Resource bounds on mental simulations: Evidence from a liquid-reasoning task

YingQiao Wang [1] and Tomer D. Ullman[1,*]

[1]Department of Psychology, Harvard University, Cambridge MA 02138

[*]Corresponding Author, tullman@fas.harvard.edu

Author Note

Tomer D. Ullman https://orcid.org/0000-0003-1722-2382.

The authors wish to thank the members of the Computation, Cognition, and Development lab for helpful discussions. TDU was supported by an NSF Science Technology Center Award CCF-1231216, the DARPA Machine Common Sense program, and the Jacobs Foundation. All data and stimuli are available at the following OSF repository: https://osf.io/pvab3/. All studies, including experimental procedures, number of participants, exclusion criteria, and analyses, were pre-registered, and see https://osf.io/pvab3/registrations. The authors declare no conflicts of interest.

An earlier version of this manuscript was shared on PsyArxiv at https://osf.io/preprints/psyarxiv/rf367.

YW and TDU conceived and planned the experiments. YW carried out the experiments. YW and TDU analyzed and interpreted the results. YW and TDU wrote the manuscript and edited it.

Correspondence should be addressed to Tomer Ullman, Harvard University, Cambridge, MA, USA; tullman@fas.harvard.edu.

Abstract

People are able to reason about the physical dynamics of everyday objects. But, there are theoretical disagreements about the computations that underlie this ability. One proposal is that people are running an approximate mental simulation of their environment. However, such a simulation must be limited in its resources. We applied the notion of a resource-bound simulation to a task of reasoning about liquids, and show that people's changing behavior can be explained by an approximate simulation that hits a resource limit after some time elapses. In Experiments 1 and 2, people performed well on tasks that asked them to estimate the time-to-fill and water level of different containers, when filled over short periods of time (1-7 seconds). Experiment 3 shows systematic biases in visual volume estimation, which further strengthens the proposal that people are using a simulation to solve the first two experiments. Experiment 4 extends the reasoning time for the time-to-fill task, and shows the existence of a 'switch point', as expected from a resource-bound simulation model. The model also accounts for individual differences: People who perform worse on a digit-span task have an earlier switch point. Our work argues for the theoretical proposal that people are using mental simulations to reason about intuitive physics, but further informs the suggestion that these simulations are limited in resources.

Public Significance Statement Our studies find that people are likely mentally simulating a scene when they imagine how fluids and containers interact. This mental simulation works reasonably well over short periods of time, but runs into resource issues over longer periods of time.

Keywords: intuitive physics, mental simulation, fluid, liquid, resource-bounded rationality

Resource bounds on mental simulations: Evidence from a liquid-reasoning task

Introduction

People are reasonable at reasoning about the behavior of everyday physical objects. In a typical day, a person may duck an errant Frisbee, spread mayo on a sandwich as intended, or fill a bottle without splashing water everywhere. Such humdrum activities seem unexciting, but only because people are so good at reasoning about typical movement and flow. Without a sense of intuitive physics, every day would be a series of small disasters.

One way to recognize how remarkable people are at intuitive physics is to consider how remarkably *bad* current machine intelligence is at it. This is despite it being a central pillar for more human-like machine intelligence (Lake, Ullman, Tenenbaum, & Gershman, 2017). While serious and significant progress has been made over the past few decades on artificial commonsense physical reasoning (For some recent examples, see Bear et al., 2021; Piloto, Weinstein, Battaglia, & Botvinick, 2022; Smith et al., 2019), no machine can currently interact with the physical world as generally, effortlessly, and flexibly as a person. We point out the current gap between human and artificial intelligence when it comes to intuitive physics not to heckle, but to illustrate that intuitive physical reasoning rests on non-obvious computations, and that we still do not fully know what those computations are.

There is an ongoing and lively debate about which model or theory best accounts for people's intuitive theories. One prominent model of the computations people may be using which has been put forth in the last decade is that people are running a kind of probabilistic mental simulation (see e.g. Battaglia, Hamrick, & Tenenbaum, 2013; Smith, Dechter, Tenenbaum, & Vul, 2013), unfolding object trajectories in their mind. The 'mental physics engine' has been used to explain people's intuitive physics in a variety of behavioral tasks (e.g. Allen, Smith, & Tenenbaum, 2020; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018). This proposal

has also been used to explain findings in cognitive development (Ullman & Tenenbaum, 2020), and has found support in neuroscience (Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016; Schwettmann, Tenenbaum, & Kanwisher, 2019). Here, we take the mental physics engine proposal seriously, but also take seriously its critics.

The mental physics engine is not the only hypothesis for explaining people's physical reasoning. Earlier work on intuitive physics suggested it was based on heuristics or rules (Gilden & Proffitt, 1989; Todd & Warren Jr, 1982), qualitative reasoning (Forbus, 1997), or possibly pre-Newtonian intuitive theories (McCloskey, Caramazza, & Green, 1980). More recently, researchers have put forward proposals for intuitive physics based on deep neural networks that learn to non-linearly combine a large number of features in order to predict everyday physical outcomes (Lerer, Gross, & Fergus, 2016; Piloto et al., 2022; Riochet et al., 2018).

Regardless of alternative theories, several researchers have pointed out potential issues with the mental physics engine (Davis & Marcus, 2016; Ludwin-Peery, Bramley, Davis, & Gureckis, 2021, 2020; Marcus & Davis, 2013). One issue raised by this past work is that a true-to-life, fully detailed probabilistic simulation of an everyday scene would mean holding in mind an extremely large number of objects and interactions, and calculating the many interactions between them, in a fine-grained way. This is taken to be very computationally expensive, and at odds with the current understanding of biology and psychology. A second issue focuses on the empirical fact that people consistently show inconsistent physical reasoning in certain situations. For example, people in some circumstances show a 'physical conjunction error' Ludwin-Peery et al. (2020), where they report that the conjunction of two physical events (e.g. one object hitting another, and the second object landing on a certain spot) is more likely than one of the events (e.g. the second object landing on a certain spot). This is akin to the classic 'conjunction fallacy', much discussed in judgment and decision making (Tversky & Kahneman, 1983). The reasoning goes that if people are using a mental simulation, they should not show such a

conjunction error, since a true-to-life, veridical physical simulation should not produce inconsistencies, just as physical reality should not produce inconsistencies. Therefore, the argument goes, the fact that people do show inconsistencies in physical reasoning suggests that they are not using a fully veridical physical simulation.

The broad theoretical issue at stake, then, is what computations support people's intuitive physics, and whether it is based on mental simulation or not. Here, we work within the general tradition of the probabilistic mental simulation engine approach to intuitive physics. We take seriously the previous criticisms of mental simulation, including the point that people deviate from the true ground state in a way unexpected by a perfect simulation, and that a perfect simulation would likely be very computationally expensive in a cognitively unrealistic way. However, we also work under the assumption that people's mental simulation (if it exists) is almost certainly not a perfect simulation of the physical world. This has been pointed out before, and approximations within the probabilistic physical simulation engine have been explored in several domains, including the idea that a mental dynamic simulation likely includes inherent noise (Smith & Vul, 2013), and the suggestion that systematic conceptual approximations and simplifications, similar to the work-arounds used by engineers to get around limitations of memory and computation in their own physical simulations (Ullman, Spelke, Battaglia, & Tenenbaum, 2017). More recently, it was shown that approximations in the form of partial simulation can quantitatively and qualitatively explain people's deviation from optimal predictions, including the physical conjunction fallacy (Bass, Smith, Bonawitz, & Ullman, 2021), that approximations of object shapes bodies help explain people's reasoning about collisions and physical causality (Li et al., 2023), and that mental simulation may update the state of a single object at a time (Balaban & Ullman, 2024).

In this paper, we examine a general resource limitation of mental simulation engines: the finite tracking of objects. We connect this to the more general theoretical proposal that mental simulation is limited and approximate, while still being a simulation

process. Such a limitation could correspond to attention or working memory limitations, as both are concerned with the number of bodies that can be actively held in mind (Vul, Alvarez, Tenenbaum, & Black, 2009). We examine the ways in which this limitation leads to performance shifting over different settings of a task, and propose a concrete mechanism for how people reason about a given task when their mental resources are taxed beyond a certain limit, using computational modeling.

As a specific task, we focus on people's behavior when reasoning about liquids. This domain has several appealing properties. First, this is a common-sense physical reasoning domain that has been studied empirically and computationally over decades, with both children and adults. Second, the results point to the basic tension we are interested in: Some studies show people are quite good at reasoning about the properties of everyday liquids (Paulun, Kawabe, Nishida, & Fleming, 2015; Van Assen, Barla, & Fleming, 2018), and that reasoning about liquids can be captured by a particle-fluid simulation (Bates, Yildirim, Tenenbaum, & Battaglia, 2019). Other studies point to difficulties, biases, and mistakes. Children have a much harder time reasoning about non-rigid bodies, both in infancy (Huntley-Fenner, Carey, & Solimando, 2002), and later in childhood. Some of the best known results in child developmental show that toddlers do not seem to believe liquids are conserved when liquid is poured between two containers (see Piaget & Inhelder, 1974, and the many follow ups). It should be noted though, that child development also displays the tension we're interested in: more recent cognitive development work shows that even infants can broadly distinguish rigid bodies and liquids, and have roughly accurate overall expectations about the behavior of liquids (Hespos, Ferry, Anderson, Hollenbeck, & Rips, 2016).

Liquids are also an interesting domain to explore from a computational perspective. Accurately simulating the behavior of fluids is a challenging computational task, one that takes more resources than simulating the behavior of rigid bodies (Gregory, 2014; Ullman et al., 2017). It seems likely that any computational limitations on predicting the behavior

of everyday entities will be especially present in reasoning about liquids.

To briefly foreshadow the experiments and results: we examined people's successes and failures when reasoning about the behavior of liquids and containers, using 4 novel (pre-registered) experiments. In Experiments 1 and 2 we establish baseline performance with liquids, asking people to judge when different-sized containers will be fully filled by a stream of water (Experiment 1), and how filled a container is when different amounts of water are poured into it (Experiment 2). People perform well on such tasks, a non-obvious result that adds to the current body of knowledge. But, the stimuli presentation time for Experiments 1 and 2 is relatively brief (under 8 seconds). This success may in principle be explained either by two very different kinds of computation: veridical mental simulation, or visual volume estimation. In Experiment 3, we asked people to make judgments of relative volume based on static scenes, and find that they have systematic biases in their volume estimation, suggesting that simple visual estimation is *not* the way that people solve the tasks in Experiments 1 and 2, as these biases are not present there. In Experiment 4, we increase the presentation time of stimuli from Experiment 1 and encounter a 'switch point' in people's behavior. Participants in Experiment 4 also completed a digit-span working memory task, and we found that people with larger digit spans have a later switch-point. We show that a mental simulation model with a limited budget of particles can account for both the successes of Experiment 1 and 2, as well as the switch point found in Experiment 4, and the relationship between digit span and switch point.

Also, to briefly foreshadow our specific cognitive model (the Bounded Fluid Simulation model): we used a bounded particle-based model, adapted from a common model used to simulate the behavior of fluids. This model takes as its starting point the approximation that fluids can be thought of as a collection of particles that are governed by the partial differential equation known as the Navier-Stokes equation. Previous work Bates et al. (2019) has already demonstrated that modification of such a model can be used to capture people's reasoning about the behavior of fluids. The Bounded Fluid

simulation model (BFS) places a bound on the possible number of particles, which we relate to people's empirically established cognitive resources. The model also introduces a strategy to handle situations in which the number of particles available to the simulation 'runs out', which we relate to people's empirically-established switch point.

Overall Procedures and Methods

The studies were approved under IRB19-1861 (Commonsense Reasoning in Physics and Psychology). All participants provided informed consent. All studies were conducted online (Peer, Brandimarte, Samat, & Acquisti, 2017), recruiting US-based participants via Prolific (https://www.prolific.co). Participants were compensated at a rate of approximately 12 USD/Hour for their time. All stimuli involved video animations or still images, and were created using Blender 3D modeling software http://www.blender.org or Unity game engine https://unity3d.com.

