al., 2016), for computational efficiency this model uses a minimal equivalent representation: each stack of blocks or shapes is assumed to be a point mass at a its center point on the beam, the beam itself has a weight, and the pivot is represented by its position and width.[2]

Therefore, for a basic trial that consists of just stacks of uniform blocks with the pivot centered, for each stack, noisy perception would encode the number of blocks as the weight of the stack and would place the stack at a position drawn from a Gaussian distribution centered around the actual position with a standard deviation of σ_d.[3] The pivot was typically assumed to be centered but could provide support to beams along a constant width of $width_{pivot}$, so that any system of the beam and blocks with a center of mass the rested within this width would be predicted to balance in simulation. The weight of the beam itself was irrelevant when the pivot was centered, and so was not taken into account for basic beams.

However, because the beams differed across the three experiments, the way that the stimuli that deviated from the most basic type were encoded also differed:

Experiment 1: Shapes The only difference between the basic balance beams and the ones

---

[2] Even if people see the pivot ending at a near point, modeling a pivot with some width could capture uncertainty in simulation about whether the center of mass of the rest of the beam would like close enough to the pivot that it would not fall.

[3] We investigated whether the perception of weight of each stack of pure blocks was uncertain as well, but the parameter fits for the noise in weight perception were indistinguishable from zero and did not increase model fit. This suggests that people do encode all blocks as identical, which could in part explain why prior experiments could explain behavior well without any perceptual uncertainty.

with shapes is that the stacks of oddly shaped blocks were not as easy to judge the weight of as a set of identical wooden blocks. The noisy perception module therefore encoded their weight probabilistically, as the true weight multiplied by a log-normal distribution $ln\ \mathcal{N}(0, \sigma_{shape})$.

Experiment 2: Materials The materials stimuli were perceived in the same way as the basic beams, except the weight of each of the blocks was judged to be proportional to its actual density. While we tested for bias and uncertainty in the relative densities of the materials, adding this flexibility did not increase the model's explanatory power at all. Because the material densities were different enough that small errors in their assessment would not materially affect judgments, and because participants were explicitly notified of the relative densities of the materials in the introduction, we believe that participants' perception was calibrated to the densities of the stimuli well enough that bias and uncertainty were not required.

Experiment 3: Pivot For the pivot stimuli, the stacks of blocks were perceived identically to the basic stimuli. However, the perception of the pivot and beam differed in two ways. First, because the weight of the beam can affect the stability when the pivot is off center, the model perceives the total weight of the beam noisily as a draw from an exponential distribution with rate parameter $weight_{beam}$. Second, because the width of the pivot can affect how unbalanced a beam can be supported, the effective width of support differed across all four different pivot widths, for four support parameters: $width_{2.5}$, $width_5$, $width_{10}$, and $width_{20}$.

A2.2 Primitive strategies

Symmetry judgments. This primitive strategy tests whether the structures on both sides of the pivot are identical and symmetrically placed, and judges the beam to 'balance' if so. Symmetry is thought to be an easy feature to detect in scenes (Wolfe & Friedman-Hill, 1992), and thus is a good candidate for one way people might quickly judge whether objects will balance.[4] Because perception is noisy, this requires testing whether all object stacks are 'close enough' to be

[4] While the symmetry rule has not been directly proposed in the prior literature, it falls out naturally: if the weight is judged the same on both sides, and then the distance is judged the same, the beam is predicted to balance.

considered the same: for every stack of objects on one side, there must be a stack on the other side of approximately the same weight (such that the difference is weights is below some 'just noticeable difference' constant, w_{jnd}), and approximately the same distance from the pivot (such that the difference in distance is below a constant, d_{jnd}). If all blocks are matched, then this rule predicts 'balance', but is undecidable if there are any mismatches.

Weight rule. The weight rule is an implementation of the rule proposed by Siegler (1976): the weights on both sides of the pivot are summed and compared, mostly irrespective of the position of the objects on each side of the beam.[5] Similar to the symmetry judgment, the weight rule accounts for uncertainty in its representation by assuming that one side is heavier only if there is a noticeable difference in weight between the two sides, using the same difference constant (w_{jnd}). If a side does have noticeably more weight, this rule predicts that the beam will fall in the direction of that side; otherwise this rule is undecidable.

