Priors, Progressions, and Predictions in Science Learning: Theory-Based Bayesian Models of Children’s Revising Beliefs of Water Displacement

Joseph Colantonio, *Rutgers University - Newark*, Igor Bascandziev, *Harvard University*,
Maria Theobald, *DIPF*, Garvin Brod, *DIPF*, Elizabeth Bonawitz, *Harvard University*

***Abstract*— Despite sometimes noisy evidence (e.g., perceptual processing errors), young children are capable of predicting and evaluating events based on complex causal representations. Children rapidly revise their beliefs and learn scientific concepts - sometimes without prior knowledge of an underlying causal system. What might we need in our computational models of belief revision to similarly simulate children’s behaviors when learning such causal systems? Building from experimental data of elementary school children’s intuitive beliefs and predictions of water displacement, we propose three aspects of human inference and belief revision that warrant attention within the subfield of computational cognition. Each aspect is described by identifying the gaps between empirical findings and current computational implementations. Then, specific implementations of these aspects are built using models of Theory-based Bayesian inference. First, we construct children’s prior beliefs at the individual level based on their prior behavior. Second, we approximate children’s learning using an “optimal” Bayesian model, revealing the dynamics of belief revision trial-by-trial. Third, we investigate the role prediction may have in facilitating learning. By performing these key computational steps, we find support for contemporary claims that children may be approximately “Bayesian” learners and increase awareness of the importance of generating predictions in active learning.**

***Index Terms*— cognitive system and development, prediction, belief revision, Bayesian inference, computational modeling, science learning, perceptual processes**

I. Introduction

Endeavors in robotics and machine learning that focus on associative, reinforcement, and model-free learning enjoy the flexibility of statistical learning without constraints. However, these “bottom-up” learning models typically struggle to construct the theories that support causal reasoning on their own, and they may not capture the human ability to construct and revise causal beliefs in a rapid fashion [1]. Yet, theory construction and revision are critical components of human learning. Abstract causal theories of the world are a foundation of human intelligence, supporting inference, prediction, counterfactual reasoning, explanation, and action planning already present in young children [2-4]. Children’s causal beliefs flexibly change in the face of new, even ambiguous evidence [5]. Beliefs about the world begin to be revised as early as infancy, as seen in studies showing that infants begin enriching “core” concepts about object solidity, continuity, persistence, and causality over the first few years of life (e.g. [6-9]).

How might belief construction and revision take place in human learners? Intuitive theories — such as children’s theories about the relationship between mass and balance [10-12], biology and vitalism [2, 13], the principles of magnetism [14], buoyancy [15,16], or about solids and liquids [17] — can take the form of a weighted space of prior causal beliefs, helping to explain how a learner could draw rich inferences quickly from limited, ambiguous data [18,19]. These intuitive theories may also include misconceptions – possibly due to their ability to explain a limited range of specific, observed phenomena [20] and due to misconceptions’ ability to coexist with other beliefs [21,22], sometimes even after learning the correct scientific concept [23]. Thus, beliefs may be captured as a space of multiple, competing prior beliefs, sometimes imposing constraints on the kinds of hypotheses a learner considers in the first place and offering another means to faster learning [24,25].

Computational models of belief revision help to precisely characterize the process by which belief construction and revision are possible, despite correct beliefs competing with the influence of misconceptions. However, even models of human cognition that consider top-down, theory-based constraints to predict the learning behavior of groups, on average, may take for granted that learners (and in particular children) have tremendous variability in their starting states (e.g., [26]). The impact of individual differences may be most apparent if

September 18th, 2021. This work was supported in part by the German Research Foundation (DFG, BR 5736/2-1) to Garvin Brod. Garvin Brod was supported by a Jacobs Foundation Research Fellowship.

J. Colantonio is with Rutgers University - Newark, Newark, NJ 07102 USA (email: colantoniojoseph@gmail.com).

I. Bascandziev is with Harvard University, Cambridge, MA 02138 USA(email: igb078@mail.harvard.edu).

M. Theobald is with DIPF - Leibniz Institute for Research and Information in Education, 60323 Frankfurt, Germany (email: theobald@dipf.de).

G. Brod is with DIPF - Leibniz Institute for Research and Information in Education, 60323 Frankfurt, Germany (email: brod@dipf.de).

E. Bonawitz is with Harvard University, Cambridge, MA 02138 USA (email:elizabeth_bonawitz@gse.harvard.edu).

.

learning is assessed trial-by-trial, in a “mini-microgenetic” experiment [27]. If prior beliefs play an important role in human and machine learning, then models must consider how differences in these initial starting states influence learning at the individual level. Thus, the first two goals of this paper are to capture the individual starting states of the learner’s prior beliefs and predict learning trial-by-trial.

