Inferring Mass in Complex Scenes by Mental Simulation

Jessica B. Hamrick
Department of Psychology
University of California, Berkeley

Peter W. Battaglia
Google DeepMind, London
and
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Thomas L. Griffiths
Department of Psychology
University of California, Berkeley

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Author Note

Address correspondence to: Jessica B. Hamrick, 3210 Tolman Hall, University of California, Berkeley, Berkeley CA 94720. Contact: jhamrick@berkeley.edu

Abstract

After observing a collision between two boxes, you can immediately tell which is empty and which is full of books based on how the boxes moved. People form rich perceptions about the physical properties of objects from their interactions, an ability that plays a crucial role in learning about the physical world through our experiences. Here, we present three experiments that demonstrate people's capacity to reason about the relative masses of objects in naturalistic 3D scenes. We find that people make accurate inferences, and that they continue to fine-tune their beliefs over time. To explain our results, we propose a cognitive model that combines Bayesian inference with approximate knowledge of Newtonian physics by estimating probabilities from noisy physical simulations. We find that this model accurately predicts judgments from our experiments, suggesting that the same simulation mechanism underlies both peoples' predictions *and* inferences about the physical world around them.

Keywords: mental simulation; physical reasoning; probabilistic simulation; inference; learning

Inferring Mass in Complex Scenes by Mental Simulation

Consider the scene in Figure 1a. Despite the difference in size, one can infer that the mass of the forklift is large compared to that of the storage container. Inferences about the physical properties of objects such as mass and friction are critical to how we understand and interact with our surroundings. While they are sometimes specified unambiguously by a small set of perceptible features such as size, material, or tactile sensations, we often access them only indirectly via their physical influence on observable objects. Here, we ask: how do people make such inferences about the unobservable physical attributes of objects from complex scenes and events?

In addition to one-off inferences about properties such as mass, people form beliefs about these properties over time. For example, through experience, people learn that certain materials (e.g., metal) are heavier than others (e.g., plastic). How is it that people learn these attributes? Certainly, people may rely on sensorimotor feedback as they hold and manipulate objects (e.g. Baugh, Kao, Johansson, & Flanagan, 2012). Can people also learn through experience if only visual information about the static and dynamic behavior of such objects is available? If so, what is the mechanism by which they do this?

There is a vast literature on whether (and if so, how) people reason about mass. People are clearly sensitive to mass when reasoning about other physical properties: for example, people's memory for the location of an object is affected by its implied weight (Hubbard, 1997); similarly, people make different judgments about how a tower of blocks will fall down depending on which blocks they think are heavier (Battaglia, Hamrick, & Tenenbaum, 2013). Previous studies of how humans *infer* mass from observed collision dynamics have examined the relative roles of perceptual invariants (Runeson, Juslin, & Olsson, 2000) and heuristics (Gilden & Proffitt, 1994; Todd & Warren, 1982), focusing on judgments about simple one- or two-dimensional (1D or 2D) situations with one or two objects. However, the real world is much more complex: everyday scenes are

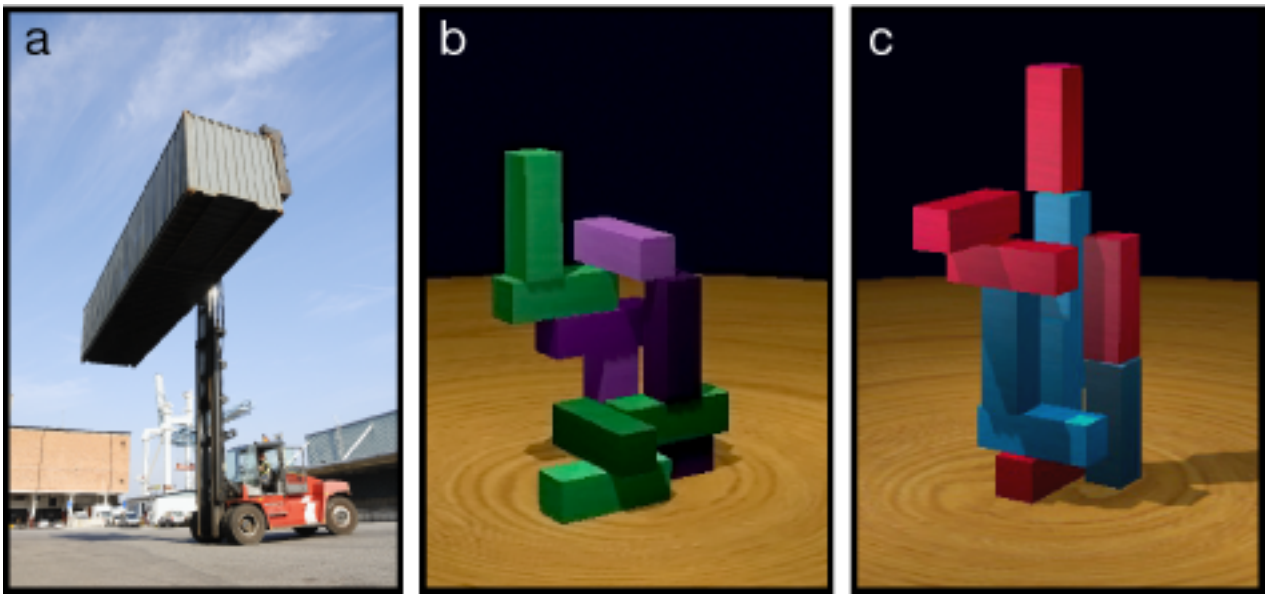

Figure 1. Three scenes that engage our ability to reason about mass. (a) The forklift's weight counterbalances the container's. (b-c) Two examples of experimental stimuli. If the green blocks in (b) are heavier than the purple blocks, you can predict that the tower will fall down rather than remain standing. If the tower in (c) stays standing, you can infer that the blue blocks are heavier.

three-dimensional (3D) and often involve many objects[1]. Moreover, collisions between objects are not the only factor affecting peoples' judgments: for example, there are no collisions in the forklift scene in Figure 1a, yet we can easily infer what the relative masses of the objects might be.

A question related to *whether* people can make accurate inferences about unobservable physical properties is *how* they make any inferences at all. Sanborn, Mansinghka, and Griffiths (2009, 2013) proposed that inferences could be characterized by a model that performs Bayesian inference over structured knowledge of Newtonian physics[2]

[1]We define a 3D scene to be any scene that contains depth information, regardless of whether it is viewed as a 2D projection. We define a 2D scene to be a scene with no depth cues (i.e., it is truly 2D and not a projection of a 3D scene).

[2]In this context, we take "structured" to mean implicit knowledge of formal physical laws, in contrast to

and noisy or uncertain perceptual inputs. In the 2D case, this "Noisy Newton" hypothesis works well for inferring properties like mass because the laws of Newtonian physics (such as conservation of momentum) can be encoded as distributions over random variables such as velocity, where the randomness comes from perceptual uncertainty (Sanborn et al., 2009, 2013; Sanborn, 2014). However, for scenes involving both statics and dynamics, it is not clear where these probabilities should come from. For example, if the forklift in Figure 1a is about to tip over, you can infer that the storage container is heavier, because if it were not, the forklift would likely remain upright. Where does this "likelihood of remaining upright" come from?

Recent research has proposed that people reason about complex environments using approximate and probabilistic mental simulations of physical dynamics (Battaglia et al., 2013; Hamrick, Battaglia, & Tenenbaum, 2011). They are *approximate* in the sense that they do not analytically solve the exact equations that underly Newtonian physics, but rather estimate the implications of those equations through an iterative process. They are *probabilistic* in that the simulations are non-deterministic, where the stochasticity reflects uncertainty that arises from noisy perceptual processes and imperfect knowledge of the scene. There is a growing body of evidence that people use such approximate and probabilistic mental simulations, including explanations of human judgments of physical causality and prediction in a wide range of scenarios (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2014; Hamrick, Smith, Griffiths, & Vul, 2015; Smith & Vul, 2013; Smith, Battaglia, & Vul, 2013; Smith, Dechter, Tenenbaum, & Vul, 2013; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2014). A similar hypothesis has also been proposed by White (2012), suggesting that simulations are the result of retrieving past perceptual experiences and extrapolating using a forward model.

If people use probabilistic mental simulations to make predictions about physical

---

implicit knowledge of naïve physics or explicit knowledge of formal physics. See the General Discussion for further discussion of how these differing forms of physical knowledge relate.

scenes, then it should be possible for people to use those simulations to estimate the probabilities of different outcomes. These probabilities can be used to make inferences about unobservable physical properties. Indeed, recent work by Ullman et al. (2014) and Gerstenberg et al. (2012, 2014) has provided examples of how simulations might be used in simple 2D scenes to estimate the necessary probabilities for Bayesian inference. However, such an approach has not been applied to the types of complex, 3D scenes that people encounter in the real world.

Using probabilistic simulation to make inferences about unobservable physical properties also suggests a unified framework both for reasoning about individual object-level properties (i.e., that the forklift in Figure 1a is heavier than the storage container) as well as class- or material-level properties (i.e., that objects made out of stone are heavier than objects made out of plastic). Historically, research has focused on how people make one-shot inferences about the properties of individual objects (Gilden & Proffitt, 1989a; Runeson et al., 2000; Sanborn et al., 2009, 2013; Sanborn, 2014; Todd & Warren, 1982), but not on how these inferences might also play a role in learning class-level properties such as the density of a particular material. We suggest that if Bayesian inference is performed using probabilities obtained through approximate physical simulation, then this could provide an account for *both* one-shot inferences and learning. Specifically, Bayes' rule dictates both how to compute inferences about individual objects, as well as how to integrate multiple pieces of information over time to learn about the properties of classes of objects.

This work is the first to explore people's ability to make inferences about mass in complex scenes that may be either static or dynamic, and addresses two questions regarding this ability. First: *can* people make accurate inferences? To answer this, we present three experiments in which we asked people to make inferences about the relative masses of objects in complex scenes involving both static and dynamic objects. We find that people can form accurate judgments about the relative mass, and that they become

increasingly fine-tuned to these properties as they accumulate multiple pieces of information. Second: *how* do people make inferences about properties like mass? We introduce a new cognitive model that uses approximate, probabilistic simulation to estimate probabilities needed by Bayesian inference to produce judgments about the relative mass of objects. When compared to data from our experiments, we find that our model is a good characterization of how people make inferences about the masses of individual objects and how they learn about the mass of a class of objects. Moreover, by replacing the model's simulations with people's own predictions about the future dynamics of the scenes, our model can predict people's inferences about mass with high accuracy. This suggests that the same simulation-based mechanism the mind uses for predicting physics is also involved in forming physical inferences about object-level properties and in learning the properties of a class of objects over time.

Experiment 1: Inferring Mass From a Single Trial

We first asked the question: *can* people infer mass in complex scenes? To answer this, we ran three experiments in which we showed people videos of towers of blocks, where each block was one of two colors, and asked them to judge which color was heavier. This section describes the first of these experiments.

Participants

We recruited participants on Amazon's Mechanical Turk using the psiTurk experiment framework (Gureckis et al., 2015). We randomly assigned participants to one of four conditions, which determined which stimuli they saw (see Design for details). Based on sample sizes from earlier pilot studies, we aimed to have 20 participants per condition that would result in 40 judgments per stimulus. However, also based on earlier pilot data, we expected about 20% of participants to fail an attention check (see the *posttest* in the Procedure section for details). Thus, we aimed for 100 participants overall, though we ended up recruiting 101 participants because one participant had completed an earlier version of the experiment. We excluded this one participant from analysis, as well as 19

Table 1

Phases of Experiment 1.

	Phase				
	Familiarization	*Training*	*Prediction*	*Inference*	*Posttest*
Stimuli	control	training	experimental	experimental	control
Block color	random	red/blue	red/blue	varied	random
Judgment	fall?	fall?	fall?	heavier?	fall?
Feedback	yes	yes	no	-	yes

participants for failing the attention check. All participants were treated in accordance with UC Berkeley's IRB protocols and were paid $1.25. All participants were 18 years or older, and completed the experiment from within the United States. Participants took a median of 14.6 minutes to perform the full experiment.