Distance rule. The distance rule is also an implementation of the rule from Siegler (1976): the side with objects further from the pivot is assumed to fall. This is implemented by comparing the objects with the greatest distance from the pivot on each side, and testing whether the difference in that distance is greater than the same noticeable distance used in the symmetry judgment (d_{jnd}). If there is a difference, this rule predicts the beam will fall towards the side with more distance, but otherwise this rule is undecidable.

Physical simulation. The simulation module is assumed to operate in a similar fashion to the Intuitive Physics Engine suggested by Battaglia et al. (2013): people run forward their internal representation of the world and use that future world to determine whether the beam will balance, or which direction the beam will fall. Because physical simulation is stochastic (Smith & Vul, 2013), we also assume that the way the physics engine resolves torque to calculate balance is noisy.

For computational simplicity, this process was implemented in the model as a numeric

[5] The only cases where distance is considered in this rule is in scenes with wide pivots. In these cases, if a stack rests so that its center of mass is above the pivot, it is assumed to be supported by that pivot and so its weight is not counted towards either side.

calculation rather than using a computer physics engine. This was accomplished by first calculating the center of mass of the combination of the beam and all object stacks (assuming the center of the beam is at $dist_{beam} = 0$), then perturbing this calculation with Gaussian noise with standard deviation σ_{CoM}:

$$CoM \sim \mathcal{N}\left(\frac{\sum weight_s * dist_s}{weight_{beam} + \sum weight_s}, \sigma_{CoM}\right) \tag{A1}$$

If this center of mass rests over the effective support width of the pivot, the physics module deems the beam to be balanced. Otherwise, it deems the beam to fall to the left or right, depending on where the center of mass rests with respect to the pivot.

A2.3 Model fitting

To obtain a probabilistic distribution over predictions for a given trial, this process was repeated 500 times each trial and the model predictions were tallied.

The model parameters listed above were fit across all three experiments at the same time.[6] Because of the stochastic nature of this model, parameter estimation was performed using the simultaneous perturbation stochastic approximation (SPSA) technique for stochastic gradient descent (Spall, 1992).

A3 Additional model results

A3.1 Additional model variants

Section 5.2 presented analyses showing that, compared to the baseline model that included only the SP and SWP strategies, any model variants that allowed for additional strategies failed to improve fits to human performance, whereas any variants that removed or replaced strategies caused the model to fit human performance worse. Here we expand this analysis by comparing the baseline model to all models that allow one additional strategy.

[6] These did not vary appreciably if they were fit to each experiment individually.

As can be seen in Figure A3, there is no individual strategy that reliably improves the model's ability to explain human performance, which, together with the analyses in Section 5.2, suggests that peoples' behavior can be sufficiently explained with only the symmetry->physical simulation and symmetry->weight->physical simulation strategies.