Simply capturing these prior beliefs and assuming all learners deploy them approximately-rationally is likely not sufficient to capture the nuance of individual differences in learning. To steer learning, these prior beliefs have to be activated and leveraged [28]. In line with this claim, a wealth of research has found that conflicting evidence alone is not sufficient for successful theory revision [29]. Specifically, successful belief revision may require acknowledging the strength of their prior beliefs relative to the conflicting evidence – with findings that show that children who already have an initial belief about a topic may only revise it if an expectancy violation cannot be “explained away” by their current beliefs [30,31]. Thus, instead of only *passively* observing evidence that may be conflicting with one’s beliefs, recent experiments find that asking learners to *actively* generate an explicit prediction has been shown to increase the perceived conflict between the learner’s theory and unexpected evidence, and it may facilitate theory revision [32-34]. Thus, making a prediction may be a key process by which humans actively leverage their prior beliefs [28]. The degree-to-which prediction-based learning engages theory-based “optimal” belief updating remains open.

Thus, we raise and attempt to answer three critical questions. First, how can we best understand the initial models that young learners start with and leverage throughout learning? Second, how does this learning process unfold over time, as new observations are acquired trial-by-trial? Third, in what circumstances do learners engage their models of the world to support learning in a top-down fashion? If we are to reverse engineer the human learner, then we must understand the role of the starting state of the causal-explanatory representations, how they change in response to evidence, and the contexts that engage model-based learning. In what follows, we will provide a brief background on the role of prior knowledge and prediction in learning, our theory-based Bayesian model approach, and results modeling empirical data from prior developmental research [35] towards answering these questions.

A. Prior knowledge in learning and development

Incorporating structured prior beliefs into models is necessary to capture the speed of learning characteristic of humans but has often been undervalued in past approaches. Bayesian inference models have gained traction, highlighting the importance of considering multiple competing hypotheses and intuitively performing live updates to the strength of belief in each of those theories. While associative models may capture how learners can update the relative strengths between cause and effect in a linear causal relationship, Bayesian models allow for inference of different types of causal relationships and whether different theories of various causal relationships are even being considered [15,36]. Thus, as interest has grown in modeling more complex cognitive processes - such as causal reasoning and theory revision - Bayesian inference has become quite prevalent despite being computationally costly. This is likely due to the parsimony of its probabilistic principles while still maintaining the ability to describe complexities.

A growing field of theory-based Bayesian modeling has begun to address the problem of complexity and search over large spaces, acknowledging the role of prior intuitive theories [37-39]. But to capture the intricacies of human learning, these approaches must also attend to how beliefs are constructed at the individual level, with careful attention to the individual differences that characterize the variability of prior beliefs in early development. For example, past research [10] measured children’s intuitive beliefs of balance before giving children the chance to explore further and explain the observation. The observed evidence was not sufficient to predict how children would behave. Instead, it was the interaction of the children’s prior beliefs of balance and the evidence that predicted children’s behavior - with children exploring and explaining events that conflicted with their individual prior beliefs. Such results highlight the need to measure and model differences in prior beliefs at the individual level.

B. Bayesian Models of Learning

Given the power of Bayesian inference to formally model the role of individual difference in prior beliefs and the step-by-step learning process during theory revision, we apply this approach to our analysis. In Bayesian inference, operations can be performed sequentially with prior theories influencing behavior before each new observation occurs. Priors are revised and updated at each step - generating posteriors that are now used as the priors from a subsequent event similar to the first one. Typically, this is formalized computationally using Bayes Rule:

$$p(h|d_1,\ldots,d_{n+1}) = \frac{p(d_{n+1}|h)p(h|d_1,\ldots,d_n)}{\sum_{h_i \in H} p(h_i|d_1,\ldots,d_n)}$$

where we calculate the updated posterior ($p(h|d_1,. . .,d_{n+1})$) of some hypothesis (h) after some number (n) of data observations ($d_1,. . .,d_{n+1}$), given a prior probability of said theory ($p(h|d_1,. . .,d_n)$).