Transparency and Openness: Our studies comply with TOP guidelines. All data and stimuli are available at the following OSF repository: https://osf.io/pvab3/. All studies, including experimental procedures, number of participants, exclusion criteria, and analyses, were pre-registered [1]. Data for the main studies was collected over the period 2022-2023, and for the supplementary materials over 2023-2024. The generality of the findings is constrained by the population that completed our studies, though given its low-level perceptual nature we believe it will generally hold across populations.

Statement on Sample Sizes: As the experiments we consider are not inherited from previous work, it was difficult to know in advance what effect sizes and power analysis would be relevant. Given that, the sample sizes for all studies were based on pilot studies, which were then used for a bootstrap analysis running the same analysis used in the full experiments. We also conducted a post-experiment bootstrap power analysis for each experiment, as detailed below.

A Note on Terminology: Our empirical tasks involved the use of (simulated)

[1] https://osf.io/pvab3/registrations

water-like *liquids* being poured into containers under gravity. Liquids are a sub-type of fluids, but not synonymous with them. Nevertheless, we refer more generally to 'fluids' in our presentation and discussion of the Bounded Fluid Simulation model, as this model is based on models that use particles to handle fluid simulations more generally, and because this use is in keeping with previous usage in tasks involving reasoning about liquids (Bates et al., 2019). We thank an anonymous Reviewer for raising this issue.

Study 1: Time-to-Fill

Our first study aimed to establish people's basic competency with a task that requires reasoning about the behavior of a liquid. Participants were shown videos of different containers being filled with water, and asked to indicate the point at which they thought the containers were full. Basic competency in this case would mean that participants' time-to-fill responses linearly increases with the volume of the containers, and is not significantly affected by container shape.

Stimuli

The stimuli consisted of 21 video animations, each 40 seconds long. Every animation first showed a cylindrical container (a cup) and a tube, panning the camera view smoothly to give a sense of depth. The panning lasted for 8 seconds, ending with the container being shown in profile (see Fig 1). Following the panning, water began to flow from the tube into the container. The containers were opaque, and participants could not see the water inside the container. The rate of flow was constant within and across videos. The container was never shown to be completely filled. Examples of the stimuli for Experiment 1 and 2 are available at this link: https://youtu.be/OCz3qkRVHxU. The full set is in the OSF Repository mentioned above.

The videos were identical except for the container, which varied in a 7x3 design of [volume] x [shape]. That is, the containers came in 7 different volumes, and a container of a particular volume could be either 'regular', 'wide' (1.5 shorter than the base container), or 'tall' (1.5 taller than the base container). The volume of the containers was chosen such

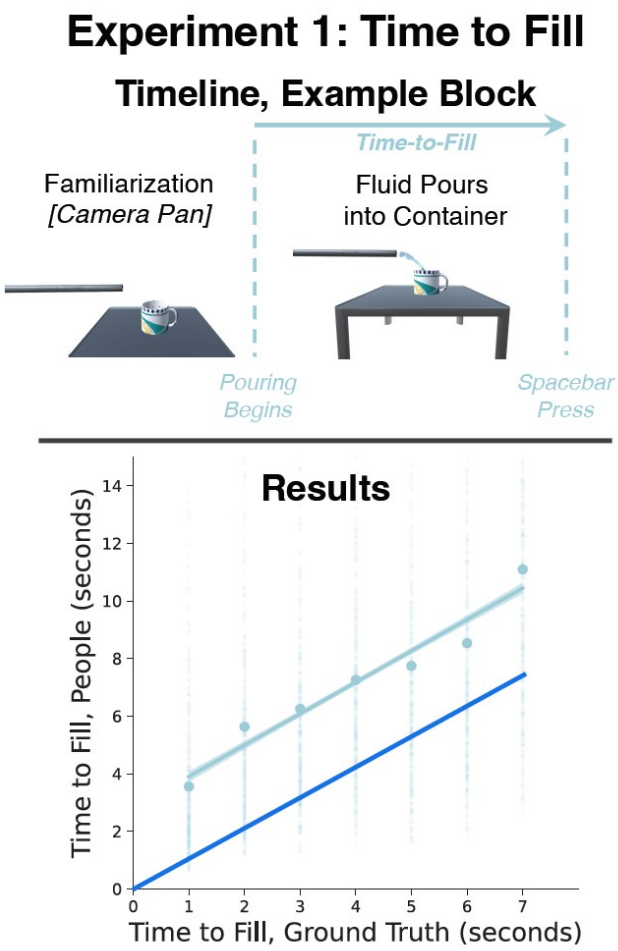

Figure 1

Experiment 1, Time to Fill. *Top: Participants were shown 21 blocks in random order. In each block, an animation familiarized participants with a container, panning the camera and ending with the container in profile. Next, liquid poured into the container, until a participant pressed the spacebar to indicate they thought the container was full. Bottom: Participants' time-to-fill estimates compared to ground truth, in seconds. Large circles indicate mean response time, small circles are individual participants. The straight line in light blue shows a linear fit model, with shaded 95% HDI surrounding it. The 95% HDI is defined as the 95% highest posterior density interval, which is used as 95% Credible Interval (CI) to characterize the uncertainty of the corresponding posterior distribution. The straight blue line shows the identity function.*

that they would take 1, 2, 3, 4, 5, 6, or 7 seconds to fill in reality, given the flow rate. We refer to the time it takes to completely fill a container as 'Time-to-Fill' (TTF). In addition to their volume and shape, the containers randomly varied in color and texture, in a way

unrelated to the questions of the study.

Participants and Methods

We recruited 140 participants, online via Prolific. The sample size was determined based on pilot data, and a post-experiment bootstrap power analysis was conducted to examine this sample size. To the degree that the relevant effect is the existence of a linear correlation between ground-truth time-to-fill and participant response, the bootstrap analysis indicated 7 participants are sufficient to reach a power of 80%, at a significance level of .05. The number of participants is far above this in order to potentially rule out shape effects and conduct model comparisons.

The average completion time was 12 minutes, with standard deviation of 5.9 minutes. The participants were paid at a rate of $12/hrs. Of the original participants, 4 participants were excluded, based on pre-registered criteria. Exclusion criteria included failing to pass a check question, or having 30% or more of the responses on the 21 trials being outliers. In addition, specific responses exceeding the maximum length of the video (40 seconds) were excluded. See https://osf.io/ne2y3 for full details on pre-registered criteria. This left 136 participants for analysis (We asked for the gender of the participants, and they were given four choices: male, female, non-binary, refuse to answer. 74 identified as women, 59 as men, 3 as non-binary. The median age was 34.5 years, with mean of 34.5 years and standard deviation 10.3 years).

Participants were instructed that they would see short videos of different containers being filled, and that for each video they should press the space bar at the moment at which they thought the container was full and about to overflow. Participants were instructed that the videos would never actually show the container being completely filled. After the instructions, participants were asked short validation questions to verify they paid attention, and understood the task. Following validation, participants were presented with all 21 animations in a randomized order. Participants used the space bar to indicate the moment they thought the container was full. After the space bar was pushed, the

experiment moved on to the next animation. At the end of all 21 stimuli, participants were given the option of answering a demographic survey, and thanked for their time.

Results

We coded the time from the beginning of water-flow in each video until a participant pressed the space-bar as a participant's Time-to-Fill for that stimuli (TTF_{Human}). We examined the relationship between this variable, and the actual Time-to-Fill for each container ($TTF_{groundTruth}$).

We fit three different hierarchical models to the relationship between TTF_{Human} and $TTF_{groundTruth}$: A linear model, and two sub-linear models (square-root and logarithmic). The hierarchical aspect of the models accounts for individual differences in overall reaction time. See the sub-section below for details regarding the model analysis, but summarize briefly: We compared the models using Leave-One-Out Cross Validation to calculate Model Weight (MW), and found the best-fitting model to be the linear model ($MW_{Linear} = .853$, $MW_{Logarithmic} = .122$, $MW_{SquareRoot} = .025$). We note that such model comparison replaces Bayes Factors, here and elsewhere. For the detail of the Leave-One-Out Cross Validation methods formulation and Model Weight calculation, please refer to the supplementary material.

Figure 1 (bottom) shows both the data from participants, and the hierarchical linear model. As can be seen in the figure, participants linearly adjusted their Time-to-Fill response with the ground-truth time, based on the container size. People are generally close to the identity function (slope of 1), but with a likely added reaction time component induced by motor-planning above and beyond mental processing related to physical reasoning.

Model analysis and comparison

We examined three candidate hierarchical statistical models that vary in their functional form, and selected the best fitting functional form based on the Leave-One-Out Cross Validation results.

The overall model formulation was:

$$TTF_{human} = a + b * f(TTF_{groundTruth}) + \epsilon \tag{1}$$

Where the functions $f(TTF_{groundTruth})$ we considered were linear, square root, or logarithmic:

Functional Forms for the Hierarchical Models	
Functional Form Names	Formula
Linear	$f(TTF_{groundTruth}) = TTF_{groundTruth}$
Square Root	$f(TTF_{groundTruth}) = \sqrt{TTF_{groundTruth}}$
Logarithmic	$f(TTF_{groundTruth}) = \log TTF_{groundTruth}$

The intercept parameter is denoted as a; the slope parameter is denoted as b; the Gaussian noise is denoted as ϵ.

We now detail the priors for the parameters. Given the fact that we have little prior information about the parameters, we intentionally chose the uniform prior with a wide range to mimic the effect of uninformative priors and make the data to dominate the inference of the posterior distribution of the parameters. The priors were the following:

Priors for Parameters			
Parameters	Linear	Square Root	Logarithmic
a	Uniform(0,4000)	Uniform(-2000,2000)	Uniform(-40000,0)
b	Uniform(0,2)	Uniform(0,140)	Uniform(0,40000)
ϵ	Uniform(0,5000)	Uniform(0,7000)	Uniform(0,10000)

We fit the model at the individual level. Considering the same functional forms, each participant i had their own a_i, b_i, ϵ_i, drawn from Gaussian distributions over the hyperpriors above. The model at the individual level was formulated as:

$$TTF_{human_i} = a_i + b_i * f(TTF_{groundTruth}) + \epsilon_i \tag{2}$$

Individual parameters included:

Individual Priors for Parameters			
Parameters	Linear	Square Root	Logarithmic
a_i	Normal(a, 1900)	Normal(a, 1500)	Normal(a, 2000)
b_i	Normal(b, 0.49)	Normal(b, 25)	Normal(b, 300)
ϵ_i	Normal(ϵ, 1000)	Normal(ϵ, 1000)	Normal(ϵ, 1000)

The three candidate models were compared using Leave-One-Out Cross Validation with Pareto-Smoothed Importance Sampling (PSIS). The model weight (MW) found following this procedure were MW(Linear) = .788, MW(Logarithmic) = .211, MW(Square Root) = .001.

Possible Effects of Container shape

With the linear model as the best fit, we next considered whether the container shape impacted volume estimation. That is, we examined whether the wide/regular/tall shapes of a container influenced Time-to-Fill. To summarize the result of the analysis below: As shown in Table 1, we found that the container's shape did not affect participants' estimation of how long it takes to fill a container.

In more detail, we fit a hierarchical linear model with additional parameters to account for the various shapes of the containers (different slopes and intercepts for 'wide' and 'tall' containers, with 'regular' containers being the baseline), as follows:

$$TTF_{Human} = (a + c * k_1 + d * k_2) + (b_e * k_2 + f * k_2) * TTF_{groundTruth} + \epsilon \tag{3}$$

Where a is the intercept, b is the slope, c is the intercept modification due to wide type container, d is the intercept modification due to tall type container, e is the slope modification due to wide type container, f is the slope modification due to tall type container, k_1 is a binary indicator for a wide container (1 for trials with container that has type wide, 0 for trials with container that does not have type wide), k_2 is another binary

Variable	Point Estimate	HDI [2.5%—97.5%]
Intercept	**2.9**	**2.5—3.2**
Intercept Modifier [Wide]	-0.5	-1.1—0.02
Intercept Modifier [Tall]	0.3	-0.3—0.9
Slope	**1.10**	**0.9—1.3**
Slope Modifier [Wide]	0.0	-0.2—0.2
Slope Modifier [Tall]	0.0	-0.2—0.2

Table 1

The result of fitting a linear model to the results of Experiment 1, Time-to-Fill. We examined if there was a variation of the slope and intercept based on container shape (wide/regular/tall). The HDI for the wide and tall modification parameters intersects with 0, indicating the container shape did not affect the estimation of the water level.

indicator for a tall container (1 for trials with container that has type tall, 0 for trials with container that does not have type tall), and ϵ is the gaussian noise.