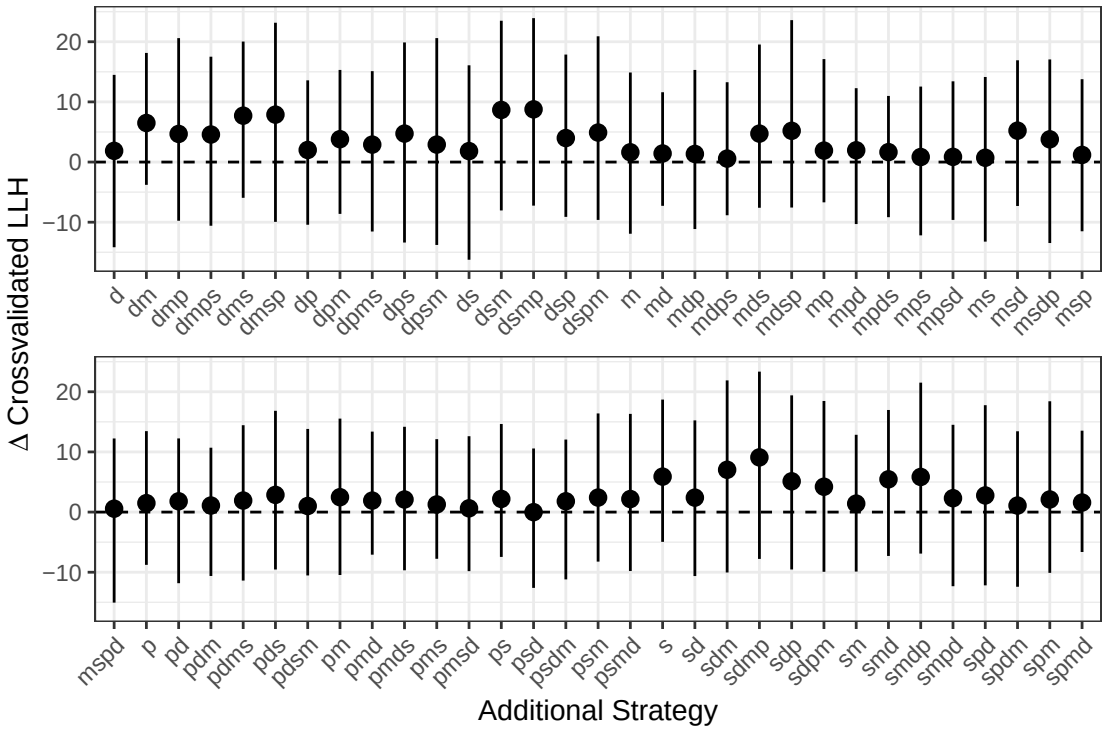

Figure A3. Differences in cross-validated log-likelihood for model variants versus the baseline model. Each model includes one strategy in addition to the baseline SP/SWP. Points indicate mean difference in log-likelihood over 50 samples, bars indicate the 10th to 90th quantiles. The dotted line indicates parity with the baseline model.

A3.2 Individual differences in strategies vs. perception / dynamics

In Section 5.4, we assumed that the mixtures of strategies or rules could vary across participants, while all other parameters were shared. Here we show that the majority of individual differences are in fact in the strategies that people use, rather than in uncertainty about perception or dynamics.

Using the same cross-validation strategy as in Section 5.4 – splitting the trials into two equal sets, optimizing parameters on one set, and comparing the likelihood on the held out set, then repeating 50 times – we compare model fits assuming that all participants can be explained by a single set of parameters with three variants: (1) allowing all parameters to vary per individual, (2) allowing only parameters defining the mixture of strategies (SP/SWP/Guess) to vary by participant, but sharing all perceptual and dynamic parameters, and (3) allowing the perception and dynamics parameters to vary by individual, but assuming consistent strategy usage across participants.

As can be seen in Figure A4, the ISR's explanatory power is improved by allowing all parameters to vary by participant ($\Delta LLH = 427$, 90% CI$= [382, 466]$), individual variation in strategy usage explains about 80% of this difference ($\Delta LLH = 338$, 90% CI$= [293, 386]$), while individual variability in perception and dynamics explains much less ($\Delta LLH = 108$, 90% CI$= [83, 132]$). Thus while there may be some individual differences in the the way that people perceive, judge, and simulate these balance beams, the biggest axis of individual variation is in how they weight each strategy.

A3.3 Rationality analyses including distance

In Section 5.5, we showed that of the three primitive strategies that people seemed to use (symmetry, weight, and simulation), under a range of assumptions about the cognitive costs of those primitive strategies – in which simulation is moderately costly and the symmetry and weight rules are relatively cheap – the integrated strategies that we observed people to use are the ones that provide the highest value of computation, in line with resource rationality.

Figure 12 showed this in aggregate, but to compare strategies we defined a "reasonable strategy to use" as one that provided the highest value of computation in at least 10% of resampled trial sets. Figure A5 shows more detail for this analysis, splitting the integrated strategies into the same groups used in Figure 12, but splitting these strategies out into individual panels to show the proportion of the time each strategy group dominates. This figure shows that in most individual cost regimes, when one strategy dominates it tends to be best across the majority