Stimuli

Stimuli were videos of computer generated 3D towers of identically sized building blocks arranged in a way such that both the position and orientation of the blocks were relevant to the dynamics of the scene (see Appendix A for details on how stimuli were constructed, and see Figure 1b-c for two example stimuli). Stimulus presentation videos showed the camera (beginning from a random angle) rotating 180° counterclockwise around the tower for 5 seconds, during which gravity was set to zero so that none of the towers would fall down. Separate feedback videos showed the physical dynamics of the tower falling or not falling for 3.5 seconds under a gravitational acceleration of $-9.81\frac{m}{s^2}$. The feedback videos began from the last frame of the stimulus presentation video that depicted the same tower.

Experimental stimuli. There were 20 experimental towers, which consisted of 5 blocks of one color and 5 blocks of another color (see Figure 1b-c for two examples). The relative mass ratio between the differently-colored blocks was either 1:10 or 10:1, which is

approximately the same as that between a metal steel block and a wooden oak block. We chose the experimental towers such that whether the tower fell or not was determined both by the geometry of the structure, as well as the mass ratio[3] For example, if a tower had a mass ratio of 1:10 and fell down, then a tower with the same geometry but a mass ratio of 10:1 would stay standing (see Appendix A for details). Note that this has the consequence that for all the towers we chose, if the tower fell under one mass ratio, it would *not* fall under the other, and vice versa (see the section on "Reasoning About Physical Properties via Simulation" for further discussion of the consequence of this choice).

Training stimuli. There were also ten training stimuli, which were of the same form as the experimental stimuli (5 blocks of one color and 5 blocks of another color, with relative masses of 1:10 or 10:1).

Control stimuli. Finally, there were six control stimuli, which consisted of 10 randomly-colored blocks that all had the same mass and which were taken from a previous experiment (Battaglia et al., 2013). We chose the control towers to be those towers which participants rated in that experiment as extremely stable or extremely unstable.

Design

The experiment consisted of five phases: *familiarization*, *training*, *prediction*, *inference*, and *posttest*. Within each phase, the trial order was randomized for each participant. Participants were assigned randomly to one of four conditions, which determined the mass ratios of the towers in the training, prediction and inference phases. Specifically, we constructed a 2×2 design in which participants observed towers with a mass ratio of either 1:10 or 10:1 in the training and prediction phases, and towers with a separate mass ratio of either 1:10 or 10:1 in the inference phase. A summary of each phase is provided in Table 1.

[3]The stability of a given tower is strongly dependent on perceptual information, such as its geometry, as well as explicit information, such as which blocks participants think are heavier. See Battaglia et al. (2013) for a detailed investigation into what determines the stability of a tower.

Familiarization phase. The familiarization phase familiarized participants with the "will it fall?" decision. They answered this question for the six control stimuli, and received feedback after responding.

Training phase. The training phase familiarized participants with the task of judging "will it fall?" when the blocks could have different masses. Participants answered "will it fall?" for the 10 training stimuli, which had blocks colored red and blue; participants were told which color block was heavier. Depending on condition, the mass ratio between the blocks was either 1:10 or 10:1, and the colors were counterbalanced. Participants received feedback after responding.

Prediction phase. In the prediction phase, participants again judged "will it fall?", but for the 20 experimental stimuli. This phase was included in order to *a priori* estimate how likely people thought it was for towers to fall under different mass ratios. The blocks in these stimuli were also colored red and blue, and had the same mass ratio as in the training phase. Participants did not receive any feedback.

Inference phase. The inference phase was designed to gather participants' inferences about the relative mass ratios of the blocks. Participants answered "which is the heavy color?" for the same 20 experimental stimuli (in a different order) as in the prediction phase after observing a video of the tower falling or not falling. The colors of the blocks changed on every trial, and no pair of colors was shown more than once, though individual colors were reused in different pairs of colors (see Appendix A for all pairs of colors). Depending on condition, the mass ratio between the blocks was either 1:10 or 10:1, and was either the same or different as the ratio in the training and prediction phases. As before, whether a given color was assigned to be heavy or not was counterbalanced across participants.

Posttest phase. The posttest phase was identical to the familiarization phase, with the exception of trial order. The purpose of the posttest was to check that participants were paying attention by asking them to perform more stability judgments.

This design was motivated by the assumption that by the end of the experiment, participants should be able to judge the easier training towers with high accuracy. We excluded participants from analysis who incorrectly judged the stability of at least one (out of six) towers in the posttest.

Procedure

There were two types of trials: *stability* trials, in which participants were asked to predict whether the towers would fall down, and *mass* trials, in which participants were asked to infer which color block was heavier. Participants initiated all trials by pressing the 'c' key, after which the stimulus presentation began.

Stability trials. On stability trials, participants were then asked the question, "On a scale from 1-7, how likely is the tower to fall down?", where 1 meant "unlikely to fall" and 7 meant "likely to fall". Participants responded by pressing the corresponding number key. Participants then saw feedback (if any, depending on the phase) immediately after responding. Feedback consisted of a video depicting the tower either falling or not falling was shown in addition to text indicating "tower falls" or "tower does not fall".

Mass trials. On mass trials, after being shown the stimulus presentation, participants were prompted to press the 'c' key to view feedback. After the feedback was complete, they were asked the question, "Which is the heavy color?", and could click one of two buttons corresponding to the block colors.

Analysis

Before presenting the results, we describe a few generic analyses that we perform several times in the subsequent results section. For analyses of participant's accuracy, we computed medians and 95% confidence intervals using 10000 bootstrap samples. To test if people's judgments were above chance on a particular stimulus, we used the same bootstrap analysis and tested whether $p\left(p(\text{correct}) \leq 0.5\right) \leq \frac{0.05}{40}$, where $p(\text{correct})$ is an empirical probability of answering correctly and where $\frac{1}{40}$ is a Bonferroni correction for multiple comparisons. That is, this equation is a test for whether the empirical frequency

of answering correctly is significantly greater than 0.5 (chance), where "significantly" is defined as $p < 0.05$, adjusted for multiple comparisons. For correlation analyses, we used a bootstrap analysis of 10000 bootstrap samples to compute the median and 95% confidence intervals of both Spearman (ρ) and Pearson (r) correlations. Any other reported confidence intervals were similarly computed using 10000 bootstrap samples with replacement.

Results

Overall accuracy in responding to "which is the heavy color?" was significantly above chance ($M = 81.9\%$, 95% CI $[79.9\%, 83.7\%]$, averaged across participants and stimuli), though there were 9 individual stimuli (out of 40) for which accuracy was not significantly above chance (corrected for multiple comparisons). We suspect that the stimuli which participants did not classify above chance were not classified as such either due to a combination of (1) insufficient power and (2) the information in the stimulus presentation genuinely not being very informative. This latter possibility is supported by our cognitive model later in the text (see the section on "Reasoning About Physical Properties via Simulation" and Figure 4). Accuracy of individual participants ranged from 45% to 100%, and 95% of participants answered correctly on at least 60% of the trials.

We also looked at how well participants' responses predicted each other, both in response to "will it fall?" and "which is the heavy color?". To compute this, we performed a bootstrap analysis in which each bootstrap sample was computed by randomly dividing the participants in half and calculating the correlation of average judgments between the two groups. For "will it fall?" judgments, this split-half correlation was $r = 0.90$, 95% CI $[0.84, 0.94]$, implying that people were very consistent with each other. Similarly, for "which is the heavy color?" judgments (where 0 corresponded to a ratio of 1:10 and 1 corresponded to a ratio of 10:1), the correlation was $r = 0.95$, 95% CI $[0.92, 0.97]$, again indicating that people were highly consistent. For accuracy in judging the mass[4] (where 0 corresponded to an incorrect answer, and 1 to a

[4]Note that here, sometimes the correct ratio was 1:10 and sometimes it was 10:1. Therefore average

correct answer), people were more variable, with a correlation of $r = 0.69$, 95% CI $[0.55, 0.80]$.

To check whether participants improved over time due to practice effects, we computed the Spearman rank correlation between trial number and mean accuracy. We found no significant effect of practice ($\rho = 0.07$, 95% CI $[-0.20, 0.34]$).

Discussion

In this experiment, participants inferred which blocks were heavier with high accuracy even though they were only shown one example tower per judgment. We next asked: can people accumulate information over time and further improve their inferences based on multiple examples? To answer this, we ran another experiment similar to Experiment 1. The main difference was that during the inference phase, the block colors did not change, and participants were told that the mass ratio remained the same.

Experiment 2: Learning From Multiple Trials (Within Subjects)

Participants

As in Experiment 1, we recruited participants from Amazon's Mechanical Turk using the psiTurk experiment framework (Gureckis et al., 2015) and assigned participants randomly to one of four conditions. We aimed to have 20 participants per condition, and expected 20% of participants to fail the attention check. Thus, we set our target sample size to 100 participants. Overall, we collected data from 111 participants because 7 participants had incomplete or duplicate data due to an experimental error and 4 participants had already done an earlier version of the experiment. We excluded these participants from analysis, as well as 15 participants for failing the attention check. Participants were treated in accordance with UC Berkeley's IRB protocols and were paid $1.00. All participants were 18 years or older and completed the experiment from within the United States. Participants took a median of 16.2 minutes to perform the full experiment.

---

judgments to "which is the heavy color?" are not the same as the average *accuracy* of those judgments.

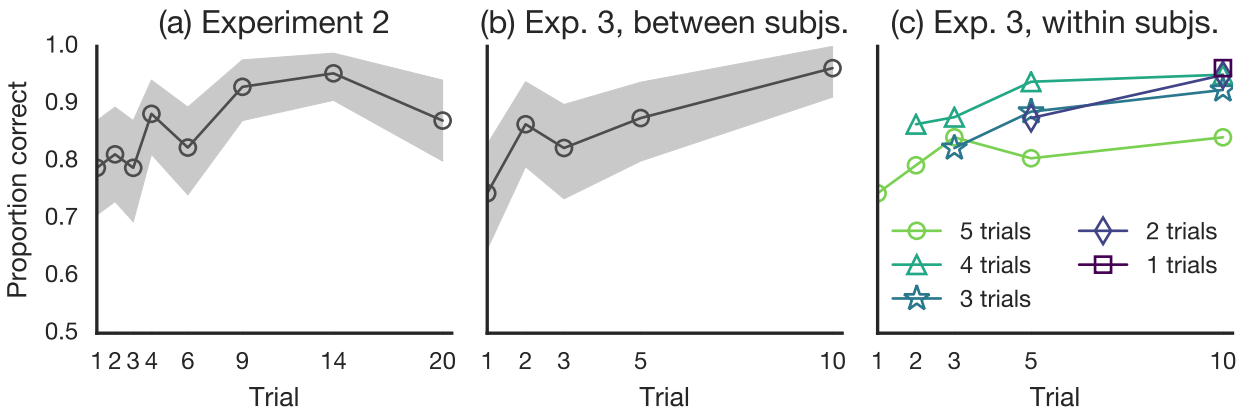

Figure 2. Inferences of relative mass in Experiments 2-3 as a function of trial. In all plots, the solid lines indicate the mean proportion of correct human responses to "which is the heavy color?". Shaded regions are 95% confidence intervals of the mean. (a) The left subplot shows accuracy as a function of trial in Experiment 2. (b) The middle subplot shows between-subjects accuracy in Experiment 3. (c) The right subplot shows within-subjects accuracy in Experiment 3.

Design

The design for Experiment 2 was identical to that of Experiment 1, except during the inference phase. As in Experiment 1, we did not tell participants which color was heavier, but we did tell them that the heavier blocks would always have the same color. Instead of having different color pairs on every trial, the colors were always purple and green (e.g., Figure 1b), and these colors were counterbalanced across conditions. Participants judged the relative mass only on trials 1, 2, 3, 4, 6, 9, 14, and 20. These trials were chosen based on the hypothesis that participants' beliefs would change a lot at the beginning of the experiment and less so at the end of the experiment; thus, it would be better to ask more frequently at the beginning of the experiment rather than after a fixed interval. On trials where they were not asked, they just watched the stimulus presentation and feedback video, and then immediately moved on to the next trial.