We now detail the priors for the parameters. Given the fact that we have little prior information about the parameters, we intentionally chose the uniform prior with a wide range to mimic the effect of uninformative priors and make the data to dominate the inference of the posterior distribution of the parameters. The hyperparameters has the following setup:

Priors for Parameters	
Parameters	Priors
a	Uniform(0,6000)
b	Uniform(0,2)
c	Uniform(-2000,2000)
d	Uniform(-2000,2000)
e	Uniform(-1,1)
f	Uniform(-1,1)
ϵ	Uniform(0,10000)

We fit the model at the individual level. Each participant i had their own priors a_i, b_i, c_i, d_i, e_i, f_i, ϵ_i drawn a Normal distribution centered on the priors above:

$$TTF_{Human} = (a_i + c_i * k_1 + d_i * k_2) + (b_i + e_i * k_2 + f_i * k_2) * TTF_{groundTruth} + \epsilon_i \quad (4)$$

Individual-level hyperprior parameters are formulated as the following:

Priors for Parameters	
Parameters	Priors
a_i	Normal(a,150)
b_i	Normal(b,1)
c_i	Normal(c,1000)
d_i	Normal(d,1000)
e_i	Normal(e,1)
f_i	Normal(f,1)
ϵ_i	Normal(ϵ,1000)

We found that the point estimates of the hyperprior parameters (means and 95% HDI) for the model were: a = 2898.537, with 95% HDI [2527.407, 3268.189]; c = -536.63, with 95% HDI [-1091.613, 24.42]; d = 278.211, with 95% HDI [-258.364, 852.354]; b =

1.096, with 95% HDI [0.909, 1.281]; e = -0.043, with 95% HDI [-0.243, 0.163]; f = 0.024, with 95% HDI [-0.174, 0.233]; ϵ = 2677.376, with 95% HDI [748.656, 4587.369].

As shown in Table 1, the point estimation of the hyperprior parameters (means and 95% credible interval) indicate no effect of object shape.

We would like to emphasize that we do not commit to whether shape effect will affect people's performance in other conditions that are different from our experimental setup, we only commit to the result about shape effect under our experimental setup. See supplementary material for more information.

Before proceeding to the next experiment, we emphasize that the (hierarchical) statistical models we considered so far, and consider for the following experiments, are not the same as *cognitive models* accounting for the behavior. The statistical models are a-theoretic tools for assessing participant performance, and can in turn be used to consider cognitive models of behavior (whether heuristic models or mental simulation). We consider our specific cognitive model, a bounded mental simulation, in the modeling section after detailing the experiments.

Study 2 : Estimations of Water Level

Study 1 established people's basic competency on a reasoning task involving liquids, showing a linear response to volume differences, and no significant variation by shape when keeping the volume consistent. Such a task could be solved by a mental simulation, but it could also be solved in other ways, especially given the constant flow rate. For example, if people have an accurate estimation of a container's volume from visual inspection, combined with an estimation of the flow, they do not need to track liquids moment by moment. Rather, people could in principle combine their estimation of volume and flow to compute time-to-fill early on, and then simply wait to execute their response.

In the second study, we examined people's competency in a task that required moment-by-moment tracking of liquids in an unfolding scene. As before, participants were shown videos of different containers being filled by water. However, this time the scene

paused at randomized intervals. Participants were then asked to indicate how full a container was. Competency in this task would mean that participants' estimation of how full a container is would be a linear function of the ground-truth amount of liquid, and would not vary based on irrelevant container information such as shapes.

Stimuli

The stimuli consisted of 45 video animations, similar in general format to Study 1. As in Study 1, every video first showed an opaque cylindrical container (a cup) and a tube, panning the camera view smoothly to give a sense of depth for 8 seconds, ending with the container in profile (see Figure 2, top). As in Study 1, after the familiarization, water begins to flow at a constant rate from the tube.

The videos varied in a 3x3x5 design of [volume] x [shape] x [duration]. As in Study 1, the containers varied in volume and shape, with a container of a particular volume being either 'regular', 'wide', or 'tall'. The duration was either 2, 3, 4, 5, or 6 seconds. The container never overflowed. The containers again randomly varied in color and texture, in a way unrelated to the relevant questions of the study.

We also created 45 control videos, which were identical to main stimuli, except that the container was transparent. These stimuli were used in a control task, as detailed below.

Participants and Methods

We recruited 90 US-based participants online (We asked for the gender of the participants, and they were given four choices: male, female, non-binary, refuse to answer. 56 identified as women, 31 as men, and 3 declined to answer. The median age was 32.5 years, the mean age was 37.9 years, the standard deviation of age was 13.9 years). Of these, 70 participants were assigned to the main study, and 20 participants were assigned to 2 control condition (10 per control). We find the absolute difference between the average of first control and second control by trial to create the ground-truth water level $WLgroundTruth$ by trial. As in Experiment 1, to the degree that effect of interest is a linear relationship between the water-height and participant response, a post-experiment

bootstrap power analysis indicated that 5 participants are sufficient to reach the power of 80% at a significance level = .05. The number of participants recruited was far above this to examine model comparisons and shape effects. The average completion time was 22 minutes, with standard deviation 8.7 minutes. The participants were paid at a rate of $12/hrs.

In the main study, participants were instructed that they would see short videos of different containers being filled, and that for each video they would be asked to indicate how full a container was when the video paused. After the instructions, participants were given a practice trial using a transparent container, and then asked short validation questions to verify they paid attention, and understood the task. Following validation, participants were presented with all 45 animations in a randomized order. At the end of each animation, a red bar was super-imposed the container, placed in the horizontal center, and running vertically from top to bottom (see Fig. 2). Participants clicked on the red bar to indicate how full the container was, and the (x, y) coordinates of their click was recorded. At the end of all 45 stimuli, participants were given the option of answering a demographic survey, and thanked for their time.

The control studies were similar to the main study, except that the animations used a transparent container. The first control study asked participants to indicate the water level (how full the container was). The second control study asked participants to indicate the lowest point of the liquid inside the container. These responses were used to establish ground-truth water levels for comparisons with the main study.

Results

We used the y-coordinate (height, in pixels) indicated by participants when they clicked on the red bar to indicate the estimated water level (WL), which we refer to as $WL_{Participant}$. We used the y-coordinates (height, in pixels) indicated by participants in the transparent control tasks to establish ground-truth water levels, which we refer to as $WL_{groundTruth}$.

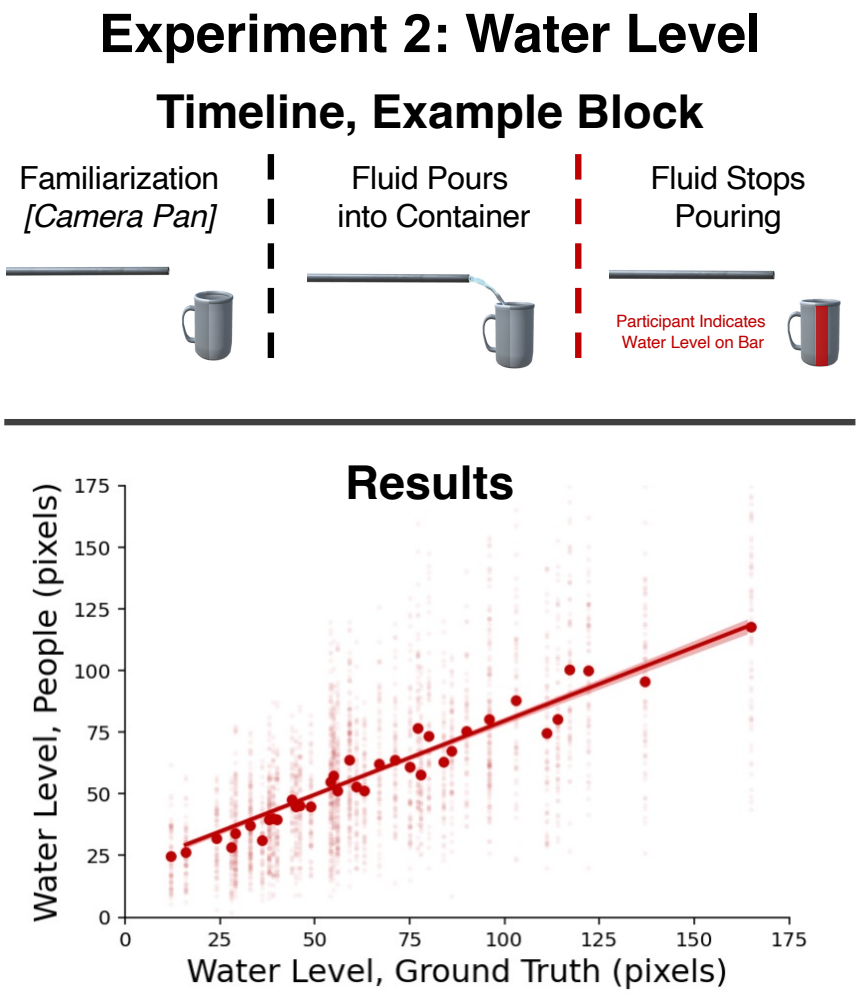

Figure 2

Experiment 2, Water Level. *Top: Participants were shown 45 blocks. In each block, an animation panned the camera and ended with the container in profile. Next, liquid poured into the container, pausing at a random interval. A red vertical bar appeared, and the participant clicked on it to indicate how full the container was. Bottom: Participants' water-level estimates compared to ground truth, in pixels. Large circles indicate participant means, small circles are individual participants. The straight line is the linear fit model, with shaded 95% HDI surrounding it.*

We originally opted to use participant judgments to establish ground-truth water levels (via the control studies), rather than the water level as calculated by ground-truth physics, in order to avoid the potential pitfall in which people's perception of the water level deviates from the physical ground truth even when the scene is fully visible. If such a deviation occurred but people match the control estimates, people could still be said to be

performing well in the main study. This whole concern turned out to be completely irrelevant in practice, as there was no deviation between physical ground truth and participant judgments in the transparent control task. In addition, the standard deviation around the ground truth was quite small (2.7 pixels on average, an order of magnitude smaller than the average standard deviation around the height variable in the main condition), so in Figure 2 we show the ground truth water levels as the comparison.

We examined the relationship between $WL_{Participant}$ and $WL_{groundTruth}$. As in Study 1, we fit three hierarchical models to the relationship: A linear model, and two sub-linear models (square-root and logarithmic). In the subsection below we describe the model analysis in full detail, but to summarize briefly: Using the same model comparison techniques as Study 1, we, found the best-fitting model to be the linear model ($MW_{Linear} = .90$, $MW_{Logarithmic} = .08$, $MW_{SquareRoot} = .02$). We note again that such model weights replace Bayes Factors.

Figure 2 (bottom) shows both the data from participants and the hierarchical linear model. As can be seen, participants linearly adjusted their water-level estimates with the ground-truth, based on the container size.

Model analysis and comparison

We used two variables, WL_{Human} and $WL_{groundTruth}$ in our analysis. The variable WL_{Human} is the difference between the Y coordinate of participant water level, and the averaged Y coordinate of the bottom of the corresponding container.

The variable $WL_{groundTruth}$ is the averaged difference between the Y coordinate of participant responded ground truth water level, and the Y coordinate of the bottom of the corresponding container.

The Y coordinate of the bottom of the corresponding container and the Y coordinate of participant responded ground truth water level were obtained by averaging across participants' response in a control experiment using transparent container. This was done due to a theoretical concern that people's estimates even when directly perceiving the

containers would vary from the ground truth. This concern turned out to not matter in practice, as people's estimation in the control was the same as ground truth.

The variables WL_{Human}, $WL_{groundTruth}$, and the X,Y coordinates were all measured in pixels.

We examined three candidate hierarchical statistical models that vary in their functional form, and selected the best fitting functional form based on the Leave-One-Out Cross Validation results.