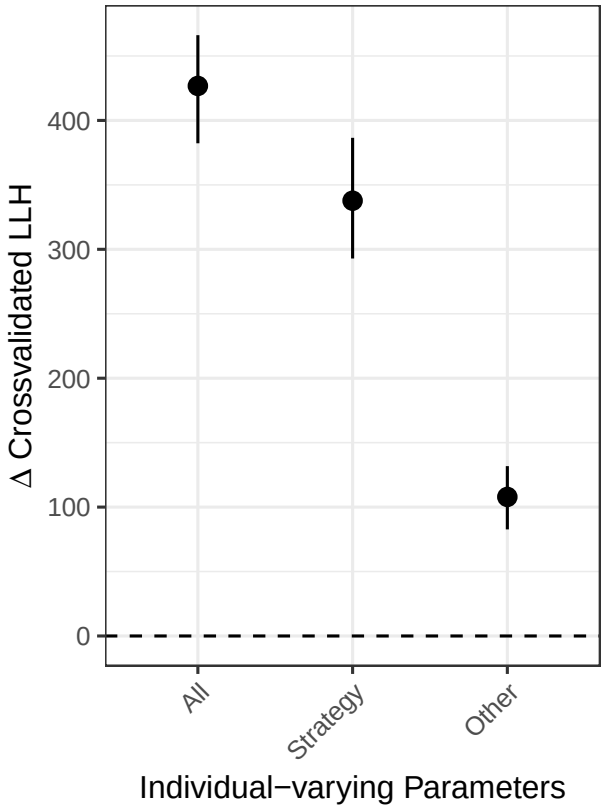

Figure A4. Comparison of individually fitting model parameters versus setting all parameters to be the same across all participants. Points represent difference in average cross-validated log-likelihood versus a model that assumes the same parameters for all participants, with bars representing 90% CI across 50 samples. Individual parameter fits included either all model parameters (*All*), parameters relating to the choice between SP/SWP/Guess strategies (*Strategy*), or all parameters describing perceptual uncertainty, simulation uncertainty, and rule thresholds (*Other*). Allowing all parameters to vary by individual provides the best explanatory power, but most of the improvement is driven by the strategy parameters (2 parameters per participant) and not the rest (9 parameters per participant).

of all trial sets. Nonetheless, when simulation and the weight rule are moderately costly and the symmetry rule is cheap, the SP rule is best; however, if simulation is moderately costly but both the symmetry and weight rules are cheap, then the SWP (or WSP) rule will dominate.

However, in the main paper we limited our analysis to just the symmetry, weight, and simulation primitive strategies, since our participants did not appear to notice that the distance rule