Results

As in Experiment 1, participants were above chance in judging which color was heavier overall ($M = 85.6\%$, 95% CI $[82.9\%, 88.2\%]$, averaged across participants and stimuli), though there were 14 individual stimuli for which this was not significant (corrected for multiple comparisons, see the Analysis section of Experiment 1). Accuracy of individual participants ranged from 25% to 100%, and 95% of participants answered correctly on at least 50% of the trials.

Unlike Experiment 1, participants in Experiment 2 were told that the mass of the blocks remained the same during the inference phase. Thus, we would predict their accuracy to generally increase as a function of trial. This trend does appear ($\rho = 0.68$, 95% CI $[0.25, 0.90]$), though as shown in Figure 2a, is not monotonic.

Participants were very self-consistent in their responses to "will it fall?", with a split-half correlation of $r = 0.90$, 95% CI $[0.84, 0.94]$. Their responses were also consistent with those in Experiment 1, with a correlation of $r = 0.92$, 95% CI $[0.86, 0.96]$.

Discussion

Why did the accuracy in judging the mass ratio go *down* between trials 14 and 20? If participants were learning over time, we would expect the last trial to have the highest accuracy. We suspected that participants may have been confused by being asked "which is the heavy color?" multiple times, perhaps thinking that they needed to change their response. Alternatively, it is possible that the delay between responding to "which is the heavy color?" might have resulted in a decay of any learning that did happen. To address these issues, we ran a third experiment similar to Experiment 2 that was shorter and in which we varied the number of times participants were asked "which is the heavy color?".

Experiment 3: Learning From Multiple Trials (Between Subjects)

Participants

We recruited participants on Amazon's Mechanical Turk using the psiTurk experiment framework (Gureckis et al., 2015). Participants were randomly assigned to one of ten conditions (see Design for details), and to keep sample sizes consistent with the first two experiments, we aimed to have 40 participants per condition. Based on Experiments 1-2, we expected that about 17% of participants would fail the attention check. Thus, we aimed for 480 participants total, though we actually collected data from 487 participants because 7 participants had completed an earlier version of the experiment. We excluded these participants from analysis, as well as 79 participants who failed the attention check. Participants were treated in accordance with UC Berkeley's IRB protocols and were paid $0.70. All participants were 18 years or older and completed the experiment from within the United States. Participants took a median of 8.3 minutes to perform the full experiment.

Design

Experiment 3 utilized nearly the same design as Experiment 2, except that participants did not complete the prediction phase, and they only completed 10 trials of the inference phase (randomly selected from the full set of 20 stimuli). Rather than make judgments on the exact same trials, five different subgroups of participants were asked to judge the mass different numbers of times: on trials 1, 2, 3, 5 and 10; on trials 2, 3, 5 and 10; on trials 3, 5 and 10; on trials 5 and 10; and just on trial 10. Crossed with the two mass ratios (1:10 and 10:1), this led to a total of ten conditions in Experiment 3. This design was chosen in order to isolate the effect of asking the question of "which is the heavy color?" multiple times. If participants were getting confused by being asked the question multiple times, then the participants in this experiment who answer the question first on the 3rd trial, for instance, should have a higher accuracy than those that answer first on the 1st or 2nd trials. Additionally, this design tests the alternate hypothesis that learning could be

decaying due to the gap between questions: if learning is decaying, then participants who answer only on the 10th trial should arguably do worse than any of the other conditions.

Results

As with the previous experiments, participants were above chance in judging which color was heavier across all stimuli ($M = 87.1\%$, 95% CI $[85.2\%, 88.9\%]$, averaged across participants and stimuli). There were only 2 individual stimuli for which people's judgments were not significantly above chance (corrected for multiple comparisons, see the Analysis section for Experiment 1).

To judge the effect of learning over time, we computed participant's accuracy as a function of trial both between and within subjects. To compute the between subjects accuracy, we took only the first responses from each condition (i.e., only the first time each participant answered "which is the heavy color?"). The Spearman rank correlation between subjects was significant ($\rho = 0.70$, 95% CI $[0.40, 1.00]$), while the rank correlation for the condition in which participants responded on five trials was not ($\rho = 0.60$, 95% CI $[-0.31, 1.00]$). Figure 2b-c shows both the between- and within-subjects accuracy as a function of trial.

Discussion

These results suggest that demand effects were at play both in Experiment 2 and within subjects in Experiment 3. However, the structure of Experiment 3 allowed us to perform between-subject analyses, revealing that when participants were not biased by being asked the same question multiple times, they did become increasingly accurate over time. The fact that people are able to incorporate this information over time suggests a way in which people might learn about the properties of classes of objects (for example, the densities of particular materials). Additionally, our results rule out the hypothesis that learning decayed in the gaps between when we asked them which color they thought was heavier, because participants who responded on fewer trials (e.g. only trial 10) did better than those that responded on more trials (e.g. trials 1, 2, 3, 5, and 10).

Reasoning About Physical Properties via Simulation

The three experiments described previously demonstrate that people *can* make accurate inferences about mass in complex scenes, and that they can accumulate evidence over time. Our next question was: *how* do people infer mass? We hypothesized that people's inferences can be characterized using Bayesian inference in which the probabilities are computed via probabilistic simulation. Here, we formalize this hypothesis in a model observer and compare judgments from the model with people's judgments in Experiments 1-3.

Observer Model

On each trial, the observer model views a stimulus (S) and receives feedback (F). The feedback is a Bernoulli random variable indicating whether the tower fell ($F = 1$) or did not fall ($F = 0$). Let κ be a random variable corresponding to the mass ratio, and let a particular hypothesis about the mass ratio be indicated by $\kappa = k$ (i.e., either $\kappa = 0.1$ or $\kappa = 10$). Then, the probability of that hypothesis given the observed feedback is computed from Bayes' rule:

$$p(\kappa|F, S) = \frac{p(F|S, \kappa = k)p(\kappa = k)}{\sum_{k'} p(F|S, \kappa = k')p(\kappa = k')}, \tag{1}$$

where $p(\kappa = k)$ is the prior probability of the hypothesis that $\kappa = k$, and $p(F|S, \kappa = k)$ is the probability of observing the feedback given that the hypothesis $\kappa = k$ is true.

Equation 1 demonstrates how the observer model computes its belief about the mass ratio after a single trial. We can further extend this model to show how an observer model should update its beliefs over time as it encounters more evidence. Specifically, the observer model should use the posterior distribution of each trial as the prior distribution of the next:

$$\begin{aligned} p_t(\kappa = k) &= \frac{p(F_t|S_t, \kappa = k)p_{t-1}(\kappa = k)}{\sum_{k'} p(F_t|S_t, \kappa = k')p_{t-1}(\kappa = k')} \\ &= \frac{p_0(\kappa = k)\prod_{i=1}^{t} p(F_i|S_i, \kappa = k)}{\sum_{k'} p_0(\kappa = k')\prod_{i=1}^{t} p(F_i|S_i, \kappa = k')}, \end{aligned} \tag{2}$$

where $p_t(\kappa = k) := p(\kappa = k|F_1, S_1, \ldots, F_t, S_t)$ is the belief about the mass ratio after observing t trials. Thus, $p_0(\kappa = k)$ is the prior, $p_1(\kappa = k)$ is the belief after the first trial, and so on. Note that each $p_t(\kappa = k)$ is defined recursively in terms of $p_{t-1}(\kappa = k)$ and thus contains all information observed so far up through trial t. This model is consistent with the optimal inference computation for an observer that aggregates information across trials.

We can contrast the model defined in Equation 2 (henceforth referred to as the *learning* model) with an observer that does not accumulate information over trials. This *static* model does make inferences according to Equation 1, but effectively starts from scratch on each trial:

$$p_t(\kappa = k) = \frac{p(F_t|S_t, \kappa = k)p_0(\kappa = k)}{\sum_{k'} p(F_t|S_t, \kappa = k')p_0(\kappa = k')}. \tag{3}$$

As above, p_0 is the prior. This model is consistent with the optimal inference computation for an observer that does not aggregate information across trials.

Estimating Probabilities with Simulation

We posit that people compute the likelihood term $p(F_t|S_t, \kappa = k)$ in Equation 1 using their "intuitive physics engine" (IPE), as proposed by Battaglia et al. (2013) and Hamrick et al. (2011). We will refer to this version of the observer model as the "IPE observer model". The IPE is a hypothetical cognitive mechanism that makes predictions by running forward noisy physical simulations. Because these simulations are non-deterministic, running many under different values of κ allows people to estimate how likely a particular outcome is under each hypothesis.

The IPE observer model runs N simulations and, for each simulation, compares the initial and final states of the stimulus. Specifically, the estimated probability of observing F_t given a stimulus S_t and the mass ratio κ is:

$$p(F_t|S_t, \kappa = k) \approx \frac{1}{N} \sum_{i=1}^{N} b^{(i)}, \tag{4}$$

where $b^{(i)} \sim \text{IPE}(S_t, \kappa)$ is the fraction of blocks (ranging from 0 to 1) that moved more than 0.25 cm from their initial positions during the i^{th} simulation. There are other possible

"queries" that could be used to define what it means for a tower to fall, such as whether any blocks moved at all. Such alternate queries yield similar results and are discussed further in Appendix C.

Empirically Estimating Probabilities

As stated previously, we hypothesize that people are using an IPE to predict how likely it is for a tower to fall. If this is the case, and if we use probabilities estimated from their predictions of "will it fall?" in our learning and static models, then these empirically-based models should predict people's inferences of mass just as well (or better) than those which use the IPE model's predictions. To test this hypothesis, we computed Equation 1 based on "will it fall?" judgments from the prediction phases of Experiments 1-2; we will refer to this version of the model as the "empirical observer model". In order to turn these judgments into probabilities, we assumed that the smallest response (1) was equivalent to 0 and the largest response (7) was equivalent to 1. By rescaling these judgments and averaging, we obtained an empirical estimate of the probability that the tower will fall, analogous to Equation 4:

$$p(F_t|S_t, \kappa = k) \approx \frac{1}{M}\sum_{i=1}^{M}\frac{J_{\text{fall}}^{(i)} - 1}{6}, \tag{5}$$

where M is the number of participants and $J_{\text{fall}}^{(i)}$ is the judgment of the i-th participant. This assumes participants are running only one simulation to make their judgments, which is not necessarily the case (Battaglia et al., 2013). While running more simulations would not change the overall mean of the likelihood, it would change the variance, and this possibility is investigated in more depth in the General Discussion.

Fitting Model Parameters

We fit both the static and learning models to human data using a Bayesian logistic regression. More formally, let the i^{th} participant's judgment of the mass ratio on trial t be a Bernoulli random variable denoted by $J_{\text{mass},t}^{(i)}$. Then, assume that a participant's

judgment is related to their belief via the logistic function:

$$p(J^{(i)}_{\text{mass},t} = k|\beta, \eta_t) = \frac{1}{1 + e^{-\beta\eta_t}}, \tag{6}$$

where β is a parameter that specifies how strongly the evidence is weighed and η_t is the posterior log odds at time t. Specifically, in the case of the learning model, we have:

$$\eta_t = \frac{p(F_t|S_t, \kappa = 10)p_{t-1}(\kappa = 10)}{p(F_t|S_t, \kappa = 0.1)p_{t-1}(\kappa = 0.1)}. \tag{7}$$

The equation for the posterior log odds is the same for the static model, except that $p_{t-1} = p_0$ for all t. Given this formulation, we are interested in finding the value of β which best fits a participant's responses, i.e., the maximum *a posteriori* (MAP) estimate over $p(\beta|J^{(i)}_{\text{mass},1}, \ldots, J^{(i)}_{\text{mass},T}, \eta_1, \ldots, \eta_T)$. To avoid overfitting, we used a Laplace prior with $\mu = 1$ and $b = 1$ (equivalent to L1 regularized logistic regression). Then, we found the MAP estimate of β for each participant separately.