The overall model formulation was:

$$WL_{Human} = a + b * f(WL_{groundTruth}) + \epsilon \tag{5}$$

Where the functions $f(WL_{groundTruth})$ we considered were linear, square root, or logarithmic:

Functional Forms for the Hierarchical Models	
Functional Form Names	Formula
Linear	$f(WL_{groundTruth}) = WL_{groundTruth}$
Square Root	$f(WL_{groundTruth}) = \sqrt{WL_{groundTruth}}$
Logarithmic	$f(WL_{groundTruth}) = \log WL_{groundTruth}$

We now detail the priors for the parameters. Given the fact that we have little prior information about the parameters, we intentionally chose the uniform prior with a wide range to mimic the effect of uninformative priors and make the data to dominate the inference of the posterior distribution of the parameters. The intercept parameter is denoted as a; the slope parameter is denoted as b; the Gaussian noise is denoted as ϵ.

We chose uniform prior with wide range to mimic the uninformative priors for our parameters, since we don't have prior information about the parameters we were fitting. The priors were the following:

Priors for Parameters			
Parameters	Linear	Square Root	Logarithmic
a	Uniform(0,1000)	Uniform(-1000,1000)	Uniform(-1000,1000)
b	Uniform(0,2)	Uniform(0,100)	Uniform(0,100)
ϵ	Uniform(0,100)	Uniform(0,100)	Uniform(0,100)

We fit the model at the individual level. Considering the same functional forms, each participant i had their own a_i, b_i, ϵ_i, drawn from Gaussian distributions over the hyperpriors above. The model at the individual level was formulated as:

$$WL_{human_i} = a_i + b_i * f(WL_{groundTruth}) + \epsilon_i \tag{6}$$

Individual parameters included:

Individual Priors for Parameters			
Parameters	Linear	Square Root	Logarithmic
a_i	Normal(a, 4)	Normal(a, 1.9)	Normal(a, 4)
b_i	Normal(b, 0.1)	Normal(b, 0.23)	Normal(b, 0.9)
ϵ_i	Normal(ϵ, 10)	Normal(ϵ, 10)	Normal(ϵ, 10)

The three model candidates were compared using Leave-One-Out Cross Validation. Following this procedure, we found that the model weights (MW) were: MW(Linear) = .901, MW(Logarithmic) = .080, MW(Square Root) = .019.

Possible effects of container shape

Having established the linear model as the best fit, we next considered whether the container shape (wide/regular/tall) made a difference in estimation. We fit a hierarchical linear model with additional parameters to take into account the possible separate effects of the container shape on slope and intercept (different slopes and intercept for 'wide' and 'tall', on top of a baseline 'regular'). The results of this analysis are shown in Table 2. To

Variable	Point Estimate	HDI [2.5%—97.5%]
Intercept	**20.2**	**17.0—23.3**
Intercept Modifier [Wide]	-2.4	-6.7—1.8
Intercept Modifier [Tall]	1.2	-3.1—5.5
Slope	**0.58**	**0.53—0.63**
Slope Modifier [Wide]	0.03	-0.05—0.11
Slope Modifier [Tall]	0.03	-0.03—0.09

Table 2

Parameter estimates from fitting a linear model to the results of Experiment 2 (water level). We examined if there was a variation in the slope and intercept based on container shape (wide/regular/tall). The HDI for the wide and tall modification parameters intersects with 0, indicating the container shape did not affect the estimation of the water level.

briefly summarize the results of the analysis below: We found that the container's shape (wide/tall/regular) did not affect the estimation of how full a container was.

In more detail, we modified our HLM to use one formulation with a linear functional form to predict WL_{Human} for wide, regular, and tall type container with different $WL_{groundTruth}$ Time-to-Fill to the certain water level:

$$WL_{Human} = (a + c * k_1 + d * k_2) + (b + e * k_1 + f * k_2) * WL_{groundTruth} + \epsilon \tag{7}$$

Where a is the intercept, b is the slope, c is the intercept modification due to wide type container, d is the intercept modification due to tall type container, e is the slope modification due to wide type container, f is the slope modification due to tall type container, k_1 is a binary indicator for a wide container (1 for trials with container that has type wide, 0 for trials with container that does not have type wide), k_2 is another binary indicator for a tall container (1 for trials with container that has type tall, 0 for trials with container that does not have type tall), and ϵ is the gaussian noise.

We now detail the priors for the parameters. Given the fact that we have little prior information about the parameters, we intentionally chose the uniform prior with a wide range to mimic the effect of uninformative priors and make the data to dominate the inference of the posterior distribution of the parameters. The hyperparameters has the following setup:

Priors for Parameters	
Parameters	Priors
a	Uniform(0,40)
b	Uniform(0,1)
c	Uniform(-20,20)
d	Uniform(-20,20)
e	Uniform(-0.2,0.2)
f	Uniform(-0.2,0.2)
ϵ	Uniform(0,50)

We fit the model at the individual level. Each participant i had their own priors a_i, b_i, c_i, d_i, e_i, f_i, ϵ_i drawn from the a Gaussian distribution centered on the priors above:

$$WL_{Human} = (a_i + c_i * k_1 + d_i * k_2) + (b_i + e_i * k_1 + f_i * k_2) * WL_{groundTruth} + \epsilon_i \quad (8)$$

Individual-level hyperprior parameters are formulated as the following:

Priors for Parameters	
Parameters	Priors
a_i	Normal(a,4)
b_i	Normal(b,0.1)
c_i	Normal(c,1)
d_i	Normal(d,1)
e_i	Normal(e,0.1)
f_i	Normal(f,0.1)
ϵ_i	Normal(ϵ,10)

As shown in Table 2, the point estimation of the hyperprior parameters (means and 95% credible interval) indicate no effect of object shape.

We would like to emphasize that we do not commit to whether shape effect will affect people's performance in other conditions that are different from our experimental setup, we only commit to the result about shape effect under our experimental setup. See supplementary material for more information.

Study 3: Estimating Container Volume in Static Scenes

In Study 2, participants performed reasonably well at estimating the water level reached after a given time. Such behavior seems to require moment-by-moment tracking of the scene, which can accounted for by mental simulation. However, it is still in principle possible that people were using static visual heuristics or visual routines to estimate the volume of the containers, which then indicates the water level of the containers. In this process, no mental simulation or knowledge of physics quantity estimation is needed. For example, the volume of the containers can be estimated with simple visual heuristics or visual routines like combining the rate of flow, and the time elapsed, to give an estimate of the water level (though this is already a more complex alternative model than the non-simulation alternative considered in Study 1). Once people established this kind of baseline by using simple visual heuristics or visual routine while doing the tasks, they could

re-scale this baseline depending on the cup presented in the task to directly compare the volume instead of estimating the volume.

In order to better assess the possibility that people are using visual heuristics (rather than mental simulation) to carry out the tasks of Studies 1 and 2, we examined whether people can reliably estimate the volume of containers given static scenes. If people are *unbiased* in their volume estimation, this would not strictly rule out mental simulation, but it would be more parsimonious to suggest that people are using visual heuristics or routines for the previous tasks. If people are *biased* in their estimations of volume, this would suggest that people are not relying on visual estimates alone to carry out the previous two tasks, and more likely that they are using a mental simulation.

Stimuli

The stimuli consisted of 54 video animations, each approximately 8 seconds long. Each video showed two containers already filled with water, placed next to each other on a table. In each the video, the camera first panned, then ended with a horizontal view of the two containers (see Fig. 3, top).

The containers in the videos varied in shape and volume, in the following way: In 18 of the videos, the two containers were of equal volume (volume ratio 1:1). In the remaining 36 videos, the containers were of unequal volume (12 videos showing a 1:1.5 volume ratio, 12 videos showing a 1:3 ratio, and 12 videos showing a 1:5 ratio). The containers in all videos were based on two base meshes ('Regular'), which were varied and shaped to be 1.5 times shorter ('Wide' shape), or 1.5 times taller ('Tall' shape) for a given volume. Each ratio used a balanced number of such pairs.

Participants and Methods

We recruited 60 US-based participants online via Prolific (We asked for the gender of the participants, and they were given four choices: male, female, non-binary, refuse to answer. 37 identified as women, 21 identified as men, 2 identified as non-binary, and one person preferred not to answer. The median age was 35 years, the mean age was 38 years,

and the standard deviation was 12.9 years.). As before, we conducted a post-experiment power analysis, examining the main effect of deviation from a symmetric response. A post-experiment power analysis indicates that 55 participants are sufficient to reach a power of 80%, at a significance level of .05.

Participants were instructed that they would see short videos of different container pairs, and that for each video they would be asked to indicate which container held more water in it. After these instructions, participants were given a practice trial, and then asked short validation questions to verify they paid attention, and understood the task. Following validation, participants were presented with all 54 animations in a randomized order. At the end of each animation, participants were asked to indicate which container held more water using a 5-point scale: (1) The left container has a lot more water, (2) The left container has a bit more water, (3) The water amount in both containers is equal, (4) The right container has a bit more water, (5) The right container has a lot more water. At the end of the main study, participants were given the option of answering a demographic survey, and thanked for their time. The average completion time was 23 minutes, with standard deviation of 9.7 minutes. The participants were paid at a rate of $12/hrs.

Results

For ease of analysis, we treat the taller container as always being on the right (even though the stimuli counterbalanced the left/right presentation). We numerically coded the categorical choices as follows: 'The left container has a lot more water': -2; 'The left container has a bit more water': -1; 'The water amount in both containers is equal': 0; 'The right container has a bit more water': 1; 'The right container has a lot more water': 2. This makes it easier to reason about any deviations from symmetry as a deviation from 0. We examined the average participant response in relation to the container shape and the volume ratio.

We first examined whether and how a container's shape influences people's volume estimation when the volume is equal. That is, we focused the analysis on the stimuli in

which the volume ratio was 1:1, and the container shapes were different (wide-regular, regular-tall, wide-tall). As shown in Figure 3 (bottom panel, black x markers), participants showed a significant bias in volume estimation: In all cases, the wider container was seen as containing more water, with all scores being significantly different than 0. The mean score for the Wide-Regular pair was -0.29, 95% CI [-0.35, -0.22]; the mean score for the Regular-Tall pair was -0.43, 95% CI [-0.50, -0.36]; the mean score for the Wide-Tall pair was -0.36, 95% CI [-0.43, -0.29]. All scores were also significantly different than 0 by t-tests with $\alpha = .05$. See the subsection below for the full analysis and results.

We next examined how judgments of relative volume change as both the container shape and the volume ratio change, summarized also in Figure 3 (bottom panel, colored o markers). There are two findings of interest here. The first finding is that as the true volume ratio increases, people correctly estimated that the container of greater volume contains more water. Ratios of 1:3, 1:5, 3:1, and 5:1 all resulted in ceiling performance. Such a finding is not surprising, but it serves as a check that participants were indeed sensitive to visual depictions of differing volumes, and can correctly carry out the task.

The second finding is that in the case of ratios for which performance was not at ceiling (1:1.5 and 1.5:1), participants exhibited the same bias as the equal-volume cases: wider containers were seen as having more water in them, when analyzed in static scenes. Put more directly: while the purple o markers for the 1.5:1 ratio are below the 0 line, and the yellow o markers for the 1:1.5 ratio are above the 0 line, the 1.5:1 markers are more below the line of symmetry than the 1:1.5 markers are above it (paired t-tests at $\alpha = .05$ on the 1.5:1 and 1:1.5 ratio containers showed significant differences for all cases. The difference in mean score for the Wide-Regular pair was -0.57, 95% CI [-0.79,-0.36]; the difference in mean score for the Regular-Tall pair was -0.26, 95% CI [-0.48,-0.05]; the difference in mean score for the Wide-Tall pair was -0.61, 95% CI [-0.84, -0.38]. See the subsection below for the full analysis and results).