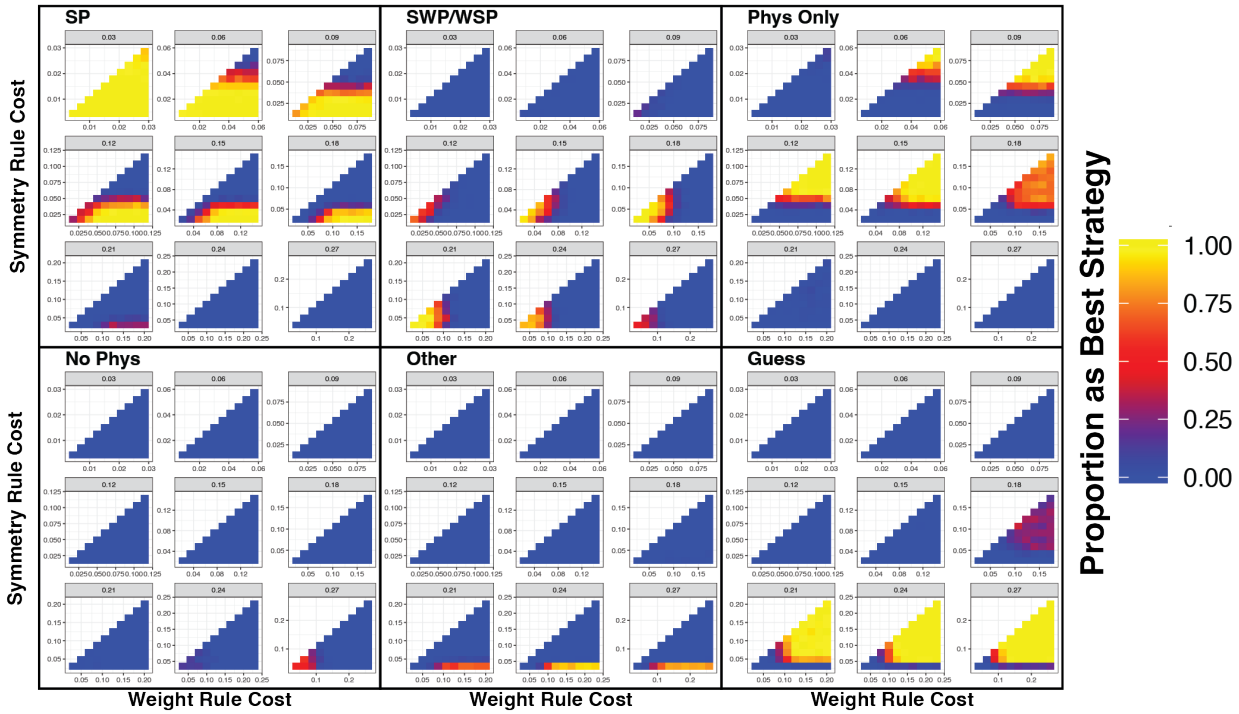

Figure A5. Proportion of resampled trial sets for which a strategy group provides the highest value of computation against all other strategies consisting of the symmetry, weight, and simulation primitive strategies. Each panel represents a separate strategy group, each structured like Figure 12: the sub-panels represent different cost assumptions for simulation, while the x- and y-axes represent different cost assumptions for the weight rule and symmetry rule respectively.

was an option. Here we further analyze the rationality of each strategy if we assume that the distance rule can be used. Again splitting integrated strategies into different groups in which the outcome will be equivalent (since the symmetry rule cannot trigger if either the weight or distance rule does), we show the proportion of the time each strategy dominates in Figure A6. For simplicity sake we assume that the weight rule and distance rule will always have identical cognitive costs, since both rules are of approximately equal complexity.

Similar to the analysis excluding the distance rule, we find that the SP strategy dominates when simulation is moderately expensive (the middle row), the symmetry rule is cheap (bottom of each subfigure), and both the weight and distance rules are moderately expensive (to the right of