In this regression, the β parameter reflects how strongly participants weigh the evidence they have observed. In the case of the static model, this translates to just how strongly they take into account the evidence on each trial; in the learning model, this translates to a learning rate. A value of $\beta = 1$ means that the model weighs evidence in accordance to Bayes' rule. A value of $0 < \beta < 1$ means that the model weighs the evidence *less* strongly than Bayes' rule does, but does still take it into account. A value of $\beta > 1$ means that the model weighs evidence *more* strongly than Bayes' rule does. A value of $\beta = 0$ means that the model ignores all evidence, and a value of $\beta < 0$ means that the model believes the *opposite* of what the evidence tells it.

Although the Laplace prior works to prevent overfitting, we were still concerned about the possibility of overfitting, particularly in Experiment 3 where we had only one or two responses from some participants. Thus, to be able to compare our models without making assumptions about which parameter values were correct, we computed Bayes factors (Kass & Raftery, 1995). Briefly, a Bayes factor is defined as the marginal likelihood ratio of the data given two different models, with the parameters of those models

integrated out. Thus, the Bayes factor gives a measure of how much better one model explains the data over another model, irrespective of the specific parameter values of the model. According to Kass and Raftery (1995), the value of the log of the Bayes factor can be interpreted as positive evidence if $1 < \log K \leq 3$, strong evidence if $3 < \log K \leq 5$, and very strong evidence if $\log K > 5$.

We computed our Bayes factors in the following manner. We first computed the marginal likelihood of participants judgments under each model by integrating over possible values of β:

$$p(J_{\text{mass}}|\eta) = \prod_{i=1}^{N} \left(\int \prod_{t=1}^{T} p(J_{\text{mass},t}^{(i)} = k|\beta^{(i)}, \eta_t^{(i)}) p(\beta^{(i)}) \, \mathrm{d}\beta^{(i)} \right), \tag{8}$$

where N is the number of participants, T is the number of trials, $J_{\text{mass},t}^{(i)}$ is the judgment of the i^{th} participant on trial t, $\eta_t^{(i)}$ is the posterior log odds of participant i on trial t, and $\beta^{(i)}$ is the parameter for participant i. We then computed Bayes factors by computing the log ratio between the marginal likelihoods for the learning model to the static model.

Results

Predictions. We first checked whether people's responses to "will it fall?" in the prediction phase were consistent with previous findings (Battaglia et al., 2013). We pooled responses to "will it fall?" from the prediction phases of Experiments 1-2, used them to compute Equation 5, and then compared them to IPE predictions (computed fation 4). We found that participants were well-predicted by the IPE model
($r = 0.75$, 95% CI $[0.57, 0.87]$), and note that this correlation is about the same as that found by Battaglia et al. (2013), which was $r = 0.80$, 95% CI [0.72, 0.86]. Figure 3 depicts this correlation.

Inferences from a single trial. As detailed in the methods section for Experiment 1, we chose towers such that the evidence for the correct mass ratio was maximized. Consequently, all the towers we chose had the feature that if the tower fell under one mass ratio, it would *not* fall under the other, and vice versa. By considering this

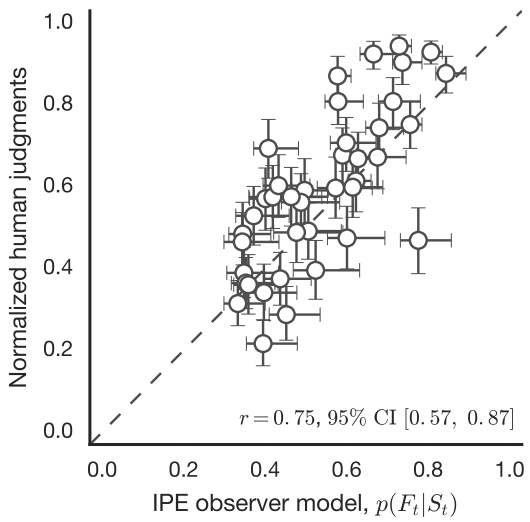

Figure 3. Responses to "will it fall?" in Experiments 1-2. The x-axis shows responses from the IPE model observer, and the y-axis shows responses from participants. Error bars are bootstrapped 95% confidence intervals. The dashed line indicates perfect correspondence between the model and people.

counterfactual, the observer gains more information than they otherwise would have; thus, Equation 1 does not tell the whole story. Let $\mathcal{L}(\kappa = k) := p(F_t|S_t, \kappa = k)$ be either the IPE likelihood or the empirical likelihood. Then, the "counterfactual likelihood" is:

$$p_{\text{CF}}(F_t|S_t, \kappa = k) = \frac{\mathcal{L}(\kappa = k)(1 - \mathcal{L}(\kappa \neq k))}{\mathcal{L}(\kappa = k)(1 - \mathcal{L}(\kappa \neq k)) + (1 - \mathcal{L}(\kappa = k))\mathcal{L}(\kappa \neq k)} \tag{9}$$

The derivation of this equation is given in Appendix D.

Figure 4a compares people's judgments with the posterior distribution calculated with the original empirical likelihood, and shows a clear sigmoidal relationship between the model and people, defined as $y = 1/(1 + \exp(-\beta(x - 0.5)))$. Using the original empirical likelihood, the best fit coefficient for sigmoid is $\beta = 11.75$, 95% CI $[9.82, 14.48]$. In contrast, when we switch to using the counterfactual likelihood (Figure 4b), this sigmoidal relationship significantly lessens ($\beta = 5.82$, 95% CI $[4.99, 6.84]$). To determine how different these sigmoid relationships are from linear, we computed best-fit sigmoid coefficients for

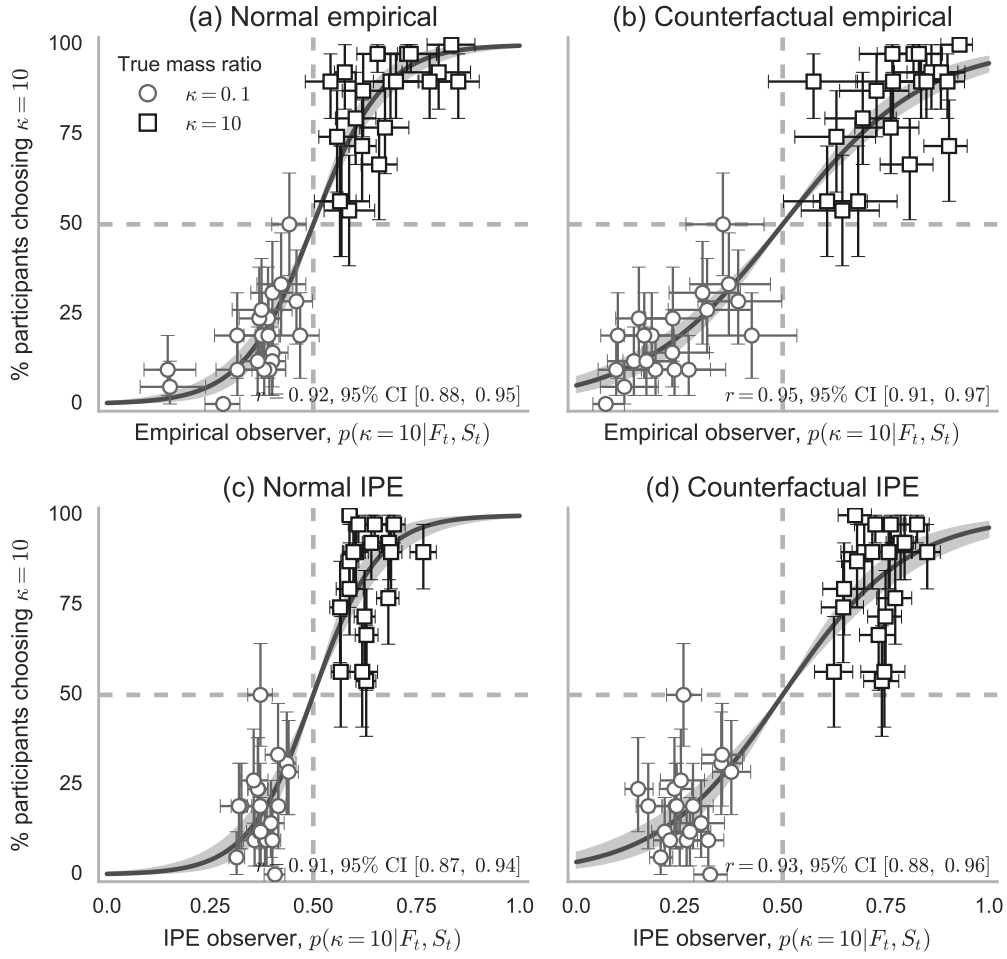

Figure 4. Comparing empirical and IPE likelihoods with and without counterfactual adjustments. Each subplot shows the posterior probability of $\kappa = 10$ in comparison to responses to "which is the heavy color?" in Experiment 1 according to different methods of calculating the likelihood. The best fitting sigmoid function for each relationship is plotted in black, with the shaded region indicating 95% confidence intervals for the best-fit coefficient. Error bars are bootstrapped 95% confidence intervals, and dotted lines show decision boundaries. (a) The posterior calculated using the original empirical likelihood. (b) The posterior calculated using the counterfactual empirical likelihood. (c) The posterior calculated using the original IPE likelihood. (d) The posterior calculated using the counterfactual IPE likelihood.

linear data with Gaussian noise, with the variance of the noise equal to the variance in the residuals between people's judgments and the likelihood. We find that a truly linear relationship gives a best-fit coefficient of $\beta = 4.94$, 95% CI $[3.55, 6.81]$ when using the residual variance for the original likelihood, and $\beta = 4.92$, 95% CI $[4.05, 5.97]$ when using the residual variance for the counterfactual likelihood, indicating that the original likelihood is distinguishable from a linear relationship while the counterfactual likelihood is not. Thus, it appears that participants picked up on this counterfactual information and exploited it when making their inferences about the mass ratio. To account for this, all results reported from this point were computed using an IPE observer model and an empirical observer model that took this counterfactual information into account.

As depicted in Figure 4d, the correlation between the IPE observer model probabilities and average human judgments of the mass ratio was $r = 0.93$, 95% CI $[0.88, 0.96]$; for the empirical observer model (Figure 4b), it was $r = 0.95$, 95% CI $[0.91, 0.97]$. The correlation between the IPE observer model accuracy and human accuracy was $r = 0.25$, 95% CI $[-0.04, 0.51]$, and for the empirical observer model, it was $r = 0.60$, 95% CI $[0.36, 0.78]$. Note that the split-half correlation of accuracy (described in the results section of Experiment 1) was $r = 0.69$, 95% CI $[0.55, 0.80]$; thus, the empirical observer model was nearly at ceiling performance in predicting the level of agreement amongst participants. We emphasize here that because the empirical observer model is computed from participants responses to "will it fall?", our results show that people's judgments of relative mass can be predicted by their judgments of the towers' stabilities. This implies that there is a systematic relationship between the mechanism that people use both to make judgments of stability *and* judgments of relative mass.