Bayesian models thus account for prior beliefs through updating the valuation and relevance of said priors trial-by-trial as individual observations are experienced continuously. While there is a long history of investigation of children’s early causal reasoning and intuitive scientific theories [2,10,40-42], studies tend to present all available data at once and determine whether children can successfully parse this influx of information to successfully converge onto the correct theory being investigated. While this past work provides important insights into children’s development of inference skills, it lacks in its ability to track the theory revision process in a continuous way. Thus, a recent method in cognitive science - a “mini-microgenetic” method [14,27] where evidence is provided trial-by-trial and prior beliefs are investigated on each trial - is being used to understand concept learning both quantitatively and qualitatively using Bayesian modeling. With Bayesian modeling methods, it becomes possible to analyze behavior in a continuous way, providing opportunities for the discovery of the types of transitions humans may make when shifting the strengths of their theories in an approximately Bayesian way.

C. Engaging model-based learning through prediction

In supporting meaningful advances in studies of both human and machine learning, investigating the role of prediction in learning scenarios may be vital for understanding theory revision during science learning. Research on science education highlights the importance of having learners engage with their prior beliefs, as doing so may highlight inconsistencies between the learners' beliefs and the conflicting evidence that they must now accommodate to satisfy their desire to learn [43]. For example, recent work points to a beneficial role of asking learners to generate a prediction before presenting them with new information that conflicts with their intuitions [32,44]. The act of predicting is thought to increase the subjective value of the outcome, which—if the outcome is different than expected—increases the perceived expectancy-violation, resulting in enhanced attention and learning [45]. Here we take up the focus of investigating the role of prediction computationally.

We aim to investigate the role of prediction using Bayesian models of theory change. In what follows, we compare this "optimal" learning model's performance to experimental data on children's behavior in a water displacement task [35]. Importantly, this entails comparing the expectations of our optimal model between two closely matched scenarios centered on the role of prediction- one scenario where children are predicting, and perhaps performing active engagement with their prior theories, and another where children are performing post hoc evaluations.

D. Theory Change & the Domain of Water Displacement

The domain of water displacement was chosen for this investigation as it is suitable for use in answering our three main questions. First, past research has found notable individual differences in children's prior theories about water displacement. Children tend to have different beliefs about the principles of water displacement [46-49]. This allows for investigation of the variability across individual children's beliefs, as well as their impact on subsequent learning.

Second, to learn the principles of water displacement, learners can be presented with a series of similar trials and receive immediate feedback on their answers. This allows for the implementation of the continuous, trial-by-trial revision of children's prior theories in our computational models that involves the discrimination of perceptually distinct - but possibly noisy - representations. Third, for our modeling, we used data from an experimental study that manipulated children's level of engagement with the task [35]. To increase engagement, half of the children in the sample were asked to generate predictions about which of two balls displaces more water before the correct answer was revealed ("Prediction condition"). The other half of the sample was first presented with the correct answer and only then stated which outcome they would have predicted ("Postdiction condition"). Thus, we may investigate whether this difference in engagement (either generating a-priori predictions or post-hoc evaluations) affects the learning process. More specifically, we tested whether children in the prediction condition performed more similarly to an optimal theory-based Bayesian learner as compared to children in the postdiction condition. In what follows, we describe the aforementioned dataset and how it was generated, followed by the computational implementation of our Bayesian Learning model and subsequent results per this model's predictions.

II. Empirical Methodology and Behavioral Data

For the present study, we used data from an experiment that investigated elementary school children's theories of water displacement (experimental procedure, data, and empirical results are those found in [35]). Children's causal beliefs of water displacement were chosen as children frequently have the misconception that water displacement depends on the weight of an object or a combination of weight and size rather than on its size only (see e.g., [47]), providing an appropriate domain for the investigation of variability across individual children's beliefs, as well as their impact on children's subsequent learning.

Ninety-four six- to nine-year-old children participated in the experiment. Children first viewed a clip of an experimenter demonstrating how water gets displaced by pressing a sphere underwater - where, importantly, the experimenter stressed that the spheres were held underwater to avoid the chance that children evaluate buoyancy instead of water displacement. Importantly, while this was the only instance of children seeing actual water displacement during the entire experiment, no concerns were raised regarding the quality of the computerized stimuli throughout the task. This is because previous literature (originating with [47], but noted by work following it) demonstrates that not even hands-on experimentation (where children could raise and lower the compared objects in two separate containers of water) seemed to be what engaged students enough to reconsider their models of displacement. Instead, work related to water displacement and revision of related beliefs supports the notion that it is instead instances of contradictory information that might be eliciting learning. Following this familiarization clip, children completed a pretest, learning phase, posttest, and transfer test (in this order). Here, we will focus on data from the pretest to model children's prior theories, and on data from the learning phase to model children's learning over time. The data from the posttest and transfer test were not modeled.