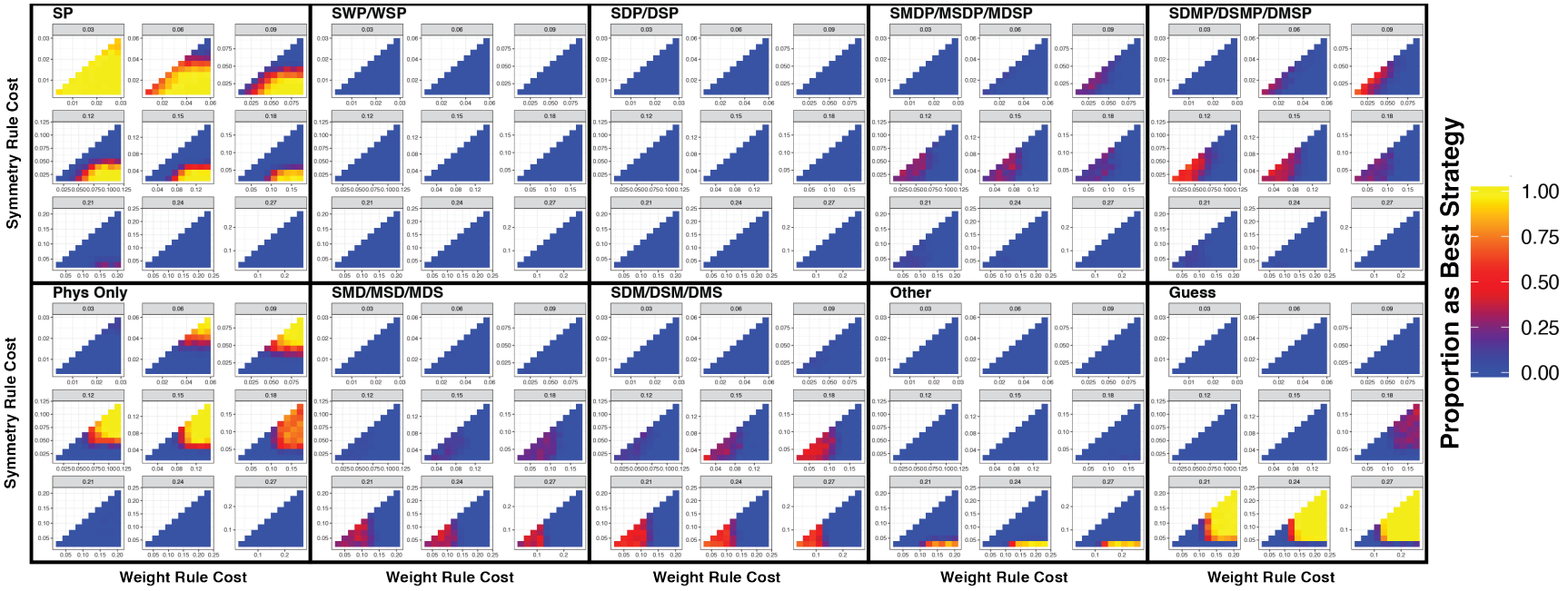

Figure A6. Proportion of resampled trial sets for which a strategy group provides the highest value of computation against all other strategies consisting of all possible sub-strategies. Each panel represents a separate integrated strategy group, each structured like Figure 12: the sub-panels represent different cost assumptions for simulation, while the x- and y-axes represent different cost assumptions for the weight rule and symmetry rule respectively.

each subfigure). However, when the weight and distance rules come less costly (left of each subfigure), we find two main differences from the analysis without the distance rule. First, in many cases it is optimal to include the distance rule as part of an integrated strategy as well as the weight rule (SDMP/DSMP/DMSP). Second, under some conditions, using simulation is no longer required to produce the most efficient strategy (SDM/DSM/DMS). This is because balance beams for which neither the symmetry, weight, or distance rules apply are ones that are particularly challenging, as the weights and configurations of each side will be very similar. In these cases, noisy simulation is unlikely to provide a reliable answer, and therefore it is optimal to simply guess and not accrue the cost of simulation. Thus it is an open question whether, if participants notice the distance rule is an option (see Section 9.2.3), that they will in some cases eschew the use of simulation and rely solely on the other three rules.

A4 Supplemental experiment: Physical versus geometric processing

We have suggested that judgments about balance beams are based on a combination of rules and simulation that work over a physically plausible representation of the world. Yet many of the prior proposed rules for making judgments about balance beams require a very sparse representation of the world, capturing (at most) the number of blocks and where they are placed on the beam, but not requiring information about physical attributes of the blocks, such as their material properties or actual weights. Although the experiments with blocks of different materials suggest that people do take into account these physical properties (Exp. 2 & 4), we have not directly tested whether those judgments could be alternately explained by a purely “geometric” parsing of the scene. In this experiment, we therefore asked participants to make judgments about balance beams with blocks of different materials (similar to Experiment 2) such that accounting for the physical material properties was required to accurately judge how the beam would fall.

A4.1 Experiment

We recruited 26 participants from Mechanical Turk, who were compensated $1.50 for their time.

The procedure was identical to the materials experiment; only the stimuli were different.

Participants each made judgments about 140 different balance beams, created in matched pairs such that the geometric representation of the beam was identical across the pairs (e.g., the same number of blocks were positioned the same distance from the pivot), but one of the pair was made of blocks of the same materials (pure) while the other beam in the pair used blocks of different materials. These configurations were designed so that the way the beam fell or balanced would be different for each member of the pair, in one of five ways (see Figure A7, Left):

1. *Balance -> Weight.* The beam had a symmetric configuration of blocks so would balance with no material differences, but would fall to the side with more weight with separate materials.

2. *Weight -> Weight.* Blocks were placed in a typical 'weight' configuration such that the distances were identical but there were more blocks on one side. When the materials differed, however, the aggregate weight of the side with fewer blocks was large enough to cause the beam to tip in that direction.

3. *Conflict-balance -> Conflict-weight.* The blocks were placed in a 'conflict-balance' configuration such that with the same material one side had more blocks and the other had greater distance, but the torques were balanced. However, when the materials differed, the side with more weight would be the one to fall.

4. *Conflict-distance -> Conflict-weight.* The blocks were placed in a 'conflict-distance' configuration such that with a single material type the side with fewer blocks and further distance would fall; however, with different materials the weight of the side with more blocks would overcome the distance and the beam would fall to that direction.