Learning over multiple trials. Figure 5 shows a comparison of the human data to the fitted models under the IPE observer, and Table 2 lists numerical values for the log-likelihood ratios and Bayes factors. Negative values indicate evidence for the static model, and positive values indicate evidence for the learning model. In Experiment 1, the

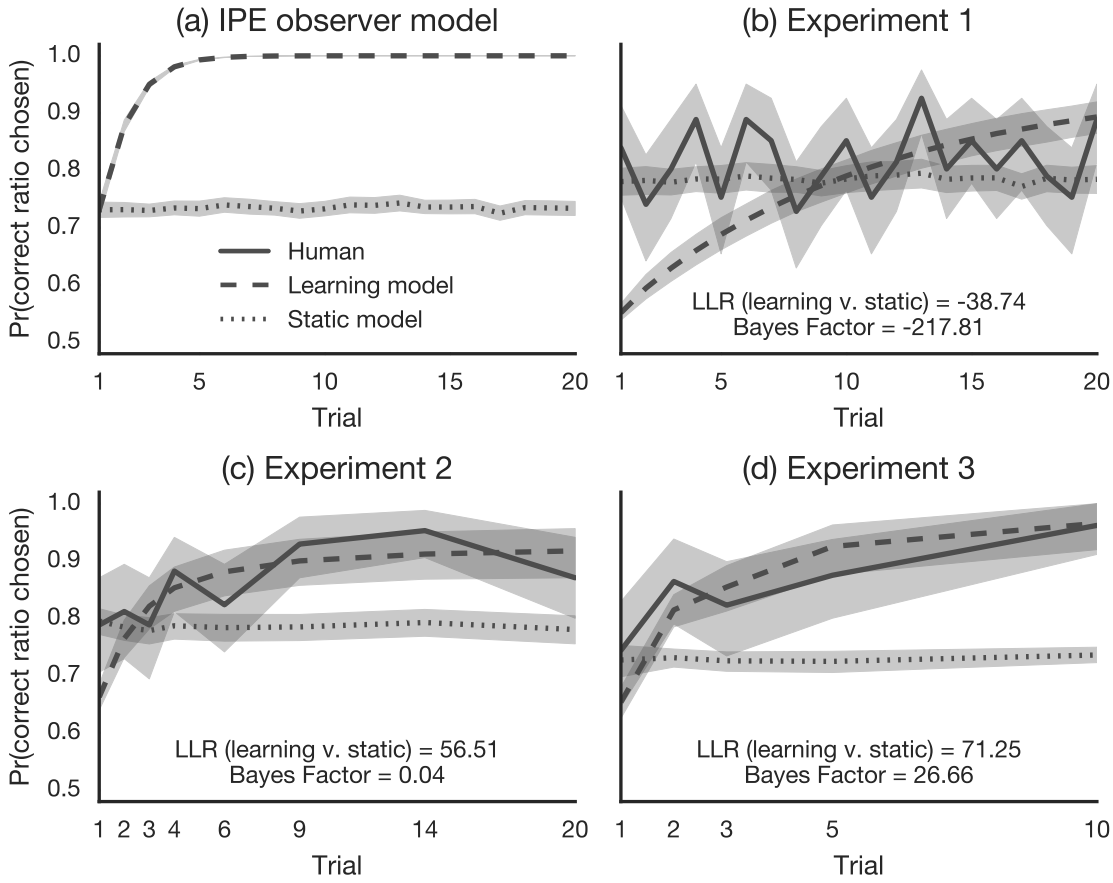

Figure 5. Model predictions of mass inferences. In all plots, dotted and dashed lines indicate the mean proportion of correct model responses computed for the IPE observer (responses for the empirical observer are nearly identical). Shaded regions are 95% confidence intervals of the mean. (a) The top left subplot shows the IPE learning and static observer models before being fit to human data (i.e., ideal observer predictions where $\beta = 1$ for all participants). (b) The static model is a better fit to the human data from Experiment 1 than the learning model. (c-d) The learning model is a better fit to human data from Experiments 2-3 than the static model.

static model was a better explanation for people's behavior than the learning model. In Experiment 2, the learning model was the better explanation of participant's behavior than the static model according to the log-likelihood, though not according to the Bayes factor.

Table 2

Log-likelihood ratios (LLR) and Bayes factors ($\log K$*) for learning vs. static models. Positive values favor the learning model, while negative values favor the static model.*

		Experiment 1	Experiment 2	Experiment 3 within subjs.	Experiment 3 between subjs.
LLR	IPE (fitted)	-38.74	56.51	43.42	71.25
	Empirical (fitted)	-70.79	60.02	41.77	60.78
$\log K$	IPE	-217.81	0.04	-0.52	26.66
	Empirical	-269.85	-6.58	-9.78	22.15

Similarly, if we look only at the condition in which participants responded on five trials in Experiment 3 (within subjects), the learning model was a better explanation of people's behavior according to the log-likelihood ratio, but not the Bayes factor. If we look between subjects in Experiment 3, using data from only the first response of each subject, we find that the learning model is a better explanation of people's behavior according to both measures. This result is consistent with the Spearman rank correlations reported in the results section of Experiment 3.

Why do the log-likelihood ratios and Bayes factors disagree on the results for Experiment 2 and within-subjects in Experiment 3? This disagreement illuminates the surprising results we obtained in Experiment 2: while the majority of participants did seem to take into account the evidence and learn over time, there was a minority of participants who seemed to be answering randomly. Figure 6 shows the distribution of fitted parameters for both the static and learning models (computed with the IPE likelihood) in all three experiments. The distributions resulting from using the empirical likelihood are very similar. Unsurprisingly, the parameters for the learning model in Experiment 1 are clustered around $\beta = 0$, indicating that the evidence is largely ignored. The parameters for the static model in Experiments 2-3 are biased towards $\beta > 1$, indicating that the evidence

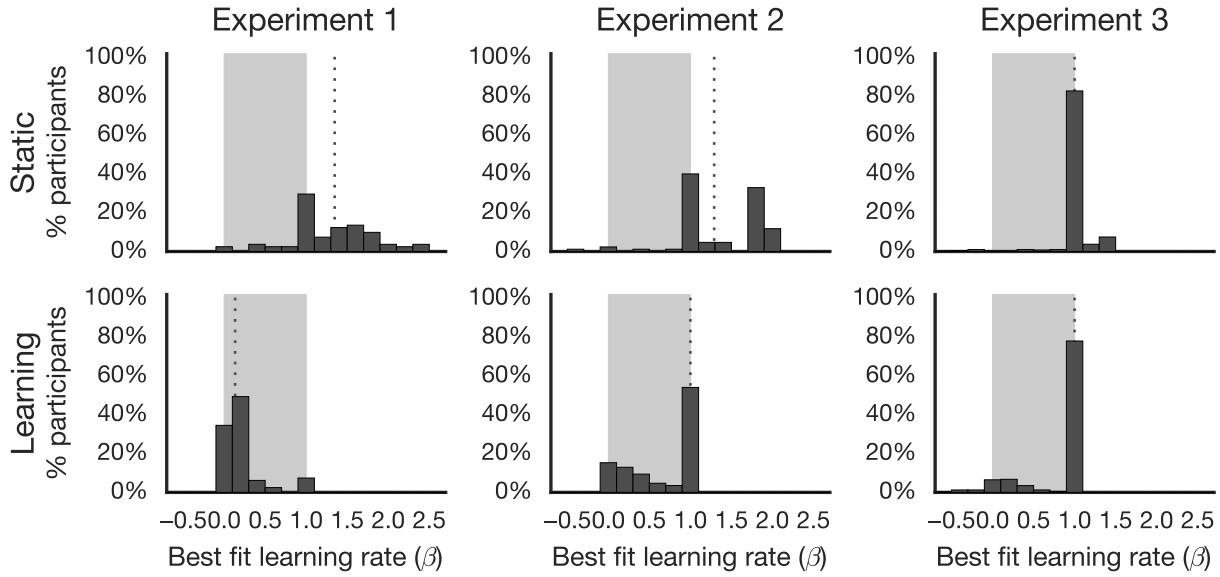

Figure 6. Distributions of parameter estimates. Each model was fit separately to each participant in each experiment. The top row shows the distributions for parameters fitted to the static model, while the bottom row shows the distributions for the learning model. The shaded gray area in each figure shows $0 < \beta < 1$, and the dotted lines show the medians of each histogram. If $\beta = 1$, the fitted model was exactly the same as the ideal observer model. If $\beta > 1$, the fitted model weighed evidence more strongly than the ideal observer. If $0 < \beta < 1$, the fitted model weighed evidence less strongly than the ideal observer. If $\beta = 0$, the fitted model did not take evidence into account at all. Finally, if $\beta < 0$, the fitted model took into account the *opposite* of what the evidence suggested.

is weighed more strongly (which is what we would expect if people were getting more accurate over time). The parameters for the learning model in Experiments 2-3 indicate a small fraction of participants who seemed to mostly ignore the evidence. In contrast, the majority of participants in Experiments 2-3 have $\beta = 1$, indicating that they are best fit to the ideal observer model.

Based on the histograms in Figure 6, we can understand the disagreement between the log-likelihood ratios and Bayes factors in Experiment 2 and within-subjects in

Experiment 3. When β is fit to each participant individually in the learning model, the resulting fits are close to zero for the participants who seemed to be ignoring the evidence. The behavior of these participants is close to random, and is therefore equally well explained by both models when $\beta \approx 0$. However, the majority of participants did actually learn over time, and had fitted coefficients of $\beta = 1$ for the learning model. The static model does a poor job of explaining the responses of these participants. Thus, overall the learning model is favored when we look at log-likelihood ratios resulting from fitted parameters.

The Bayes factors disagree with the log-likelihood ratios, however, because the fits under the learning model for people who did not learn ($\beta \approx 0$) have relatively low prior probability under the Laplace prior (with parameters $\mu = 1$ and $b = 1$). So, even though these coefficients result in higher log-likelihood ratios, they do not get much weight when the Bayes factors are computed. For the non-learning participants, the coefficients that do have higher prior probability yield responses that have low probability under the learning model and higher probability under the static model. This is enough to outweigh the fact that the learning model is a good model for people who were in fact learning, thus resulting in Bayes factors that favor the static model in Experiment 2 and within-subjects in Experiment 3.

General Discussion

We asked whether people can infer unobservable physical properties in complex scenes. We ran three experiments to answer this question, and proposed a new class of models that can capture this phenomenon by combining probabilities generated from simulations with Bayesian inference. In Experiment 1, we found that across participants and stimuli, 80% of the judgments of relative mass were correct, even though these inferences were informed by only a single trial's worth of information. This result contradicts previous studies suggesting that people are poor at inferring hidden properties like mass (e.g. Gilden & Proffitt, 1989a, 1989b) in all but the simplest of scenarios. If

anything, it suggests that people are remarkably *good* at inferring such properties. Moreover, Experiments 2-3 suggest that people can accumulate information and become increasingly fine-tuned to the properties of their environment the longer they observe it.

As illustrated by Figure 4b, we also found a systematic relationship between people's predictions about the future of a physical scene and their inferences about that scene's parameters. This suggests that people make predictions about whether the tower will fall under different possible parameter values in order to infer which parameter values were more likely to have produced the actual outcome they observed. This supports the idea that people rely on the same mechanism or source of physical knowledge both for predicting the future of physical scenes, and for inferring properties about objects in those scenes.

In the remainder of this article, we discuss questions for future research, as well as how these results relate to the larger literatures on physical reasoning.

Can people make more detailed inferences?

One question for future research regards whether people can make more detailed inferences beyond the ratios that were used in the present experiments. While within the range of objects that people normally interact with, the ratios of 10:1 and 1:10 that we used are relatively high. Can people distinguish between smaller mass ratios as well? Moreover, can people infer the *specific* ratio from a number of alternatives (e.g., 1:2 vs. 1:3 vs. 1:4)? We speculate that people could probably tell the difference between very small and very large mass ratios (e.g., 1:2 vs. 1:10) but have difficulty distinguishing between similar mass ratios (e.g., 1:2 vs. 1:3). In terms of whether people can infer a specific ratio, we suspect that in general people will be good at determining whether one type of object is heavier than other, but may have difficulty determining precisely by how much.

How do size and material properties influence peoples' inferences of mass?

In the present experiment, we asked participants to reason about the relative mass of objects. However, our stimuli consisted of blocks that were all the same size, and which were all subject to the same gravitational acceleration; thus, it could be that what we were

measuring was not people's ability to reason about mass but rather an ability to reason about density or weight. An important direction for future research will be to disentangle these confounding factors.

If people reason are sensitive to both mass and density, then their responses should influenced by factors such as size and material properties. There is ample evidence that people have strong expectations about physical properties based on their size and perceived material. Expectations about material densities are generally correct and lead to accurate predictions about the stability of objects (Lupo & Barnett-Cowan, 2015). Thus, we would expect people to be biased in their inferences if visual cues provide evidence for different sizes or materials.

It is not *a priori* clear in what direction people's inferences should be biased, however. For example, the size-weight illusion (SWI) is a well documented phenomena in which people perceive the smaller of two equally-weighted objects to be heavier after lifting them (Charpentier, 1891; Flanagan & Beltzner, 2000; Flanagan, Bittner, & Johansson, 2008; Grandy & Westwood, 2006; Murray, Ellis, Bandomir, & Ross, 1999; Ross, 1969). The explanation for this effect is that people expect the smaller object to be lighter than it actually is, and consequently perceive it to be heavier. Would visual evidence—rather than sensorimotor evidence—produce the same effect, in which people perceive smaller objects to be heavier when they are actually the same weight, because the visual evidence is surprising[5]? Or, would people's prior expectations simply override the evidence, causing them to perceive the smaller objects as lighter?