All of the trials during each phase of the main experiment were similar to one another. However, the pretest was completed using paper and pencil. Meanwhile, the learning phase was completed with a computerized version of the task and additionally included feedback for children's responses and usage of eye tracking apparatus for collecting an additional measure, pupil dilation.[1] During each trial, children were presented with pairs of spheres that varied in size (small, medium, large) as well as in material (polystyrene, wood, lead), and therefore also varied by weight (assumed via a combination of the object's size and material). For each trial, children indicated which of two spheres they think displaces more water with a confidence scale from one to five (1 = certainly the left sphere, 2 = maybe the left sphere, 3 = equal amounts of water

[1] As this is not the focus of the current manuscript, it is not discussed further.

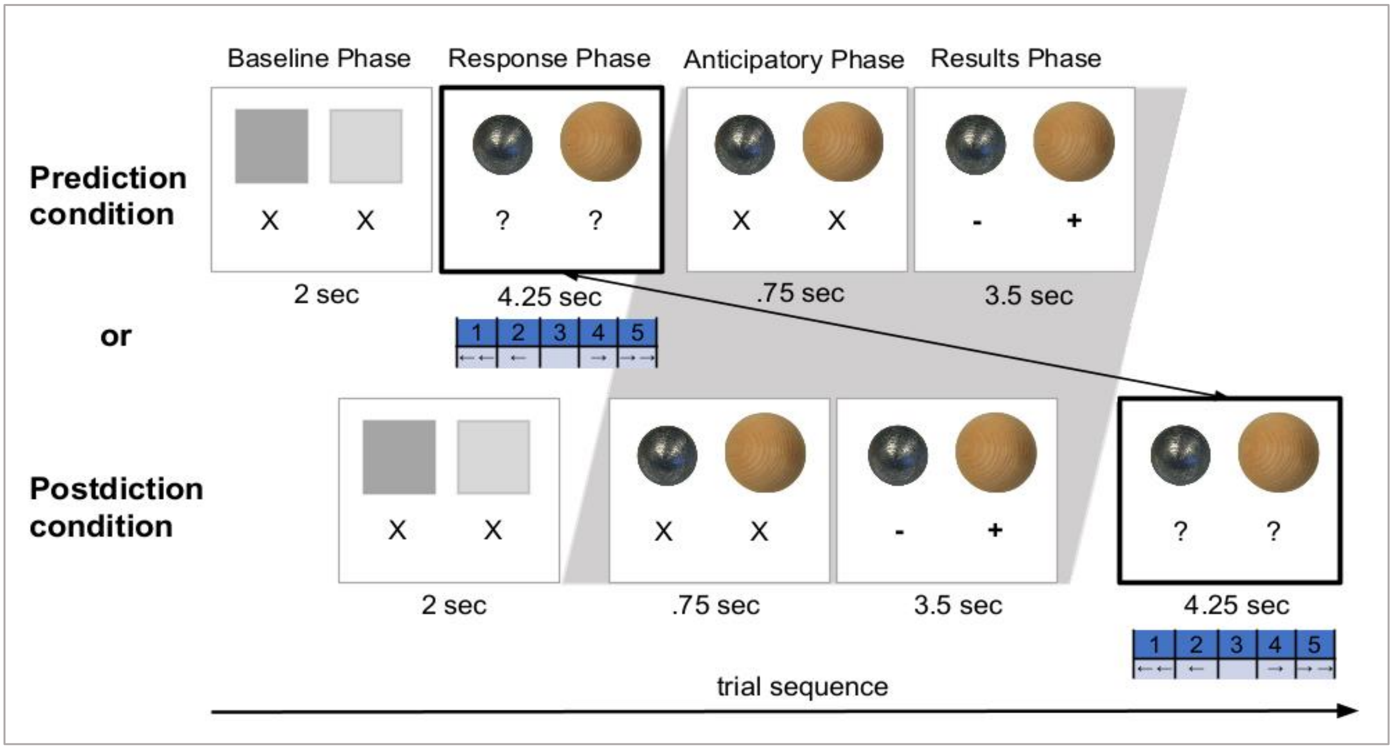


Fig.1. Schematic overview of how the experimental data being modeled was acquired from the original experimental manipulation (Theobald & Brod, 2021). The learning phase consisted of 34 trials that were presented to children that were randomly assigned to one of two experimental conditions (between-subjects): the Prediction or Postdiction condition. The figure shows an example trial that is shown to children in both conditions, but with emphasis on differences in the timing of children's response and the timing of related feedback. Specifically, children in the Prediction condition would first provide a response – stating their expectation about which sphere displaces more water before the results of the event and related feedback were presented to them. In contrast, children in the Postdiction condition first saw the results of the presented trial, then stated their expectations by providing an evaluation

for both, 4 = maybe the right sphere, 5 = certainly the right sphere; see Figure 1 for a trial example).