5. *Other -> Conflict-balance.* These beams were designed around the mixed blocks, so that there was more weight on one side and more distance on the other, but the torques were equal on both sides. On the other hand, the matched configurations with a single material

type would fall, but the beam configuration was not well controlled – these could either be conflict problems, or trivial problems that were never classified because either a simple focus on weight or on distance would provide the correct answer.

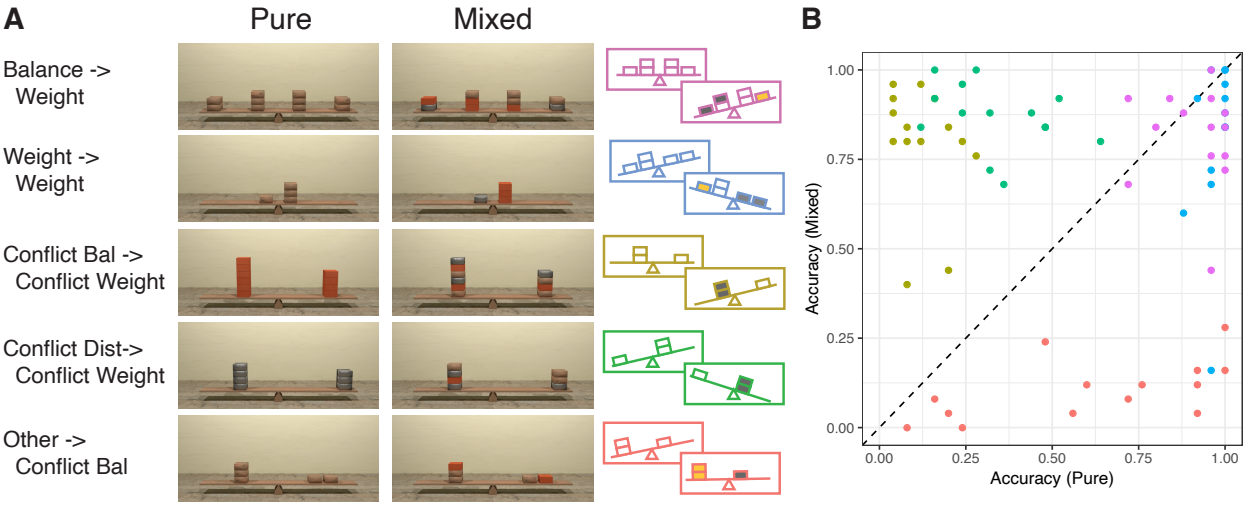

Figure A7. *Left:* Example trials from the experiment. Across all matched trials, the configuration of blocks was identical; only the distribution of materials differed. *Right:* Plot of accuracy on 'pure' material trials (x-axis) versus accuracy on the matched 'mixed' materials trials, with each point representing a single matched trial pair. A negative correlation would be expected if people were using purely geometric reasoning, but no correlation was observed ($r = -0.017$).

A4.2 Results

If participants were using purely geometric information to make judgments about the balance beams, then we would expect their judgments to be similar across the matched trials since the only difference is the non-geometric material attributes of each block. And because the trials were designed so that the correct answer differed across the matched pairs, we would then expect accuracy to be anti-correlated across those pairs. However, as can be seen on the right side of Figure A7, the correlation between the pure and mixed accuracies for each of these matched trials is effectively nonexistent ($r = -0.017$, $p = 0.90$), which suggests that participants were not treating the pairs identically.

However, we also do not expect accuracies to be well correlated across the matched trials either – even if participants are appropriately accounting for the different materials, people are more accurate on some beam configurations than others. This can be observed in the right side of Figure A7: with matched trials where the configurations are simple (purple: balance -> weight and blue: weight -> weight), accuracy is very high for both the pure and mixed trials. Conversely, when the pure blocks are in a difficult conflict configuration, but the mixed blocks make a conflict-weight beam (which people are fairly accurate on), accuracy on the pure trials is low but accuracy on the mixed trials is high. And in the 'other -> conflict-balance' trials, accuracy is low in the mixed, conflict-balance trials, but varies on the pure trials because some configurations are trivial while others are in difficult conflict patterns.