In a similar vein, the material-weight illusion (MWI) is another effect in which people perceive objects of heavier-looking materials (e.g., metal) to be lighter than objects of lighter-looking materials (e.g., styrofoam) even when those objects are actually the same

[5]For example, if the observer expects the smaller object to be lighter, then the observer would expect the smaller object to move faster than the larger object after a collision in which the initial velocities are the same. However, if they are actually the same weight, then they will move at the same speed after the collision.

weight (Buckingham, Ranger, & Goodale, 2011; Ellis & Lederman, 1999; Seashore, 1899; Wolfe, 1898). We can ask the same question: would this effect persist when people only have visual evidence, or would expectations override the visual evidence? Based on our experience with blocks of different materials (see Battaglia et al., 2013), we hypothesize that expectations would tend to overrule the visual evidence (i.e., that people would report objects with heavier-looking materials to be heavier, even when they are not).

Can these results be reconciled with "naïve physics" errors?

Many studies suggest that people's ability to reason about the physical world is relatively impoverished and error-prone. For example, some researchers have argued that people can reason about only a single dimension of physical information at once (e.g. the position of the object's center of mass over time) and have trouble incorporating multiple dimensions of information (e.g. mass as well as position) (Gilden & Proffitt, 1989b). Rather than accurately incorporating these multiple dimensions of information, Gilden and Proffitt (1989a) suggested that when reasoning about mass, people rely only on simple heuristics such as that the faster object after a collision is the lighter object. Todd and Warren (1982) also suggested that people rely on limited information (such as the final speeds of the objects) which is accurate in certain cases, but not in all situations (such as when elasticity is very high or low). While varying accounts of different heuristics and biases seem to conflict with each other in which effects they can predict, they can be reconciled by the noisy Newton hypothesis. Specifically, these effects can be explained as being the result of perceptual uncertainty (Sanborn et al., 2013; Sanborn, 2014). Moreover, given the results of the present paper, it seems unlikely that people only pay attention to a single dimension of information when making inferences about mass: in our experiments, participants must pay attention to the position, orientation, and color of multiple three-dimensional objects.

Other research has shown that people sometimes seem to rely on a naïve theory of "impetus", resulting in incorrect beliefs such as that if someone drops an object while they are walking, the object will fall straight down (rather than in a parabolic curve due to the

combination of horizontal and vertical velocity) (McCloskey, 1983; McCloskey, Washburn, & Felch, 1983). McCloskey et al. (1983) argued that this particular effect is due to a perceptual illusion involving the frame of reference of the motion (that is, that the object appears to move straight down with respect to the motion of the carrier). This interpretation is not inconsistent with the approximate simulation hypothesis, however: if approximate simulations are learned from perception, then perceptual illusions *should* affect the resulting dynamics models.

Other errors documented in the naïve physics literature may be a result of engaging different forms of physical knowledge. One classic error involves a pendulum task, in which a participants are asked to consider a pendulum consisting of a bob on a string. They are told that the string breaks, and are asked to draw the resulting trajectory of the bob. Participants' responses to this task tend to be inconsistent and often incorrect. However, Smith, Battaglia, and Vul (2013) gave people analogous tasks of *catching* the bob in a bucket or *cutting* the string such that the bob would go into a fixed bucket, and found that people were quite accurate in these tasks. When their responses did deviate from ground truth, they were strongly predicted by a noisy Newton simulation model.

Perhaps, then, certain tasks engage accurate predictive knowledge of object dynamics, while others engage more error-prone conceptual knowledge. This hypothesis has previously been suggested by Schwartz and Black (1999), who found a similar dissociation between explicit conceptual knowledge and implicit motor knowledge. Schwartz and Black (1999) asked participants to judge which of two glasses of water would need to tilt further before the water reached the rim of the cup. When given this judgment explicitly, participants were overwhelmingly incorrect; however, when asked to tilt an empty cup and *imagine* the water reaching the rim of the glass, participants tilted the cups to the appropriate angle. White (2012) further discusses the dissociation between the people's accuracy in tasks that seem to engage motor knowledge versus their inaccuracy in tasks that seem to only engage visual knowledge. Similarly, representational momentum

tasks seem to engage a type of physical knowledge that is dissociated from explicit formal knowledge of physics (Freyd & Jones, 1994; Kozhevnikov & Hegarty, 2001).

A pertinent question is: why do some tasks seem to invoke one system of knowledge, while others invoke another? It is not the case that people only use motor knowledge when the task calls for taking actions, and only visual knowledge when the task does not. For example, the classic mental rotation task by Shepard and Metzler (1971) is a good example of what seems like a purely perceptual task, yet that still engages the motor system (Parsons, 1994). One potential answer to this question is that the mind attempts to find a "perceptual match" to stored representations of actions on objects, and falls back on purely visual knowledge only when there is no motor representation to be found (White, 2012). An alternate hypothesis is that the mind engages in a metacognitive task of strategy selection (e.g. Lieder et al., 2014) and picks the strategy or domain of knowledge that has the higher expectation of being useful in the given situation.

How does approximate physical simulation relate to internal models for sensorimotor and perceptual prediction?

It is critical for the motor system to be able to accurately predict and respond to novel object dynamics; if it could not do so, then we would be unable to effectively interact with those objects. Indeed, there is a wealth of literature on motor and perceptual learning that suggests that people are quite sensitive to many dimensions of physical information and that they take this information into account in a reasonable way. There is evidence that the sensorimotor system learns both inverse models of control (i.e., what actions to perform to achieve a particular state) as well as forward models of interactions with the environment (i.e., what state to expect when an action is taken in the current state) (Flanagan, Vetter, Johansson, & Wolpert, 2003; Kawato, 1999; Wolpert & Kawato, 1998). These models incorporate dynamical information along multiple dimensions; for example, by including both gravity and mass, the motor system can predict the momentum of an oncoming object and appropriately contract the relevant muscles in order to catch it (Zago

& Lacquaniti, 2005).

The types of approximate physical simulations hypothesized in this paper and by Battaglia et al. (2013) are a type of forward dynamics model: given the current state, they predict the next state. Of course, it is not clear whether the forward models used by the motor system are exactly the same as those engaged in the higher-level cognitive tasks investigated here, especially given that our tasks did not involve a motor component. Given the evidence for multiple forward and inverse models in the motor system (Wolpert & Kawato, 1998), it seems likely that there could be additional forward models used by other aspects of cognition such as the perceptual system. Moreover, there is good evidence to believe that there is at least some degree of decoupling between the motor and perceptual systems. For example, illusions such as the size-weight illusion and material-weight illusion tend to persist, even after the motor system has adapted (Flanagan & Beltzner, 2000; Grandy & Westwood, 2006), and are influenced by top-down knowledge (Ellis & Lederman, 1998).

If the motor system is not directly involved in the approximate physical simulations explored in this paper, what is? An alternate explanation might be the perceptual forward models underlying an interesting class of phenomena known as *displacement* or *representational momentum* effects (Freyd & Finke, 1984; Hubbard, 2005). In these experiments, people's memory for the location of an object is distorted in the direction of implied motion: if someone sees an object moving towards the left, then they remember the object as being slightly more to the left than it actually was. These effects extend beyond just objects *in* motion but also to static objects that *could* move, such as an object that would fall due to gravity or be pushed upwards by a spring (Freyd, Pantzer, & Cheng, 1988). In fact, displacement effects have been found involving many types of physical information, including friction, linear velocity, centripetal force, barriers, and even top-down knowledge of properties such as mass (see Hubbard (2005) for a review).

These results from the literature on displacement suggest that the perceptual system

has a rich and detailed knowledge of how objects behave. Yet, it also appears that displacement effects are more consistent with the impetus theory of physical reasoning than with Newtonian physics. For example, there are greater displacements for objects moving up when the objects are small rather than when they are large (Kozhevnikov & Hegarty, 2001); this is similar to the Aristotelean belief that heavier objects fall faster than lighter objects. Similarly, for objects moving in a curved path, there are displacements in the direction of centripetal force (Freyd & Jones, 1994; Hubbard, 1996); this result is similar to the naïve belief that objects moving in a curved tube will continue moving in a curved path after exiting the tube (McCloskey & Kohl, 1983). Other evidence similarly suggests the presence of impetus principles underlying displacement effects in Michotte-type launching experiments (Hubbard, 2004, 2013a, 2013c, 2013b; Hubbard & Ruppel, 2002).

There are two hypotheses that come to mind in reconciling approximate physical simulation with displacement. The first hypothesis is that approximate physical simulation is a different cognitive process from that underlying displacement. This argument is plausible, though lacks in parsimony as it requires positing that the mind has two forward models for the same physical phenomena. The second hypothesis is that approximate physical simulation and displacement are related, and that the impetus effects arise naturally from learning forward dynamics from noisy data. Whether either of these hypotheses is correct is an open question for future research.

Is simulation too computationally intensive?

The noisy Newton hypothesis is largely posited at the computational-level of analysis (Marr, 1982), while approximate physical simulation is an algorithmic-level solution to the computational-level problem. As an algorithmic-level model, approximate physical simulation comes with several dimensions of computational constraints that allow us to form testable hypotheses about how it is used. One question we can ask is: given the ability to run approximate physical simulations, how many simulations should be run? In particular, one critique of approximate physical simulation is that it is too computationally

intensive to be plausible as a cognitive mechanism (Davis & Marcus, 2014). It seems unlikely that people run hundreds or even tens of simulations per decision; a more realistic hypothesis is that people run just one or two simulations per decision.

Battaglia et al. (2013) performed an analysis of the standard deviation of participants' "will it fall?" judgments to estimate the number of simulations. We perform a similar analysis here. If people take n samples per judgment, then the variance of their responses should be:

$$\sigma^2_{\text{judgments}} = \frac{\sigma^2_{\text{sims}}}{n} + \omega^2, \tag{10}$$

where σ_{sims} is the standard deviation of simulations from the IPE and ω is the standard deviation of other sources of uncertainty (such as general decision-making noise). We used a least-squares linear regression to estimate ω in Equation 10 for each of $n \in \{1, 2, 3, 4, 5, 6\}$, and from each regression computed the mean squared error between predicted variance and the actual variance of human judgments. Figure 7 shows the variance of participant responses as a function of the variance of IPE samples. The dotted lines correspond to the predicted variance for each value of n. The solid line has the lowest MSE, and corresponds to $n = 1$ and $\omega = 0.25$, indicating that the variance of participants' judgments is consistent with them running one simulation per judgment. This result is consistent with research suggesting that it is optimal to only take a small number of samples before making a decision (Vul, Goodman, Griffiths, & Tenenbaum, 2014), and that this is in fact what people do in the case of physical prediction (Hamrick et al., 2015).

Running a single simulation per judgment seems much more tractable than, say, ten simulations. However, there are additional questions regarding computational limitations beyond just the number of simulations: for example, it seems unlikely that people would be able to run a detailed simulation of ten objects, especially given that people can only track a small number of objects simultaneously (Pylyshyn & Storm, 1988). While it is beyond the scope of this paper to answer this question, we emphasize that our hypothesized

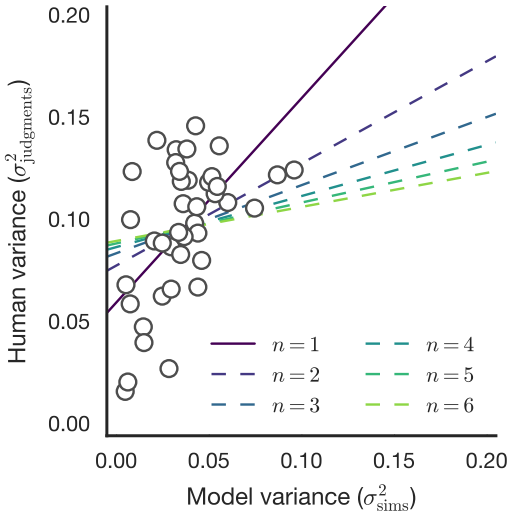

Figure 7. Analysis of human variance. This plot shows the variance of human predictions as a function of the variance of model predictions computed from different numbers of samples (n). The solid line indicates the fit with the lowest mean squared error, corresponding to $n = 1$.

simulations are *approximate* and thus may sacrifice accuracy for performance[6]. How this trade-off actually manifests is an important question that we hope to address in the future.