The pretest contained 8 trials as described above, asking children to evaluate which of two spheres displaced the most water with some confidence (per the described scale, from 1 to 5). After responding on each trial, the next trial of the pretest phase began. Critically, during this pretest phase, children did not receive feedback on the accuracy of their responses. This allows us to derive representations of children's prior theories on water displacement without the risk of evidence-based revision occurring part-way through these trials.

After the pretest, children completed a learning phase consisting of 34 trials that were similar to the pretest trials. The critical difference between the learning and pretest phase was that learners received feedback ("+" under the correct response) on their answers. The learning phase thus measured children's responses on each trial and provided feedback on the correct outcomes, providing the evidence needed for children to learn the correct theory of water displacement (that the size of the objects determines how much water it displaces). This allowed us to model the process by which children may be learning the true concept of water displacement as they incorporate the continuously incoming evidence trial-by-trial.

For the experimental manipulation, children were randomly assigned to either a Prediction (*n* = 48) or Postdiction (*n* = 46) condition that differed only in the learning phase trials on the computer. For all trials, feedback on the outcomes of each trial were presented to children by highlighting the correct option on the computer screen (a "+" would appear under the ball that would displace the most water). However, the timing at which children received feedback differed between conditions. In the Prediction condition, children predicted the outcome of a trial before receiving feedback. In contrast, the children in the Postdiction condition saw this feedback of the trial's correct outcome before responding. This experimental manipulation served to influence children's engagement with the task.

Here, we predicted that children in the Prediction condition would engage more in the task as they had to actively choose and commit to an option [28], while children in the Postdiction condition would engage less because they only made post-hoc evaluations after they already knew the correct outcome. Our goal was to gain insight into differences in the dynamics of learning processes in theory revision that may vary by the level of engagement that children have with their prior intuitive theories. For this purpose, we will compare model predictions to children's performance between these two conditions.

III. Bayesian Learning Model of Water Displacement

Here, we discuss the design and implementation of a computational endeavor that involves three main modeling stages. The first modeling stage entails capturing different representations of children's possible prior beliefs. Then, the second modeling stage aims to capture the process by which they evolve throughout learning. Finally, the third modeling stage entails a comparison of learning outcomes under different contexts of task engagement. In what follows, we outline the computational methods implemented for these three stages, followed by results that relate the predictions made by our computational model to children's actual behavior during the described experiment.[2]

A. Capturing a space of prior models of water displacement

The first goal of our model is to capture how children's prior beliefs (at the individual level) might motivate their predictions during the water displacement task. Here, we consider a hypothesis space distribution of four different, competing theories based on object-feature variance that prior literature has typically identified children maintaining in the early elementary school years (e.g., [46,47]). Critically, this past research highlights that while children's theories of scientific concepts (such as water displacement) are complex, they construct them based on their observations and the specific information that they can parse from them. Specifically, when children complete water displacement tasks, they typically make decisions or include descriptions of their naive theories by emphasizing the importance of object weight (e.g., its "Mass") or density (e.g., its "Material) as the determining factor for how much water is displaced, before eventually learning that the object's volume ("Size") is the correct feature to investigate.

Per these prior findings, object discrimination within our model relies on balancing three competing theories that may determine how much water an object displaces: a Size theory, a Material theory, and a Mass theory. We also include a fourth comparison to a Random Guessing model that captures a phase in which children might have no intuitions about the correct outcome and so respond randomly.

The first three theory-based models generate responses based on their respective feature values, such that objects with the "higher" value within the respective feature will "win" and displace more water (e.g., the correct Size theory will almost always predict that the larger object will displace more water). The Random guessing theory generates predictions from a uniform distribution, placing equal weight on each of the five potential outcomes of each trial.