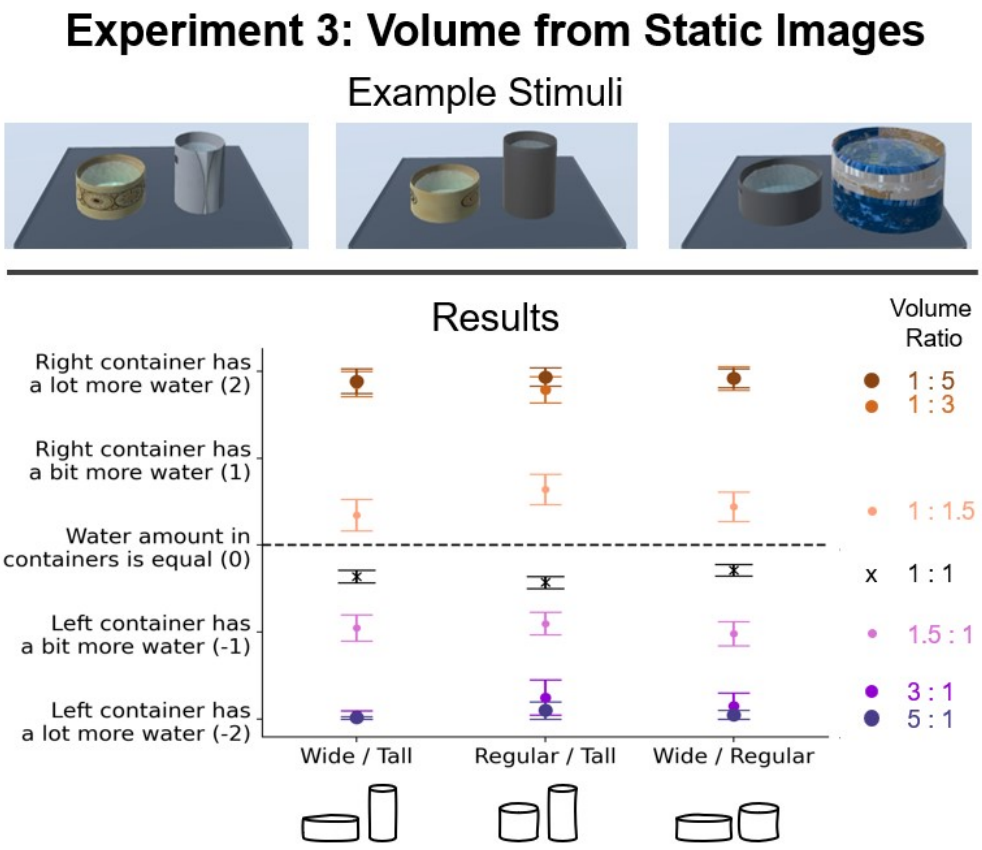

Figure 3

Experiment 3: Estimation of Volume in Static Scenes. *Top: Examples of different stimuli pairs. Participants were shown 54 container pairs, one pair at a time, and were asked to judge which container had more water in it. Bottom: Participant estimations of container volume, varying by container shape (wide/regular/tall) and volume. For ease of presentation here, wider containers are shown on the left. However, in the study the left/right position was counterbalanced. The y-axis shows the ordered scale participants used to respond, which was transformed into linear ratings ranging from -2 to +2. The x-axis shows the different possible shapes. Participants were overall correctly sensitive to changes in the volume ratio, in the sense that as the volume ratio grew more towards one one container, their response changed towards that container. However, participants were also systematically biased, such that wider containers were seen as holding more water.*

Analysis Details

We were interested people's M_{Score}, the mean of the answer they gave using the 5-point scale above, and how this changed by volume or shape. For ease of analysis, we re-code the 1-5 scale above to the range [-2, +2], with 0 indicating the participant thought

both container had the same amount of liquid. We used t-tests at the $\alpha = .05$ level on response data from the participants by trial type, which created 21 sub-variables, named in the following way: $M_{Score,TrialType,VolumeRatio}$. The 'TrialType' indicator was from the set [WR, RT, WT], corresponding to 'Wide-Regular', 'Regular-Tall', and 'Wide-Tall' container pairs. The 'VolumeRatio' indicator was from the set [1:1, 1:1.5, 1:3, 1:5]. Again, the left/right aspects were counterbalanced.

We first analyzed whether people's static estimates of volume were biased in cases where the ground truth volume was the same. That is, we examined whether $M_{ScoreWR1:1}$, $M_{ScoreRT1:1}$, and $M_{ScoreWT1:1}$ were significantly different from 0. T-tests indicated all are significantly different from 0 ($T(365) = -9.6039$, p = $2.2 * 10^{-16}$ in wide-tall 1:1 ratio condition; $T(365) = -12.182$, p = $2.2 * 10^{-16}$ in regular-tall 1:1 ratio condition; $T(365) = -8.7115$, p = $2.2 * 10^{-16}$ in wide-regular 1:1 ratio condition). This indicated people were biased in their estimation, such that wider containers were seen as holding more liquid.

Then, we analyzed whether $M_{ScoreWR1:1.5}$, $M_{ScoreRT1:1.5}$, $M_{ScoreWT1:1.5}$, $M_{ScoreWR1:3}$, $M_{ScoreRT1:3}$, $M_{ScoreWT1:3}$, $M_{ScoreWR1:5}$, $M_{ScoreRT1:5}$, $M_{ScoreWT1:5}$, $M_{ScoreWR1.5:1}$, $M_{ScoreRT1.5:1}$, $M_{ScoreWT1.5:1}$, $M_{ScoreWR3:1}$, $M_{ScoreRT3:1}$, $M_{ScoreWT3:1}$, $M_{ScoreWR5:1}$, $M_{ScoreRT5:1}$ and $M_{ScoreWT5:1}$ are significantly different from 0. T-tests indicated all are significantly different from 0 ($T(60) = -14.921$, p = $2.2 * 10^{-16}$ in wide-regular 1.5:1 ratio condition; $T(60) = -13.896$, p = $2.2 * 10^{-16}$ in regular-tall 1.5:1 ratio condition; $T(60) = -12.597$, p = $2.2 * 10^{-16}$ in wide-tall 1.5:1 ratio condition. $T(60) = -22.634$, p = $2.2 * 10^{-16}$ in wide-regular 3:1 ratio condition; $T(60) = -17.06$, p = $2.2 * 10^{-16}$ in regular-tall 3:1 ratio condition; $T(60) = -53.577$, p = $2.2 * 10^{-16}$ in wide-tall 3:1 ratio condition. $T(60) = -68.725$, p = $2.2 * 10^{-16}$ in wide-regular 5:1 ratio condition; $T(60) = -27.101$, p = $2.2 * 10^{-16}$ in regular-tall 5:1 ratio condition; $T(60) = -121$, p = $2.2 * 10^{-16}$ in wide-tall 5:1 ratio condition). This indicated people are correctly adjusting their answers in response to changes in volume.

Finally, we analyzed whether the wide-container bias held in trials with 1.5:1 and 1:1.5 volume ratio containers. We performed paired T-test to examine the difference in mean score between 1.5:1 and 1:1.5 ratio containers. The paired T-test indicated that in all cases, the wider container was seen as containing more water, with all the differences being significantly different than 0 and negative, with $\alpha = .05$ ($T(60) = -5.2986$), p = $1.756 * 10^{-6}$ in wide-regular condition, $T(60) = -2.4545$, p = .017 in regular-tall condition, $T(60) = -5.266$, p = $1.982 * 10^{-6}$ in wide-tall condition).

To summarize, participants were biased in their estimation of container volume when looking at static scenes, such that wider containers were seen as holding more water. This bias did not exist for the first two studies, suggesting that people were not using (only) static visual volume estimation to calculate time-to-fill (Experiment 1) and water levels (Experiment 2), and that more likely they were relying on an ongoing mental simulation[2].

Study 4: Resource Limitations in Longer Task

In the next study, we examined people's reasoning about liquids over longer periods of time. We hypothesized that there may be a *switch point* in people's reasoning, such that up until that point people have sufficient mental resources to competently track the amount of liquid, but not past it. That is, if people are using 'particles' to track the fluid (similar to a how many physical simulations work), then the number of 'particles' at their disposal is finite and will run out at some point.

We further hypothesized that if such a switch point exists, its specific value will vary by individual capacity: if a person has more mental resources available in general, the switch point for them should be further in time. That is, some people may have more 'particles' at their disposal than others, and so they will run out of them later than others.

---

[2] As an aside, we note that while our statistical, pre-registered analysis did not find an effect of width in Experiment 1, it is possible under a post-hoc analysis with different parameters to potentially infer a small, inconsistent width effect. However, such an effect, even if it exists, importantly goes in the opposite direction to the effect of the width in Experiment 3

The current study had two parts. The first part involved reasoning about liquids, and followed the overall logic, design, and stimuli of Study 1, but for longer periods. The second part involved a non-visual digit span task (Baddeley, 1992), to assess people's differing capacity.

Stimuli

The stimuli for the reasoning-about-fluids part of the study consisted of 13 video animations, each one approximately 35 seconds long. As before, every animation showed a cylindrical container and a tube. The camera panned, and then liquid began to flow at a constant rate from the tube. We created 13 containers, scaling a base container up or down such that the ground-truth time-to-fill was 1, 2, 3, 4, 5, 6, 7, 8, 10, 13, 16, 19, or 22 seconds. Given that we did not find an effect of container shape in Studies 1 and 2, and that possible shape effects were not the main concern of this study, we dispensed with creating shape variations for the same volume. The containers varied in color and texture, in a way unrelated to the task.

Participants and Methods

We recruited 220 US-based participants online. We excluded 42 participants based on our pre-registered exclusion criterion. Reasons for exclusion included failing to pass a catch question, and having 30% of one's response (or more) to the 13 trials being outliers. In addition, specific responses were excluded if TTF was longer than 35000ms (the maximum length of the video in the first part of the study), and see https://osf.io/cn5js for the detail of the exclusion criterion. This left 178 participants for analysis (We asked for the gender of the participants, and they were given four choices: male, female, non-binary, refuse to answer. 125 identified as women, 52 identified as men, one person declined to answer. The median age was 34, the mean age was 41, and the standard deviation of age was 13.8 years). We also conducted a post-experiment bootstrap power analysis, examining the existence of the switch point as the effect of interest. The analysis indicates 20 participants are sufficient to reach a power of 80% at a significance

level of .05. The average completion time of the study was 20 minutes, with standard deviation of 10.7 minutes. The participants were paid at a rate of $12/hrs.

The study had two parts (see also Figure 4, top). The first part of the study was similar to Study 1: participants were instructed that they would see short videos of different containers being filled, and that for each video they should press the space bar at the moment at which they thought the container was full and about to overflow. Participants were instructed that the videos would never actually show the container being completely filled. After the instructions, participants were given a practice trial and asked short validation questions. Following validation, participants were presented with all 13 animations in a randomized order.

In the second part of the study, participants completed two rounds of a classic digit-span task. Each round consisted of 9 trials. Each trial increased the number of digits the participants had to recall. In each trial, participants saw digits one at a time, randomly drawn from the 0-9 set. Each digit was visible for 1 second, followed by a blank screen for 1 second, followed by the next digit. Once all digits in a trial were presented, participants were instructed to use a virtual keyboard to recall the numbers presented, in the correct order. Participants then proceeded to the next trial. The number of digits presented for recall increased from trial to trial, beginning with 3 digits on trial 1, and increasing to 11 digits on trial 9. At the end of the study, participants were given the option of answering a demographic survey, and thanked for their time.

Results

As in Study 1, we coded the time from the beginning of the water flow in each video until a participant pressed the space-bar as the participant's Time-to-Fill for that stimuli (TTF_{Human}), and examined it in relation to the actual Time-to-Fill for each container ($TTF_{groundTruth}$).