While we have argued that there are situations in which simulation is computationally tractable—such as the prediction and inference tasks presented in this paper—there are certainly cases where simulation is always going to be too expensive. For example, Smith, Dechter, et al. (2013) describe a task in which participants have to predict whether a ball will reach a green target or a red target first. Interestingly, while some of their stimuli seem

[6]We distinguish our approximate simulations from heuristics in that heuristics are typically thought of as simple rules that tend to do a good job, but which are not necessarily approximating an optimal solution. Although some research suggests that some heuristics may indeed be approximating an optimal solution (Lieder, Griffiths, & Goodman, 2012; Lieder et al., 2014; Tenenbaum & Griffiths, 2001), this has not been demonstrated for *all* heuristics, and thus we feel it is more informative to label our approach as an approximation to an optimal solution rather than as a heuristic.

to lend themselves to a physical simulation, others clearly do not (such as if the ball is in a box with the green target in the box and with the red target outside of the box). In such cases, it may be that conceptual or qualitative knowledge is more appropriate (Forbus, 1983, 2011). How the mind decides between these approaches could perhaps be thought of as another instance of metacognitive strategy selection (e.g. Lieder et al., 2014).

Conclusion

In sum, we found that people can learn the relative masses of objects after observing their interactions in complex scenes. Our results both confirm recent reports that people's physical scene understanding is driven by probabilistic physical simulation, and provide additional evidence that the *same* mechanism used in making predictions is also used when making inferences about underlying physical parameters. Beyond simply making one-shot inferences, our results suggest how mental simulation as a cognitive resource can also be used to aggregate evidence and learn about properties of the world over time. Mental simulation thus truly offers an "infinite use of finite means" by simultaneously serving the needs of prediction, inference, and learning. Despite this, mental simulation may not always be the most appropriate strategy as it requires much time and effort. Key questions for the future are, then: how do people decide when to use mental simulation, and how do they determine which simulations to run?

Author Contributions

All authors designed the research. J. B. Hamrick and P. W. Battaglia wrote the modeling code, and J. B. Hamrick ran the experiments and analyzed the data. J. B. Hamrick wrote the paper and P. W. Battaglia, T. L. Griffiths, and J. B. Tenenbaum provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

This work was supported by a Berkeley Fellowship and NSF Graduate Fellowship awarded to J. B. Hamrick, as well as grant number N00014-13-1-0341 from the Office of

Naval Research. We would also like to thank Michael Pacer and Tobias Pfaff for helpful discussions and feedback.

References

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 of the United States of America*, *110*(45), 18327–18332. doi: 10.1073/pnas.1306572110

Baugh, L. A., Kao, M., Johansson, R. S., & Flanagan, J. R. (2012). Material evidence: interaction of well-learned priors and sensorimotor memory when lifting objects. *Journal of Neurophysiology*, *108*(5), 1262–1269. doi: 10.1152/jn.00263.2012

Buckingham, G., Ranger, N. S., & Goodale, M. A. (2011). The material-weight illusion induced by expectations alone. *Attention, Perception & Psychophysics*, *73*(1), 36–41. doi: 10.3758/s13414-010-0007-4

Bullet Collision Detection and Physics Library [Computer software manual]. (2013). Retrieved from `http://www.bulletphysics.org/`

Charpentier, A. (1891). Analyse experimentale quelques elements de la sensation de poids [Experimental study of some aspects of weight perception]. *Archives de Physiologie Normales Pathologiques*, *3*, 122–135.

Davis, E., & Marcus, G. F. (2014). The Scope and Limits of Simulation in Cognitive Models. *arXiv:1506.04956 [cs.AI]*, 1–27.

Ellis, R. R., & Lederman, S. J. (1998). The golf-ball illusion: evidence for top-down processing in weight perception. *Perception*, *27*(1917), 193–201. doi: 10.1068/p270193

Ellis, R. R., & Lederman, S. J. (1999). The material-weight illusion revisited. *Perception & Psychophysics*, *61*(8), 1564–1576. doi: 10.3758/BF03213118

Flanagan, J. R., & Beltzner, M. A. (2000). Independence of perceptual and sensorimotor predictions in the size-weight illusion. *Nature Neuroscience*, *3*(7), 737–741. doi: 10.1038/76701

Flanagan, J. R., Bittner, J. P., & Johansson, R. S. (2008). Experience can change distinct size-weight priors engaged in lifting objects and judging their weights. *Current*

Biology, *18*(22), 1742–1747. doi: 10.1016/j.cub.2008.09.042

Flanagan, J. R., Vetter, P., Johansson, R. S., & Wolpert, D. M. (2003). Prediction precedes control in motor learning. *Current Biology*, *13*(2), 146–150. doi: 10.1016/S0960-9822(03)00007-1

Forbus, K. D. (1983). Qualitative Reasoning About Space and Motion. In *Mental models* (pp. 53–73).

Forbus, K. D. (2011). Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(4), 374–391. doi: 10.1002/wcs.115

Freyd, J. J., & Finke, R. A. (1984). Representational Momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 126–132. doi: 10.1037/0278-7393.10.1.126

Freyd, J. J., & Jones, K. T. (1994). Representational Momentum for a Spiral Path. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 968–976. doi: 10.1037/0278-7393.20.4.968

Freyd, J. J., Pantzer, T. M., & Cheng, J. L. (1988). Representing Statics as Forces in Equilibrium. *Journal of Experimental Psychology: General*, *117*, 395–407. doi: 10.1037/0096-3445.117.4.395

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (pp. 1–6).

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 1–6).

Gilden, D. L., & Proffitt, D. R. (1989a). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 372–383. doi: 10.1037/0096-1523.15.2.372

Gilden, D. L., & Proffitt, D. R. (1989b). Understanding natural dynamics. *Journal of*

Experimental Psychology: Human Perception and Performance, *15*(2), 384–393. doi: 10.1037/0096-1523.15.2.384

Gilden, D. L., & Proffitt, D. R. (1994). Heuristic judgement of mass ratio in two-body collisions. *Perception and Psychophysics*, *56*(6), 708–720. doi: 10.3758/BF03208364

Grandy, M. S., & Westwood, D. A. (2006). Opposite Perceptual and Sensorimotor Responses to a Size-Weight Illusion. *Journal of Neurophysiology*, *95*(6), 3887–3892. doi: 10.1152/jn.00851.2005

Gureckis, T. M., Martin, J., McDonnell, J., Alexander, R. S., Markant, D. B., Coenen, A., ... Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*, 2–16. doi: 10.3758/s13428-015-0642-8

Hamrick, J. B., Battaglia, P. W., & Tenenbaum, J. B. (2011). Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1–6).

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 Conference of the Cognitive Science Society* (pp. 1–6).

Hubbard, T. L. (1996). Representational momentum, centripetal force, and curvilinear impetus. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(4), 1049–1060. doi: 10.1037/0278-7393.22.4.1049

Hubbard, T. L. (1997). Target size and displacement along the axis of implied gravitational attraction: Effects of implied weight and evidence of representational gravity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(6), 1484–1493. doi: 10.1037/0278-7393.23.6.1484

Hubbard, T. L. (2004). The Perception of Causality: Insights from Michotte's Launching Effect, Naïve Impetus Theory, and Representational Momentum. In A. M. Oliveira, M. Teixeira, G. F. Borges, & M. J. Ferro (Eds.), *Fechner Day* (pp. 116–121).

Coimbra, Portugal: The International Society for Psychophysics.

Hubbard, T. L. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review*, *12*(5), 822–851. doi: 10.3758/BF03196775

Hubbard, T. L. (2013a). Launching, Entraining, and Representational Momentum: Evidence Consistent with an Impetus Heuristic in Perception of Causality. *Axiomathes*, *23*(4), 633–643. doi: 10.1007/s10516-012-9186-z

Hubbard, T. L. (2013b). Phenomenal Causality II: Integration and Implication. *Axiomathes*, *23*, 485–524. doi: 10.1007/s10516-012-9200-5

Hubbard, T. L. (2013c). Phenomenal Causality I: Varieties and Variables. *Axiomathes*, *23*, 1–42. doi: 10.1007/s10516-012-9198-8

Hubbard, T. L., & Ruppel, S. E. (2002). A possible role of naïve impetus in Michotte's "launching effect": Evidence from representational momentum. *Visual Cognition*, *9*(1-2), 153–176. doi: 10.1080/13506280143000377

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, *9*(3), 90–95.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinions in Neurobiology*, *9*(6), 718–727. doi: 10.1016/S0959-4388(99)00028-8

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

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems*, *25*, 1–9.

Lieder, F., Hamrick, J. B., Hay, N. J., Plunkett, D., Russell, S. J., & Griffiths, T. L. (2014). Algorithm selection by rational metareasoning as a model of human strategy

selection. *Advances in Neural Information Processing Systems*, *27*, 2870–2878.

Lupo, J., & Barnett-Cowan, M. (2015). Perceived object stability depends on shape and material properties. *Vision Research*, *109*, 158–165. doi: 10.1016/j.visres.2014.11.004

Marr, D. (1982). The Philosophy and the Approach. In *Vision* (pp. 8–37).

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–130. doi: 10.1038/scientificamerican0483-122

McCloskey, M., & Kohl, D. (1983). Naive physics: the curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 146–156. doi: 10.1037/0278-7393.9.1.146

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive Physics: The Straight-Down Belief and Its Origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636–649. doi: 10.1037/0278-7393.9.4.636

McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56).

Murray, D. J., Ellis, R. R., Bandomir, C. A., & Ross, H. E. (1999). Charpentier (1891) on the size-weight illusion. *Perception & Psychophysics*, *61*(8), 1681–1685. doi: 10.3758/BF03213127

Panda3D v1.9.0 [Computer software manual]. (2013). Retrieved from `https://www.panda3d.org/`

Parsons, L. M. (1994). Temporal and kinematic properties of motor behavior reflected in mentally simulated action. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 709–730. doi: 10.1037/0096-1523.20.4.709

Pérez, F., & Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, *9*(3), 21–29. doi: 10.1109/MCSE.2007.53

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence

for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197. doi: 10.1163/156856888X00122

Ross, H. E. (1969). When is a weight not illusory? *The Quarterly Journal of Experimental Psychology*, *21*(4), 346–355. doi: 10.1080/14640746908400230

Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: Cue heuristics versus direct-perceptual competence. *Psychological Review*, *107*(3), 525–555. doi: 10.1037/0033-295X.107.3.525

Sanborn, A. N. (2014). Testing Bayesian and heuristic predictions of mass judgments of colliding objects. *Frontiers in Psychology*, *5*(938), 1–7. doi: 10.3389/fpsyg.2014.00938

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2009). A Bayesian framework for modeling intuitive dynamics. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1–6).

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 1–77. doi: 10.1037/a0031912

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 116–136. doi: 10.1037/0278-7393.25.1.116

Seashore, C. E. (1899). Some psychological statistics. 2. The material-weight illusion. *University of Iowa Studies in Psychology*.

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, *171*(3972), 701–703. doi: 10.1126/science.171.3972.701

Smith, K. A., Battaglia, P. W., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1–6).

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over

time. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1–6).

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). The Rational Basis of Representativeness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1–6).