Operationally, our model starts by assigning feature weights for objects that are being judged based on their feature variance. Specifically, for any given trial, two balls are compared to one another in terms of how much water they may displace. Here, we consider variations of three key features of the balls: their Size, Material, and Weight. A ball's Size (Small, Medium, and Large) and Material (Polystyrene (Styrofoam), Wood, Metal) are assigned possible feature values from 1 to 3 based on their relative feature (e.g., a Medium-sized Metal ball has a Size of 2 and Material of 3). A ball's Weight is assigned a value based on the product of its Size and Material values. For example, a Medium-sized (2) Metal ball (3) will have a Weight of

$$2 \times 3 = 6.$$

Critically, our models also account for perceptual noise when discriminating between objects. The models incorporate Weber's Law [50] as a cause of perceptual discrimination noise between two objects on any given trial. Noise was added to the perceptual task by multiplying the feature values being compared (e.g., a "Size" of 1 versus a "Size" of 2) by 0.22, motivated by past research on the Weber's ratio for children in this data's collected age range [51]. Importantly, recent research has found that perception as affected by Weber's Law tends to be similar across features and modalities (e.g., time, space, quantity; [52]). Thus, the model of each belief will allow for potential errors in decision making (choosing the wrong answer despite a clear winner) and representations of uncertainty (the ability to claim a tie; that both options displace equal amounts of water) when this object discrimination is performed.

An example trial of how each of these competing beliefs may generate predictions of a trial's outcome can be found in Figure 2. For the models of the Size, Material, and Mass theories, we performed 1000 pairwise comparisons of two random samples from normal distributions. The parameters of the normal distributions are defined based on the earlier described feature values, where the feature value (e.g., of a Large Ball with Size of 3) is taken as the mean, and the feature value affected by the Weber's ratio as the standard deviation (SD), where

$$SD = Feature\ Value\ \times Weber's\ Ratio = 3\ \times 0.22.$$

Here, samples would either "win", "lose", or "tie" with one another, based on the overlap of the two distributions. Then, a final tally is taken of the comparison results, generating a new distribution analogous to the children's confidence ratings (from 1 to 5). For the Random Guessing model, a Uniform Distribution was used instead, such that each potential choice (1 to 5) held equal weight, such that

$$\forall c \in [1,5], \qquad p(Choice\ c\ |Random\ Guessing) = \ 0.2.$$

The described method for object discrimination, performed trial-by-trial during the pretest (eight trials), was performed and fit to each child's behavioral responses for each of the four competing theories (Size, Material, Mass, and Random). That is, we evaluated the probability of the observed pattern of independent responses in the Pretest for each child, given the predictions of each model. The responses were treated as independent due to the absence of feedback on each trial. Thus, for each trial, we obtained the probability of the choice made by the child according to each of the competing theory model's predictions of the outcomes of each trial,

$$p(Size\ Theory\ |\ Child's\ Decision\ on\ Trial\ t)$$

- then, took the product of these eight trials to compute the overall probability for each belief.

By inferring the best fitting model of these data at the individual level, we obtained not only a quantifiable representation of each child's best fitting prior theory (the model that assigned the highest probability to the child's responses), but also a distribution of relative weights for each of the four competing theories that may be guiding their predictions. We return to these outcomes in the Results. Importantly, these individualized distributions are used to inform the predictions made by our model during the Learning Phase.

[2] The source code of each modeling component is freely available at https://osf.io/3k68n/?view_only=3b444e83f86a490e9fc4e18eb2556b51.