In addition, we coded performance in the digit-span task ($digitSpan$) in the following way: Each trial in the digit span task was considered a success if and only if the

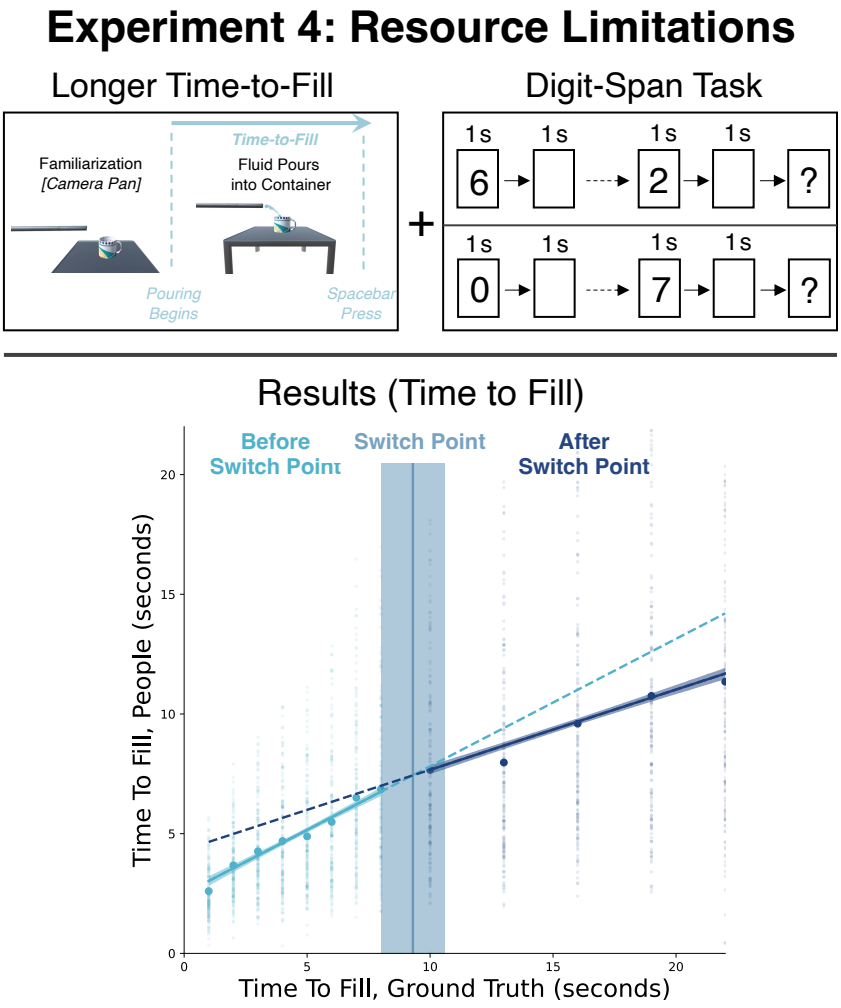

Figure 4

Experiment 4 *Top: Participants first completed a time-to-fill task similar to Study 1, with longer times. Participants then completed 2 blocks of a digit-span task, reporting back sequences of numbers. Bottom: Participants' time-to-fill estimates compared to ground truth, in seconds. Large circles indicate means, small circles show individual responses. The best-fitting model suggests a switch-point in reasoning, with different linear fits before and after. Diagonal linear lines show best model fit to data, with dotted lines continuing the trend before and after the switch to show the off-set. Shaded areas indicate 95% HDI.*

participant recalled all the numbers in the correct order. We averaged together the highest-performance trial from each of the 2 rounds, per participant. For example, if a participant's best performance was correctly recalling 7 digits in one trial and 9 in a second trial, their *digitSpan* would be 8. The internal reliability analysis was conducted on this

task. Please refer to the supplementary material for the detail.

We were interested in whether and how the relationship between TTF_{Human} and $TTF_{groundTruth}$ changes, as the container takes longer to fill. In addition, we were interested in whether this change is related to capacity limitations as captured by *digitSpan*. We analyzed whether there was a *switch-point* in people's reasoning by comparing different hierarchical models: three models *without* a switch-point (linear, logarithmic, square-root), and three models *with* a switch-point. The switch-point models assume that people's reasoning up to the switch-point is linear (in line with Study 1), and that after the switch point the response is either linear (but potentially different) or sub-linear.

We compared the models using Leave-One-Out Cross Validation to calculate Model Weight (MW), and found the best-fitting model to be the switch point model (see Figure 4, bottom), with a linear functional form both before and after the switch point, and mean switch point at 9.3 seconds, 95% HDI: [8.0 – 10.7]. Overall, the model weight in favor of a switch point is .998. See the subsection below for full model details and analysis.

Having found support for a switch point in people's reasoning, we next examined whether the position of the switch point was related to mental capacity. We used a median split by *digitSpan*, creating one group of participants with low performance on the digit span task ($digitSpan \leq 7$), and another group with higher performance on the digit span task ($digitSpan > 7$). We fit the previously-best-fitting hierarchical model separately to these two groups, and found that there is a significant difference between the switch points of the two groups: The participants with better performance on the digit span task had a later switch-point, (10.4 seconds, 95% HDI [8.8 – 12.1], compared to participants with worse performance on the digit span task (7.4 seconds, 95% HDI [6.0, 8.7]). Again, see the subsection below for full model details and analysis.

Analysis details

We considered several candidate hierarchical statistical models to predict TTF_{Human} from $TTF_{groundTruth}$, some with a switch point, and some without.

The overall model formulation for models without a switch point was:

$$TTF_{human} = a + b * f(TTF_{groundTruth}) + \epsilon \tag{9}$$

Where the functions $f(TTF_{groundTruth})$ we considered were linear, square root, or logarithmic:

Functional Forms for the Hierarchical Models	
Functional Form Names	Formula
Linear	$f(TTF_{groundTruth}) = TTF_{groundTruth}$
Square Root	$f(TTF_{groundTruth}) = \sqrt{TTF_{groundTruth}}$
Logarithmic	$f(TTF_{groundTruth}) = \log TTF_{groundTruth}$

We now detail the priors for the parameters. Given the fact that we have little prior information about the parameters, we intentionally chose the uniform prior with a wide range to mimic the effect of uninformative priors and make the data to dominate the inference of the posterior distribution of the parameters. The intercept parameter is denoted as a; the slope parameter is denoted as b; the Gaussian noise is denoted as ϵ.

The priors were the following:

Priors for Parameters			
Parameters	Linear	Square Root	Logarithmic
a	Uniform(0,5000)	Uniform(-1000,1000)	Uniform(-30000,0)
b	Uniform(0,2)	Uniform(0,100)	Uniform(0,4000)
ϵ	Uniform(0,5000)	Uniform(0,5000)	Uniform(0,5000)

We fit the model at the individual level. Considering the same functional forms, each participant i had their own a_i, b_i, ϵ_i, drawn from Gaussian distributions over the hyperpriors above. The model at the individual level was formulated as:

$$TTF_{human_i} = a_i + b_i * f(TTF_{groundTruth}) + \epsilon_i \tag{10}$$

Individual parameters included:

Individual Priors for Parameters			
Parameters	Linear	Square Root	Logarithmic
a_i	Normal(a, 1250)	Normal(a, 250)	Normal(a, 950)
b_i	Normal(b, 0.3)	Normal(b,3)	Normal(b, 110)
ϵ_i	Normal(ϵ, 1000)	Normal(ϵ, 500)	Normal(ϵ, 1000)

Because the hierarchical linear model best accounted for people's behavior over short periods of time in Experiment 1, our switch point models below all included a linear component up until the switch point, followed by different possible behaviors past the switch point.

The model candidates with a switch point was a hierarchical model that included a switch point, and the model is formulated as:

$$TTF_{human} = a + b * TTF_{groundTruth} + c * f(TTF_{groundTruth} - SwitchPoint) * k + \epsilon \quad (11)$$

Where the functions $f(TTF_{groundTruth} - SwitchPoint)$ we considered were linear, square root, or logarithmic:

Functional Forms for the Hierarchical Models	
Name	Formula
Linear	$f(TTF_{groundTruth} - SwitchPoint) = TTF_{groundTruth} - SwitchPoint$
Square Root	$f(TTF_{groundTruth} - SwitchPoint) = \sqrt{TTF_{groundTruth} - SwitchPoint}$
Logarithmic	$f(TTF_{groundTruth} - SwitchPoint) = \log(TTF_{groundTruth} - SwitchPoint)$

The intercept parameter is denoted as a; the slope parameter is denoted as b; the slope modification depending on switch point position is denoted as c, the switch point position is denoted as $SwitchPoint$, and the Gaussian noise is denoted as ϵ.

We now detail the priors for the parameters. Given the fact that we have little prior information about the parameters, we intentionally chose the uniform prior with a wide range to mimic the effect of uninformative priors and make the data to dominate the inference of the posterior distribution of the parameters. The priors were the following:

Priors for Parameters			
Parameters	Linear	Square Root	Logarithmic
a	Uniform(0,5000)	Uniform(0,5000)	Uniform(0,5000)
b	Uniform(0,2)	Uniform(0,2)	Uniform(0,2)
c	Uniform(-10,10)	Uniform(-100,100)	Uniform(-4000,4000)
$SwitchPoint$	Uniform(2000,19000)	Uniform(4000,18000)	Uniform(4000,16000)
ϵ	Uniform(0,4000)	Uniform(0,4000)	Uniform(0,5000)

We fit the model at the individual level. Considering the same functional forms, each participant i had their own a_i, b_i, c_i, $SwitchPoint_i$, ϵ_i, drawn from Gaussian distributions over the hyperpriors above. The model at the individual level was formulated as:

$$TTF_{human_i} = a_i + b_i * TTF_{groundTruth} + c_i * f(TTF_{groundTruth} - SwitchPoint_i) * k + \epsilon_i \quad (12)$$

Individual parameters included:

Individual Priors for Parameters			
Parameters	Linear	Square Root	Logarithmic
a_i	Normal(a, 1250)	Normal(a, 1250)	Normal(a, 1250)
b_i	Normal(b, 0.24)	Normal(b, 0.24)	Normal(b, 0.24)
c_i	Normal(c, 0.24)	Normal(c, 19)	Normal(c, 200)
$SwitchPoint_i$	Normal($SwitchPoint$, 1000)	Normal($SwitchPoint$, 1000)	Normal($SwitchPoint$, 1000)
ϵ_i	Normal(ϵ, 1000)	Normal(ϵ, 1000)	Normal(ϵ, 1000)

We compared the six models using Leave-One-Out Cross Validation. We found that the model weights were: MW(Linear) = .0009; MW(Square Root) = 0; MW(Logarithmic) = 0; MW(Linear&Linear) = .809; MW(Linear&Square Root) = .190; MW(Linear&Logarithmic) = .0001.

The estimated hyperprior parameters (mean and 95% HDI) for the model candidate with most probability weights, which is the Linear model with a linear switch point position, were: a = 2490.173, with 95% HDI [1923.493, 2054.486]; b = 0.533, with 95% HDI [0.477, 0.59]; c = -0.201, with 95% HDI [-0.272, -0.134]; $SwitchPoint$ = 9419.097, with 95% HDI [8125.487, 10775.788]; ϵ = 1990.285, with 95% credible interval [271.102, 3604.735].

We next tested whether people's switch point is related to their mental capacity, as measured by the digit span task.

Participants' capacity in the digit task was an integer in the interval $[5, 9]$. We separated the participants in two groups: smaller or equal to 7 (72 participants), and greater than 7 (106 participants)

The Linear model with switch point position was fitted separately to the two groups' TTF. The estimated parameters (mean and 95% HDI) for the model fitted on participants with working capacity smaller or equal to 7 were: a = 2316.688, with 95% HDI [1908.577, 2728.295]; b = 0.554, with 95% HDI [0.345,0.767]; c = -0.233, with 95% HDI [-0.338, -0.124]; $SwitchPoint$ = 7362.606, with 95% HDI [6029.247, 8716.867]; ϵ =

4795.454, with 95% HDI [3.379, 9418.36].

The estimated parameters (mean and 95% HDI) for the model fitted on participants with capacity greater than 7 were: a = 2507.968, with 95% HDI [2179.564, 2835.759]; b = 0.542, with 95% HDI [0.474, 0.613]; c = -0.200, with 95% HDI [-0.300, -0.097]; $SwitchPoint$ = 10402.743, with 95% HDI [8795.350, 12050.770], ϵ = 4048.124, with 95% HDI [0.133, 9258.424].

After we inferred the switch point for participants in both groups, we conducted bootstrap test (1000 samples) over the difference between the mean of individual switch points for the participants within two groups. The bootstrapped mean of the difference between individual switch points for the participants within two groups is -3010.2ms, with 95% CI [-5391.1, -629.3], which does not overlap 0. This indicated that the difference between the switch points of the two groups is statistically significant.

In addition to the binary, median-split analysis, we also conducted an ordinal analysis in which we separated participants by their *digitSpan* score, into the 5-score, 6-score, 7-score, 8-score, and 9-score groups. The detail of the analysis and the result is documented in the Supplementary Materials under *Experiment 4 Additional Analysis*.