Todd, J. T., & Warren, W. H. (1982). Visual perception of relative mass in dynamic events. *Perception*, *11*(3), 325–335. doi: 10.1068/p110325

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2014). Learning physics from dynamical scenes. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 1–6).

van der Walt, S., Colbert, S., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, *13*(2), 22–30. doi: 10.1109/MCSE.2011.37

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and Done? Optimal Decisions From Very Few Samples. *Cognitive Science*, *38*(4), 599–637. doi: 10.1111/cogs.12101

Waskom, M., Botvinnik, O., Hobson, P., Cole, J. B., Halchenko, Y., Hoyer, S., ... Allan, D. (2014). Seaborn v0.5.0 [Computer software manual]. doi: 10.5281/zenodo.12710

White, P. A. (2012). The experience of force: The role of haptic experience of forces in visual perception of object motion and interactions, mental simulation, and motion-related judgments. *Psychological Bulletin*, *138*(4), 589–615. doi: 10.1037/a0025587

Wolfe, H. K. (1898). Some effects of size on judgments of weight. *Psychological Review*, *5*(1), 25–54. doi: 10.1037/h0073342

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for

motor control. *Neural Networks*, *11*, 1317–1329. doi: 10.1016/S0893-6080(98)00066-5

Zago, M., & Lacquaniti, F. (2005). Visual perception and interception of falling objects: a review of evidence for an internal model of gravity. *Journal of Neural Engineering*, *2*(3), S198–208. doi: 10.1088/1741-2560/2/3/S04

Appendix A

Tower Stimuli

Construction

Subjects were presented with virtual, 3D tower scenes such as those in Figure 1b-c, which contained 10 rectangular blocks each with dimensions of 20cm×20cm×60cm. The blocks were stacked within a 120cm×120cm square column by a sequential random process such that when placed on the tower, no single block would fall off of its support (although, when later blocks were added, they might cause previously placed blocks to fall). This construction is the same as that used by Battaglia et al. (2013).

Selection

To choose the tower stimuli, we began with a set of 270 randomly generated geometries as well as the 30 tower geometries used in Experiment 3 from Battaglia et al. (2013). From each of these base towers, we generated 20 different versions in each of which 5 blocks were randomly assigned a label of "A" and 5 blocks were randomly assigned a label of "B". We ran model simulations for each of these 6000 towers under two mass ratios, $\kappa = 0.1$ and $\kappa = 10$, where the mass ratio is defined as the the ratio between the masses of "A" and "B" blocks.

For each stimulus $S_{i,j}$ (where $S_{i,j}$ is the j^{th} version of the i^{th} base tower), we computed whether the tower fell (F) under the true mass ratio. Then, according to Equation 4, we computed for both $\kappa = 0.1$ and $\kappa = 10$:

$$p(F|S_{i,j}, \kappa = k) \tag{11}$$

$$p(F|S_{i,j}, \kappa \neq k). \tag{12}$$

Equation 11 is the likelihood given the *correct* hypothesis, while Equation 12 is the likelihood given the *incorrect* hypothesis. If Equation 12 is greater than Equation 11 for a particular $S_{i,j}$, then our model will believe the *wrong* hypothesis. Based on earlier pilot data, people appear to be sensitive to this effect; thus, we excluded towers from

consideration if Equation 12 was greater than Equation 11. Finally, we chose towers that maximized the likelihood ratio

$$\mathrm{LHR}(S_{i,j}, k) := \frac{p(F|S_{i,j}, \kappa = k)}{p(F|S_{i,j}, \kappa \neq k)} \tag{13}$$

for both $\kappa = 0.1$ and $\kappa = 10$. To do this, we found the best version S_i^* of each base stimulus:

$$S_i^* = \arg\max_{S_{i,j}} \ \mathrm{LHR}(S_{i,j}, \kappa = 0.1) \cdot \mathrm{LHR}(S_{i,j}, \kappa = 10) \tag{14}$$

and then chose the top 20 of those.

Rendering

Videos of the towers were rendered using "Panda3D v1.9.0" (2013). In the training and prediction phases of Experiment 1, we rendered towers with red and blue blocks. In the inference phase of Experiment 1, we rendered towers with the following pairs of colors: blue-green, orange-blue, blue-yellow, gray-cyan, cyan-magenta, orange-cyan, cyan-purple, red-cyan, cyan-yellow, green-gray, gray-magenta, purple-gray, gray-red, magenta-green, green-orange, purple-green, magenta-yellow, orange-purple, yellow-purple, yellow-red. In the inference phases of Experiments 2 and 3, the towers were always rendered with purple and green blocks.

Appendix B

Fitting IPE Model Parameters

To obtain estimates from the IPE model, we ran 100 model samples for each stimulus for each of the parameter settings of

$\sigma \in \{0.0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1\}$ and

$\phi \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.6, 2.0, 2.5, 3.0\}$, where σ represents perceptual uncertainty, and ϕ represents force uncertainty. We then computed correlations between people's responses to "will it fall?" and the model's responses for each one of these parameter combinations (Figure B1). The parameters that best fit people's responses to "will it fall?" were $\sigma = 0.04$ and $\phi = 0.0$. Battaglia et al. (2013) found $\sigma = 0.04$ and $\phi = 0.2$ to be the best fitting parameters; because these are close to our best fit parameters, we opted to perform all our analyses with $\sigma = 0.04$ and $\phi = 0.2$.

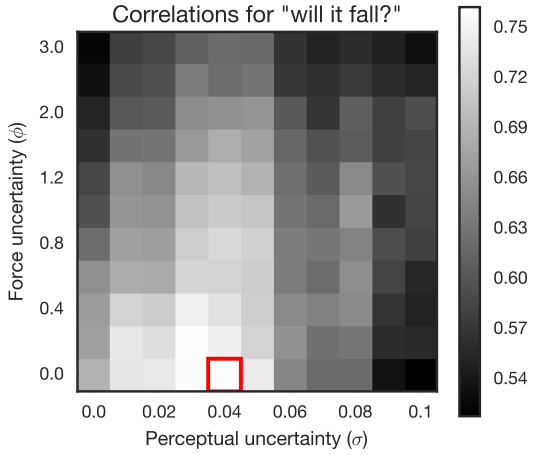

Figure B1. Correlation as a function of model parameters. We queried the IPE model for "will it fall?" responses for different settings of its parameters, finding a similar pattern as that from Battaglia et al. (2013).

Table C1

Log-likelihood ratios and Bayes factors for different query types.

		Experiment 1	Experiment 2	Experiment 3 within subjs.	Experiment 3 between subjs.
LLR	At least one	34.50	77.35	49.22	63.28
	More than half	-32.39	46.70	27.91	26.29
	Percent	-38.74	56.51	43.42	71.25
$\log K$	At least one	-163.41	9.51	0.56	26.72
	More than half	-247.21	-41.23	-25.44	11.89
	Percent	-217.81	0.04	-0.52	26.66

Appendix C

IPE Model Queries

What does it mean for a tower to "fall down"? One interpretation is that at least one block moved:

$$p(F_t|S_t, \kappa = k) \approx \frac{(\sum_{i=1}^{N} I_{b^{(i)}>0)}) + 0.5}{N+1}, \tag{15}$$

where $b^{(i)}$ is the proportion of blocks that moved during the i^{th} simulation, and where $I_{b^{(i)}>0}$ is one when at least one block has moved, and zero otherwise.

Another interpretation is that several blocks moved:

$$p(F_t|S_t, \kappa = k) \approx \frac{(\sum_{i=1}^{N} I_{b^{(i)}>0.5}) + 0.5}{N+1}, \tag{16}$$

where $I_{b^{(i)}>0.5}$ is one when more than half the blocks moved, and zero otherwise.

Yet another interpretation is that the tower falls "more" as the number of blocks that moved increases:

$$p(F_t|S_t, \kappa = k) \approx \frac{1}{N} \sum_{i=1}^{N} b^{(i)}, \tag{17}$$

where as before $b^{(i)}$ is the proportion of blocks that moved during the i^{th} simulation.

The "queries" presented in the previous paragraph are all plausible ways that people

compute the answer to "will it fall?". Looking at correlations between people's responses to "will it fall?" in Experiments 1-2 and model responses, we found that the "at least one" query had a correlation of $r = 0.62$, 95% CI $[0.48, 0.75]$; the "more than half" query had a correlation of $r = 0.73$, 95% CI $[0.55, 0.85]$; and the "percent" query had a correlation of $r = 0.75$, 95% CI $[0.57, 0.87]$.

All of the queries also do a good job at predicting people's inferences in Experiment 1. The IPE observer model using the "at least one" query agreed with participants' inferences of which color was heavier 78.5% of the time. The correlation coefficient between the model's probability of choosing $\kappa = 10$ and the proportion of people that chose $\kappa = 10$ was $r = 0.63$, 95% CI $[0.47, 0.75]$. For the "more than half" query, the model agreed with people 81.9% of the time, and the correlation coefficient was $r = 0.89$, 95% CI $[0.83, 0.93]$. For the "percent" query, the model agreed with people 81.9% of the time, and the correlation coefficient was $r = 0.91$, 95% CI $[0.87, 0.94]$.

If we fit the IPE learning and static observer models with each of the queries to participants' data, we also find similar log-likelihoods for each of the queries, with the exception of the "at least one" query for Experiment 1 (Table C1). The Bayes factors for each query are also consistent, including the "at least one" query for Experiment 1. Based on these results, we suspect the anomalous log-likelihood ratio is due to overfitting, which the Bayes factors do not suffer from.

Appendix D

Derivation of the Counterfactual Likelihood

Here we derive Equation 9. First, we have two binary variables: the mass ratio (κ), and whether the tower falls or not (F). Based on how the towers were constructed, we know that if $F = 1$ and $\kappa = k$, then $F = 0$ and $\kappa \neq k$ (and vice versa; if $F = 0$ and $\kappa = k$, then $F = 1$ and $\kappa \neq k$). However, our simulation model only allows us to sample independently from the marginal probability distributions $p(F_{\kappa=k})$ and $p(F_{\kappa\neq k})$, where $F_{\kappa=k}$ means the feedback obtained when $\kappa = k$ and $F_{\kappa\neq k}$ means the feedback obtained when $\kappa \neq k$ Thus, our joint samples will follow the proposal distribution:

$$g(F_{\kappa=k}, F_{\kappa\neq k}) = p(F_t|S_t, \kappa = k)p(F_t|S_t, \kappa \neq k).$$

The acceptance probability is:

$$q(F_{\kappa=k}, F_{\kappa\neq k}) = \begin{cases} 1, & F_{\kappa=k} \neq F_{\kappa\neq k} \\ 0, & F_{\kappa=k} = F_{\kappa\neq k} \end{cases}$$

According to the definition of rejection sampling, the true value of the joint distribution $p(F_{\kappa=k}, F_{\kappa\neq k})$ is:

$$p(F_{\kappa=k}, F_{\kappa\neq k}) \propto q(F_{\kappa=k}, F_{\kappa\neq k})g(F_{\kappa=k}, F_{\kappa\neq k}).$$

As in the main text, letting $\mathcal{L}(\kappa) := p(F_t = 1|S_t, \kappa = k) = 1 - p(F_t = 0|S_t, \kappa \neq k)$, we can expand this out to get the full distribution:

$$\begin{aligned} p(F_{\kappa=k} = 0, F_{\kappa\neq k} = 0) &= 0, \\ p(F_{\kappa=k} = 1, F_{\kappa\neq k} = 1) &= 0, \\ p(F_{\kappa=k} = 1, F_{\kappa\neq k} = 0) &= \frac{1}{Z} \cdot \mathcal{L}(\kappa)\left(1 - \mathcal{L}(\bar{\kappa})\right), \\ p(F_{\kappa=k} = 0, F_{\kappa\neq k} = 1) &= \frac{1}{Z} \cdot \left(1 - \mathcal{L}(\kappa)\right)\mathcal{L}(\bar{\kappa}), \end{aligned}$$

where the normalization constant is:

$$Z = \mathcal{L}(\kappa)\left(1 - \mathcal{L}(\bar{\kappa})\right) + \left(1 - \mathcal{L}(\kappa)\right)\mathcal{L}(\bar{\kappa}).$$

Appendix E

Software

The tower stimuli were rendered using the Panda3D video game framework ("Panda3D v1.9.0", 2013), and were simulated using the Bullet physics engine ("Bullet Collision Detection and Physics Library", 2013). All analyses were performed in the Python programming language using the following scientific computing libraries: NumPy and SciPy (van der Walt, Colbert, & Varoquaux, 2011), Pandas (McKinney, 2010), and IPython (Pérez & Granger, 2007). Figures were generated with Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2014).