But despite the range of differences between trial pairs, the ISR model can explain participants' predictions well (see Figure A8, left). Just as with Experiments 4 and 5, we can re-use the ISR model without re-fitting any parameters, and this model explains participants' predictions well in aggregate ($r = 0.91$), and across both the pure ($r = 0.94$) and mixed trials ($r = 0.88$).

We can also test how well a purely geometric model can explain participants' predictions. This model is identical to the ISR model, except it treats materials as all having identical density. Even though this geometric model was fit on the data from this experiment, it still could not explain participants' predictions as well as as the hybrid model ($r = 0.67$, see Figure A8, right). Though it does an adequate job explaining how people perform on the 'pure' trials where material information is not required ($r = 0.89$), it fails to explain predictions on the 'mixed' trials ($r = 0.46$), as it expects participants to do poorly on the 'balance -> weight' and 'weight -> weight' mixed trials where accuracy is very high. Thus people are using physical information such as density in their predictions rather than simply attending to the geometric configuration.

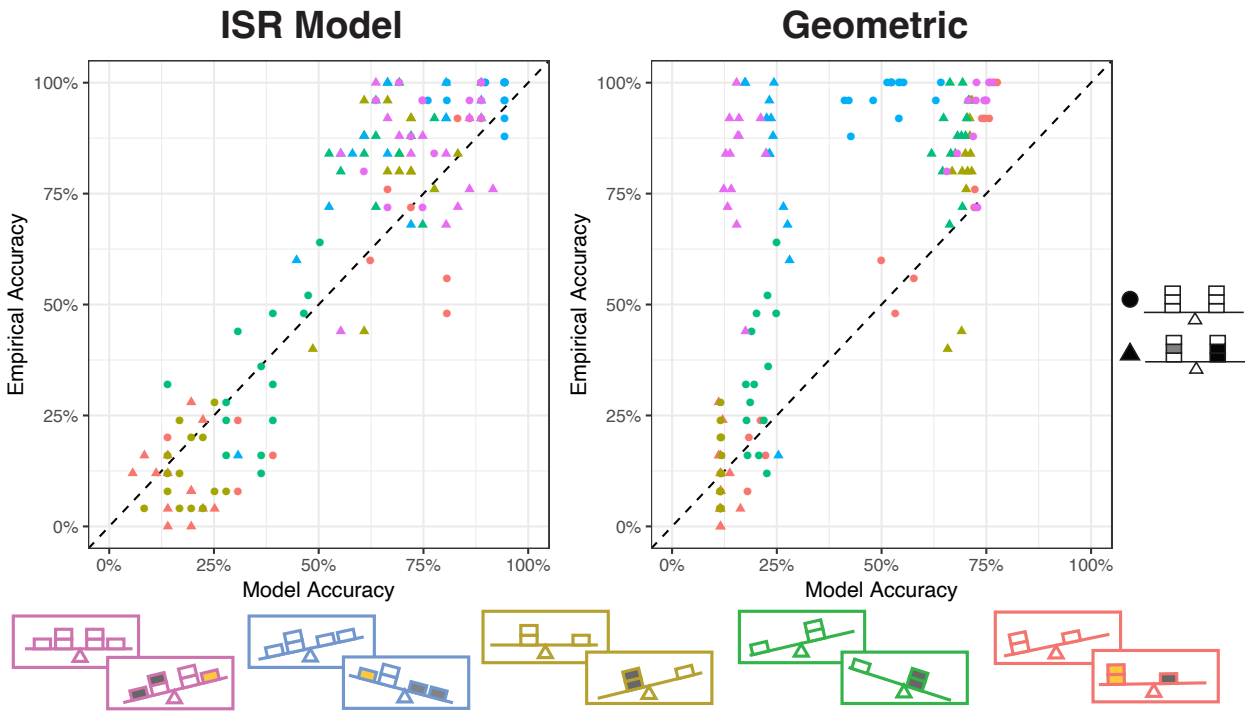

Figure A8. Plot of model predicted accuracy for each trial (x-axis) versus observed empirical accuracy (y-axis). Each point represents a trial, with colors representing the pair configuration and the shape representing the material type (circles: pure; triangles: mixed). *Left:* Comparison of the ISR model. *Right:* Comparison of the geometric processing model.