Bounded Fluid Simulation Model

The existence of the switch point is generally in line with the idea that people's capacity for simulation hits resource limit, and the relation of this switch point to the digit span task suggests a link with overall capacity limitations. However, the mechanism of the mental physics engine that is simulating the physical tasks in studies 1, 2, and 4, and how the mental physics engine runs differently past the switch point is not directly specified by the studies above. So, we next consider a specific mental simulation proposal for tying together the findings of Studies 1, 2, 4, and propose a hypothesis of how the mental physics engine might be running differently past the switch point found in Study 4.

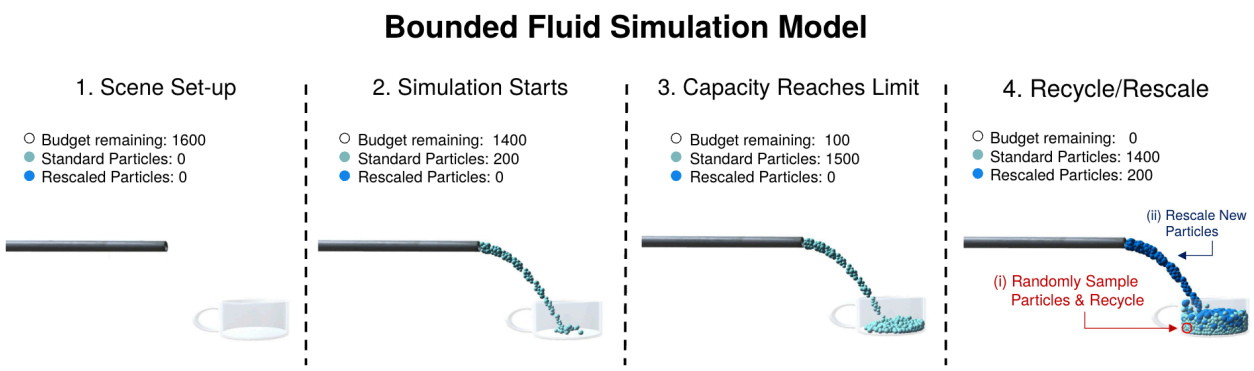

Figure 5

Illustration of The Bounded Fluid Simulation Model. *The proposed framework is modeled on a particle-based fluid simulation engine. We show four representative stages of the model unfolding over time. Stage one: the scene is set by stipulating an emitter and container. Stage two: the simulation starts, with the emitter producing standard particles, marked in light blue, that flow into the container. Stage 3: The number of particles used in the simulation is reaching the limit of the model's budget. Stage 4: Standard particles in the scene are randomly selected and recycled back into the budget. Any new particle put into the scene at this point is re-scaled (marked in dark blue) such that it is bigger than the particles emitted before the recycling step started. Larger particles represent greater uncertainty, with the particular scaling parameter used across the studies being fit to data. The flow rate does not change when recycling starts. The simulation continues until a criterion is met (e.g. the particles reach the top of the container, or the simulation is paused by an external signal.)*

Method

We implemented a Bounded Fluid Simulation model (BFS), using a particle-based fluid simulation engine, as shown in Figure 5. This was the same engine used for the creation of the stimuli, and it is generally in line with a common way of simulating fluids in real-time engines (Gregory, 2014).

The BFS model makes the following assumptions, which could be relaxed: First, the fluid particles move into the container at constant speed. Second, the velocity of the particles corresponds to the ground-truth velocity. Third, the container shape and size correspond to ground truth. In other words, our model does not solve the visual-reconstruction aspect of perception. Rather, it assumes that people's perception

accurately reconstructs a given visual scene, and that it is this representation that is handed off to mental simulation.

In this model, we are following the general principles of previous work on simulation models in intuitive physics, such as Battaglia et al. (2013) and Bates et al. (2019). More specifically, our model is based on Bates et al. (2019), which also used a particle-based simulation model to account for people's intuitive reasoning about liquids. We note, however, that the model in Bates et al. (2019) does not specify what happens if/when the budget of particles runs out, and did not tie this behavior to individual differences in general cognitive capacity.

Our BFS model inherits the standard parameters of a particle-based fluid simulation, and adds two additional parameters: $MaxParticles$, and $ParticleScaling$. The first parameter, $MaxParticles$, indicates the maximum amount of available particles for a given simulation, accounting for cognitive resource limitation. When the current number of particles in a simulated scene exceeds $MaxParticles$, the particles emitted by the source are recycled from the currently available particles in the scene, and scaled up by a fixed factor. The degree to which recycled particles are scaled is the second parameter, $ParticleScaling$. Larger particles represent greater uncertainty, as a small number of large particles coarsely approximates the behavior of a large number of small particles. The two parameters together potentially correspond to mental capacity.

The BFS model can estimate a container's Time-to-Fill by simply running a simulation forward, up to the point that a container is filled with particles. The model can also estimate the given water level of a container after some length of time, by transforming the amount of particles in a container to the corresponding water height. We assume that people's response corresponds to the output of the BFS model, up to a linear scaling (as the BFS model does not account for motor error and such).

We infer the parameters of the BFS model using the Random Walk Metropolis Algorithm. We formulated the joint posterior distribution of the parameters of the BFS

model conditioning on the experimental data, and taking into account the prior distribution of the parameters.

From the posterior distribution over the model parameters we derive point estimates of the parameters, which were then used to produce the final model predictions for Studies 1, 2, and 4. In addition to fitting the aggregate data for Study 4, we also separately fit the *MaxParticles* and *ParticleScaling* parameters for the two groups of participants who scored below-the-median and above-the-median on the digit span task. For more details about the parameter fitting, the inference method, and the model, please see the supplemental material.

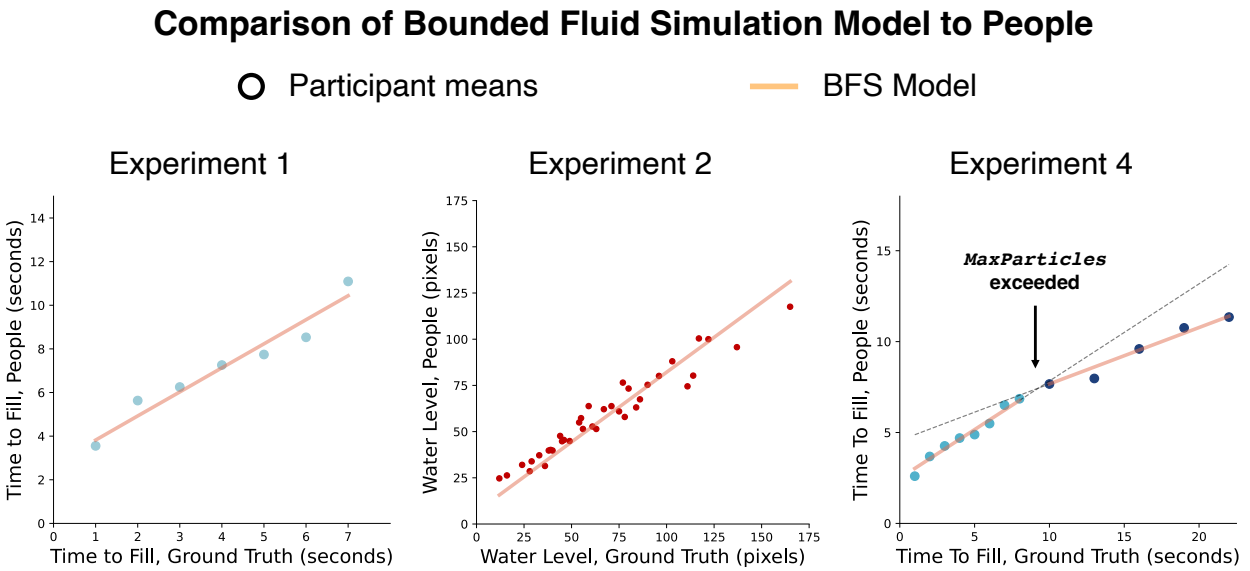

Figure 6

Model fits to participant data. *The Bounded Fluid Simulation model (BFS) uses a particle simulation with a maximum particle bound, beyond which particles are randomly recycled and scaled up to account for uncertainty. The model can relatively easily account for Experiment 1 and 2 where the bound is not reached and participants behave similar to a veridical simulation (up to a linear transformation). The model also accounts for the switch point in Exp 4, as the point at which maxParticles is exceeded.*

Results of fitting the BFS model

We first fit the BFS model to the data of Study 4, as this required the full range of model parameters (the other studies could be fit by a non-bound model). We found that

participant behavior was best explained by a BFS model with about 1940 standard-sized particles (95% CI of [1668, 2217]), and with a *ParticleScaling* parameter of about 2.7 (95% credible interval [2.43,3.01]). As can be seen in Figure 6, this parameter setting was able to recapture people's behavior, and account for the switch-point.

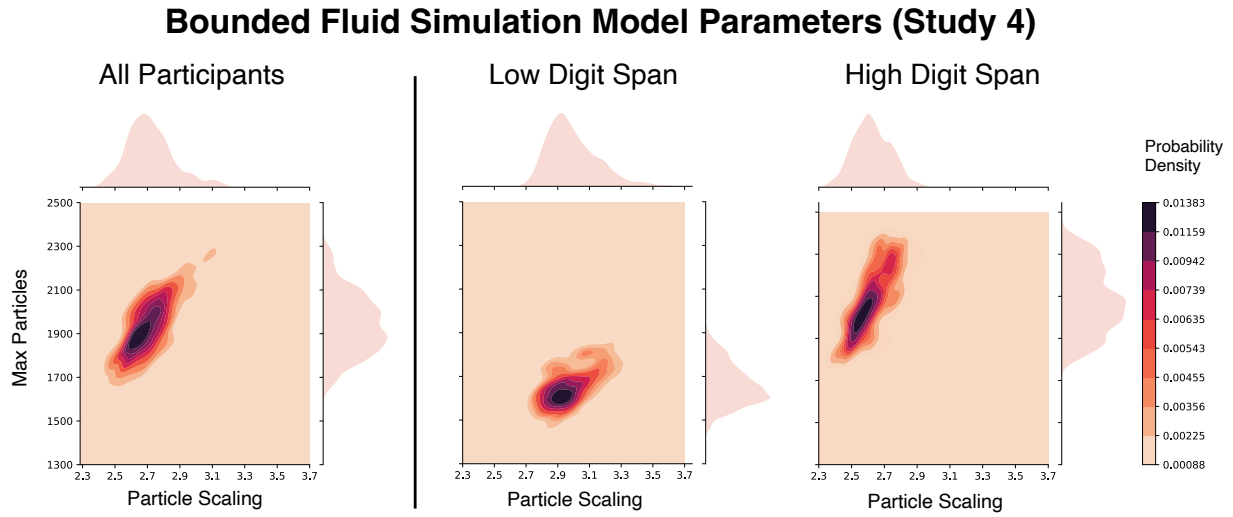

Figure 7

Posterior probability distributions over model parameters *Contour plots show the joint posterior probability distributions over the maxParticles and particleScaling model parameters for Exp 4, conditioning on participant data, both in aggregate (left), and split into two groups by digit span (right). Density plots at the top and right of contour plots show marginals over the model parameters.*

We next considered again the fit of the model parameters to the results of Study 4, but separately for the participants who scored above and below the median-split on the digit-span task. We found that for people with lower capacity (as measured by the digit span task), the best fitting model had approximately 1660 particles (95% CI [1477,1844]), and a *ParticleScaling* parameter of approximately 3.0 (95% CI [2.68,3.30]). For people with higher capacity on the digit span task, the best fitting model had approximately 2085 particles (95% CI [1779, 2387]), and a *ParticleScaling* parameter of approximately 2.6 (95% CI [2.40, 2.84]). Both parameters indicate that people who showed a higher capacity on the digit span task corresponded to greater capacity in the BFS model. Contour plots visualizing the joint posterior distributions and the corresponding marginal distributions of

MaxParticles and *ParticleScaling* under the conditions mentioned above are shown in Figure 7.