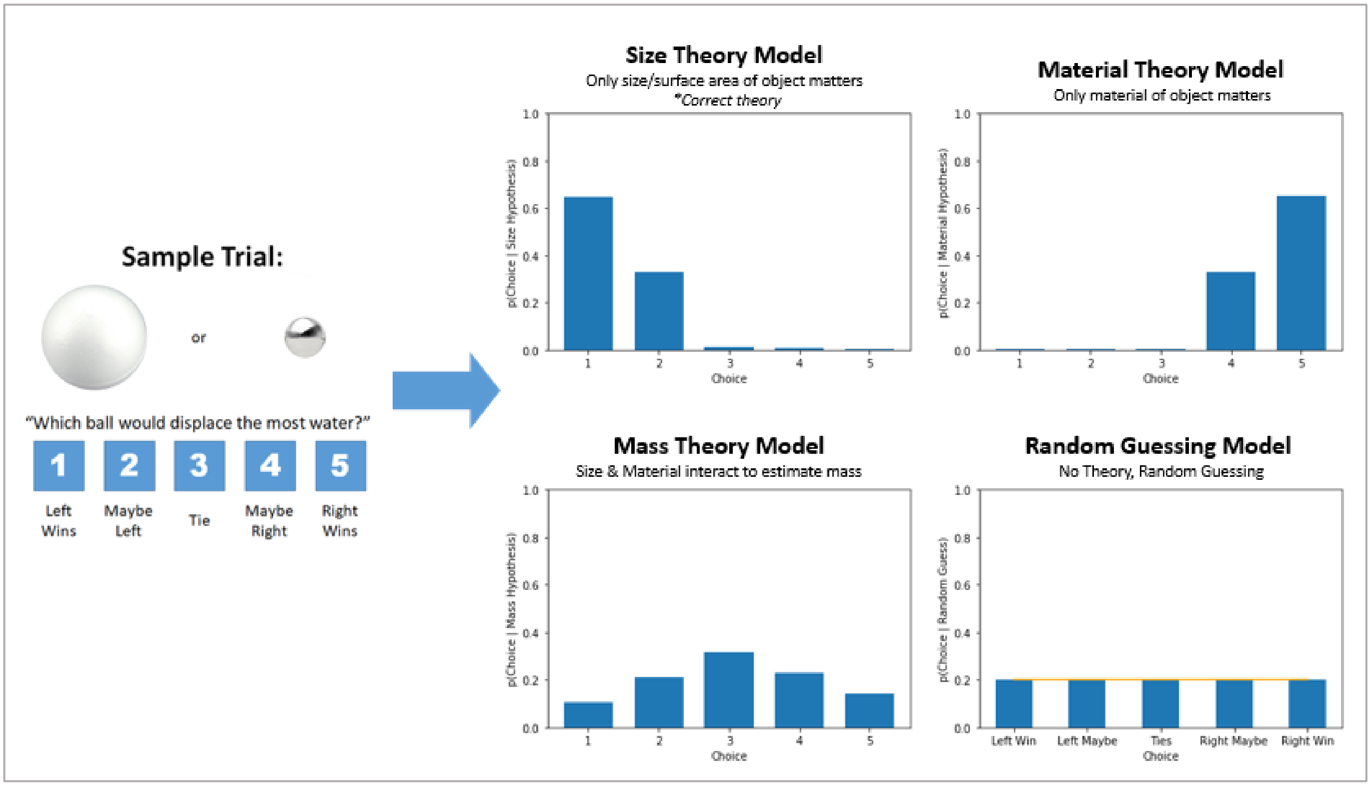


Fig. 2. Visualization of how each individual prior theory model generates predictions of event outcomes (Sample Trial) probabilistically. Each of the four unique, competing prior theories (Size, Material, Mass, Random Guessing) may generate different predictions based on variance in features of the objects being discriminated - dependent on the titular variable being considered by each theory.

B. Building Learning Models

One focus of this paper is on the role of prior beliefs in learning. Three competing models are compared, varying in how evidence and prior beliefs are incorporated. The first "Uniform-priors" model serves as a baseline comparison to behavior that does not incorporate learning or prior beliefs throughout trials and considers the impact of all competing beliefs equally. The second "Sticky-prior" model allows for an individual's initial priors to be considered, but is strictly dominated by the best-fitting prior theory from the first modeling stage and does not include trial-by-trial belief updating. The third model is the Theory-based "Optimal Bayesian model" that learns trial-by-trial and includes each individual's prior distribution of beliefs over the four described competing prior theories (Size, Material, Mass, Random) collected in the first modeling stage. These models are all compared on their performance in predicting children's actual responses during the learning phase (34 trials) of the original experiment. First, children's theory distributions were calculated trial-by-trial using traditional Bayesian Posterior Updating. We calculated an updated posterior probability on a given trial t+1 as

$$p(h_i|d_{t+1}) = \frac{p(d_{t+1}|h_i)p(h_i|d_t)}{\sum_{h_i\in H} p(h_i|d_{t+1})} ,$$

for each of the four competing theories (h_i) after some number of trials (t) with observations of data (d_t), given a prior probability of said theory given the prior data p(h_i|d_t), where at p(h_i|d_t = 0) = p(h_i). Thus, starting with the prior probability distributions calculated based on each child's individual responses during the Pretest Phase (t = 0), we generated a table of theory posteriors for each of the 34 trials of the learning phase.