General Discussion

We examined people's ability to reason about the everyday behavior of containers and liquids, and found that they were reasonable over short periods of time. People linearly adjusted their response in a task that asked them how long it would take to fill a container with water (Time-to-Fill), and a task that had them estimate how full a container was after a random time duration (Water Level). People's responses were affected by the *volume* of the container, but not the *shape* of the container (tall/regular/wide). When estimating volumes statically, people exhibited systematic biases not present in dynamic tasks. Over longer periods of time, people's behavior showed a switch point, with the best-fitting model suggesting different responses before and after the switch. The overall pattern of people's results supports the theoretical position that people are using a resource-limited mental simulation when reasoning about intuitive physics.

While our work focused specifically on reasoning about liquids, it speaks to bigger theoretical issues of ongoing current interest to cognitive science. There is currently no agreement on the basic format of the computations underlying people's intuitive physical reasoning. Several researchers have proposed mental simulation as one cornerstone of this reasoning (e.g. Battaglia et al., 2013; Ullman et al., 2017), but other researchers have raised serious concerns and issues for this account (e.g. Davis & Marcus, 2016; Ludwin-Peery et al., 2021, 2020; Marcus & Davis, 2013). Many of the concerns branch from the same root: a veridical mental simulation should correctly align with reality, whereas people in fact deviate from reality. In response to this, supporters of mental simulation suggest that mental simulation may make use of principled approximations, and that such approximations lead to systematic quantitative and qualitative deviations (Bass et al., 2021). In this work, we elaborated on the idea that people may be using approximate mental simulations, and examined specifically the notion that mental

simulations may be resource-bound in their object-tracking capacity.

People's overall performance across our tasks can be explained by a bounded simulation model, which is able to reasonably align with reality, but has a limited budget of 'particles' beyond which it must rely on coarser approximations. We take these limited particles to stand in for an overall mental capacity, without committing strongly to whether the specific issue is due to working memory, attention, uncertainty, or a combination of factors. In this, we follow previous previous work on capacity limitations in object tracking (Vul et al., 2009). Also, we note that our computational bounded simulation model found that people's budget on average corresponds to roughly 2,000 particles, but stress that such numbers should be treated with caution. Our model was strongly influenced by Bates et al. (2019), who found that a few hundred particles (or even fewer) best account people's data on reasoning tasks that involve liquids. We do not view this as a discrepancy, as the specific absolute number of particles estimated will depend (within reason) on details such as the baseline particle size, the physics engine software used, task context, and so on. We put 'within reason' in parenthetical, as even with such considerations we would not expect any engine to do well with, say, 5 particles, and similarly we would find it a discrepancy if someone suggested their model found people needed a million particles. Our more important contributions here are not the specific absolute number of particles estimated, but the tying in of this number to other cognitive capacity limits, and the difference before and after a particle budget runs out. Interestingly, one could also consider two other possibilities on how our bounded simulation model could be affected. First is whether people's judgment about time could be affected by cognitive load or differences in cognitive capacity; Second is whether people's simulations could be affected by the different flow rates of the liquid. We detailed experiments examining those two possibilities in the supplementary material, labeled as *"Experiment S1"* and *"Experiment S2"*. In short, we did not find that people's judgment about time could be affected by cognitive load or differences in cognitive capacity, and

people's simulations are not affected by the different flow rates of the liquid.

In addition to the bounded simulation model, we considered the alternative that people are using simple visual routines and/or heuristics to first estimate the volume of the containers, and then combine that estimate with a time-estimate to arrive at the final estimates for when a container is full, and how full it is. This perceptual-volume estimation proposal encounters several challenges given our data: First, a simple version of such a proposal does not explain the moment-by-moment reasoning needed for estimating water-levels in Study 2. However, the simple volume estimate can be amended to overcome this. A bigger challenge is to explain the existence of a switch point in reasoning over longer periods of time, and no simple amendment presents itself. Another big challenge is that a direct volume-estimation task (Study 3) shows people have systematic biases in their volume estimation of the containers used in our study. If people are relying on perceptual volume estimation alone, such a bias should show up in the other tasks, and it does not.

The fact that people differ in their responses when estimating volume from static scenes (Study 3) compared to the dynamic scenes (Studies 1, 2, 4) suggests that people are relying on different cognitive processes for these different tasks. But, this also presents a puzzle: The visual perception of a container should happen earlier in the perception/reasoning pipeline than a mental simulation. Even if people are relying on mental simulation, this simulation should inherit its starting point from the visual estimate. So, if the visual estimate is biased, how come the simulation is not biased? One possible solution to this puzzle would be that an object's perceived volume is not an inherent, explicit part of its 3D representation. That is, vision may accurately re-construct a 3D representation of an object, but answering the question 'what is the volume of this object' could require additional computations over this representation. As a simple example, consider that a cylinder could be represented in a graphics program by its radius and height (r, h). Such parameters would be explicitly and readily available, and are sufficient for constructing the cylinder visually, and for having the cylinder interact with

other objects. But, nowhere in the program is there the volume of the cylinder. Such an example is hardly unusual, this is the case with the graphics program used to generate the stimuli for our studies, for example. One *could* compute the volume of a cylinder given its radius and height, but this would be an extra computation, and for oddly-shaped bodies it would be a non-obvious computation that may call for heuristic approximations.

Our work sheds light on people's use of mental simulation in general, and its use under limited resources more specifically. It also connects to a broader research interest in people's ability to mentally 'evolve' the state of a representation. In particular, it is interesting to consider it in light of the well-studied phenomenon of 'representational momentum' (Freyd & Finke, 1984; Hafri, Boger, & Firestone, 2022; Hubbard, 2005). This term refers to the finding that people continue to evolve and extrapolate the state of mental objects and representations beyond their perceived states: a moving body that disappears from view is remembered as being further along than it actually was; ice melting is remembered as more melted than it was last seen, and so on. Given that people generally move objects forward in state space beyond where they were last perceived, why do people in our experiments seem to overall 'lag behind' the expected ground truth? The answer is that these two methods of examining mental extrapolation are different. In the case of representational momentum, people are asked to remember a state S at time T, and the finding is that they recall an S' that corresponds to a T+t for some positive t (they evolved S to S' even when not asked to do so). In our case, we showed people a state S at time t, and explicitly asked them to evolve it further, at the same time that S itself is indeed changing. The possible delay between the true S' and people's S' relates to the pace at which people evolve S to S', which is orthogonal to representational momentum itself, and could in principle have proceeded at a pace that is slower than, similar to, or faster than the ground truth. It is interesting to consider why the pace of simulation was the way that it was, independently of representational momentum, but the two findings are not in conflict.

Our work leaves open diverse directions for future research, and we consider a few of them in turn. First, while people were not significantly affected by different object shapes (of equal volume) in our studies, this may not hold in general. The objects we considered were relatively simple and regular, and oddly shaped bodies may lead to various biases and deviations from the ground truth. In particular, other work suggests that people may be using approximate bodies for physical reasoning and tracking (Li et al., 2023; Ullman et al., 2017). If this is true, people may consider only approximate bodies for fluid simulations. The use of approximate bodies would not be in contrast to the BFS model, and its formulation does not rely on people being unaffected by a container's shape. Second, and more directly relevant to the BFS model, the details of how people behave past the switch point need to be investigated further, both empirically and from the modeling stand-point. Our best-fitting statistical analysis heavily argues in favor of a switch point, but it places much less certainty on the specific functional form past the switch point. There are different possible ways a principled-approximation model could handle its limited budget running out, and these may be context sensitive. Third, whether their estimation of shape and flow matches reality or not, we assumed a static model in which the relevant physical parameters are available as soon as the scene is perceived. It is likely that people go through an initial period in which they estimate the physical parameters of a given scene (Ullman et al., 2018), including in particular the flow rate. While the flow rate was constant throughout our stimuli and the estimation was likely done in the first moments of the first scene and then generalized, it would be interesting to consider a more complete model that also begins by estimation of physical properties. A more general test of the BFS' recycled use of a limited number of particles could examine people's tracking of a liquid as it moves and flows (similar to Bates et al., 2019). The use of 'larger' particles (representing greater uncertainty) past the limits of the particle-budget can lead to systematic deviations regarding the predicted end-point of a majority of the liquid, for example.

Beyond variations that target the bounded-simulation model's details, a major direction for future research would be to examine the combination of a simulation-based model with more cached-based behavior. The idea that people use memory-based heuristics, simplifications, and abstractions alongside more detailed stimulation processes or world models has been a topic of formal and empirical study in different cognitive domains, including for example decision making (Stewart, Chater, & Brown, 2006), intuitive physics (Sosa, Gershman, & Ullman, 2025; Ullman et al., 2018), and hypothesis generation (Dasgupta, Schulz, & Gershman, 2017). In this specific case, it may be that people deploy more cached-based reasoning when considering tasks involving liquids and volume that occur within everyday time-frames, but switch to more simulation-based reasoning when this cache is not available. Such an idea is also intriguingly close to the notion of 'cached simulation' specifically for the simulation of fluids, which exists in many current engineered systems (Gregory, 2014). While we do think this is a useful direction for research, we note that a simple notion of 'switch from cache to simulation' model does not on its own easily quantitatively account for many of the specific findings in our current investigation, including the relationship between the switch point and independently-derived limits of mental resources, and so would need explication to be able to consider it formally as an alternative.

The books of Psalms gives us a well-known image of abundance and gratitude: *My cup runneth over*. Our results suggest that people indeed have what they need to mentally simulate the physical world, up to a point.

Constraints on Generality: Our data was collected within a representative US population, using online data collection. The generality of the findings is therefore constrained by the population that completed our studies, and further work is needed to definitively establish them outside this context. However, given the low-level perceptual nature of our studies and findings, we have no reason to believe our findings are specifically affected by US cultural norms and practices.

References

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

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Balaban, H., & Ullman, T. (2024). The capacity limits of tracking in the imagination.

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

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

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

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., ... others (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive psychology*, *96*, 1–25.

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, *113*(34), E5072–E5081.

Forbus, K. D. (1997). *Qualitative reasoning.*

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 126.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936.

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 372.

Gregory, J. (2014). *Game engine architecture*. AK Peters/CRC Press.

Hafri, A., Boger, T., & Firestone, C. (2022). Melting ice with your mind: Representational momentum for physical states. *Psychological Science*, *33*(5), 725–735.

Hespos, S. J., Ferry, A. L., Anderson, E. M., Hollenbeck, E. N., & Rips, L. J. (2016). Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychological science*, *27*(2), 244–256.

Hubbard, T. L. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic bulletin & review*, *12*, 822–851.

Huntley-Fenner, G., Carey, S., & Solimando, A. (2002). Objects are individuals but stuff doesn't count: Perceived rigidity and cohesiveness influence infants' representations of small groups of discrete entities. *Cognition*, *85*(3), 203–221.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

Lerer, A., Gross, S., & Fergus, R. (2016). Learning physical intuition of block towers by example. In *International conference on machine learning* (pp. 430–438).

Li, Y., Wang, Y., Boger, T., Smith, K. A., Gershman, S. J., & Ullman, T. D. (2023). An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*.

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

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken Physics: A

Conjunction-Fallacy Effect in Intuitive Physical Reasoning. *Psychological Science*, *31*(12), 1602–1611. doi: 10.1177/0956797620957610

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*, *24*(12), 2351–2360.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*(5), 1139–1141.

Paulun, V. C., Kawabe, T., Nishida, S., & Fleming, R. W. (2015). Seeing liquids from static snapshots. *Vision research*, *115*, 163–174.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Piaget, J., & Inhelder, B. (1974). *The child's construction of quantities: Conservation and atomism* (Vol. 2). Psychology Press.

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, *6*(9), 1257–1267.

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2018). IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning. *arXiv:1803.07616 [cs]*.

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

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

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

Processing Systems. Vancouver, Canada.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Sosa, F. A., Gershman, S. J., & Ullman, T. D. (2025). Blending simulation and abstraction for physical reasoning. *Cognition*, *254*, 105995.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, *53*(1), 1–26.

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

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive psychology*, *104*, 57–82.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558.

Van Assen, J. J. R., Barla, P., & Fleming, R. W. (2018). Visual features in the perception of liquids. *Current biology*, *28*(3), 452–458.

Vul, E., Alvarez, G., Tenenbaum, J., & Black, M. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in neural information processing systems*, *22*.