As with the Pretest Phase, we constructed choice distributions (from "picking 1" to "picking 5") for each of the four competing belief models (Size, Material, Mass, Random Guessing) across the 34 trials included in the Learning Phase. Then, given the generated priors and posteriors at each trial, models of each child's decisions were generated. Here, model predictions of choice were generated as follows: first, at each trial (t), the probability ($p_t(c_h)$) of the model predicting some outcome (c_h) given one of the four competing beliefs (h) was $p_t(hi)p_t(c_h|\ h)$. This generates a 5x4 (belief h X choice c) table for each trial for each child, detailing the possible outcomes given each potential belief. Then, to consider the impact of multiple competing beliefs, the probability that the model would predict a given choice (c) on a trial (t) was taken as the summation of each of the four competing belief probabilities,

$$p(c_t) = \sum_{h\epsilon H}\ p(c_h),$$

such that for each outcome c, we now had a summed probability across that considers the weight of each of the competing theories. Then, the final model prediction of a child's choice for each trial generated by the model was taken as the predicted outcome (from 1 to 5) with the highest probability within this summed column.

While the model may eventually "learn" to converge on the correct theory (that the size of an object determines how much water it will displace), it may still fail to correctly weigh the predictions. Our model implements not just the notion of one dominating belief, but a weighted space or distribution of beliefs. Some beliefs may not be as "strong" as others at various points during the task. This is captured by the trial-by-trial revision of individual children's priors and posteriors, where children begin with an initial distribution across the beliefs considered in our model (Size, Material, Mass, and Random Guessing) based on their pretest behaviors, but change over time given observed evidence in the feedback trials.

1) Uniform-priors model

Starting with the simplest model framework, we first define a model that does not rely on any bias for prior theories or preconceptions. Specifically, given the four competing beliefs described earlier (Size, Material, Mass, and Random Guessing), the relative weights of each belief were set equally (to 0.25, each) before calculating the likelihoods of potential choices ($p_t(c_h)$) across all 34 trials.

2) Sticky-prior Model

The Sticky-prior model accounts for the calculated priors from the first modeling stage, but only considers the predictions related to the dominant prior theory following the pretest. For example, consider the hypothetical model of a child immediately following the pretest with the priors for the Size, Material, Mass, and Random models as 0.20, 0.40, 0.25, 0.15, respectively. This model might be considered a "Strongly Material-belief Holder". Given the described initial prior theory distribution, this model remains "stuck" and only responds by choosing the option weighted in proportion to this initial distribution – tending towards Material-belief consistent responses.

3) Optimal Bayesian Model

Finally, the third model accounts for both individual differences in initial prior beliefs and flexible learning as performed via Bayesian Posterior Updating. Starting from the initial priors as found in the first modeling stage of the Pretest, the model of each child's performance is then updated trial-by-trial using Bayes Rule, accounting for both the influence of children's specific prior beliefs as well as the dynamics of active trial-by-trial learning in respect to evidence.

In generating the predictions of the three models for the choices made by children at each trial, two methods were used. First, in generating discrete predictions for use in determining the "fit" of our model to children's actual behavioral data, the maximum summed probability of each choice was taken, $\max(\boldsymbol{\Sigma}_m\, p_t(c_m))$. Second, in investigating model uncertainty, the calculation of a weighted mean was performed, generating a continuous prediction that may lie between each of the discrete options. Thus, on a given trial (t) for each possible choice option (c), the model's expected choice was calculated as: $\boldsymbol{\Sigma}\, cp_t(c)$.

C. Role of prediction in model likelihood

Finally, we also investigated potential differences in learning dynamics by analyzing model performance (of the best fitting model) between the two experimental conditions of the original study. Here, we posit that learning may have unfolded differently for children who were more actively engaged with the task in the Prediction condition than in the Postdiction condition. In the Prediction condition, children were required to generate predictions that may have been drawn from their prior beliefs and helped them to both recognize conflicting evidence and explicitly consider other potential beliefs. In contrast, children in the Postdiction condition, passively observed evidence as it occurred and were thus potentially less engaged both with the task and their own prior beliefs. We hypothesize here that the performance and correlation of our model in relation to children's actual performance will be stronger for those who are actively engaging their prior beliefs in the Prediction condition, compared to children making post-hoc evaluations in the Postdiction condition.

IV. Comparing Models to Behavioral Results

When discussing results regarding each model's proficiency in predicting children's behavior, three separate measures are used to evaluate performance. First, we calculate log-likelihood scores, capturing the probability that each model would predict the choices made by children during the task. Second, we look at the correlation between children's actual responses and the responses each of the three models would predict the child would make at each trial. Third, we evaluate the accuracy of the three models from the "mini-microgenetic" experiment, scoring each model based on whether it correctly identifies the choice made trial-by-trial.

For each of these three measures, 34 trials are analyzed for each of the 94 modeled children, resulting in 3,196 trials analyzed. However, due to technical issues with materials or apparatus during the original experiment, 270 trials during the learning phase were not properly recorded, resulting in 2,926 trials available for analysis (see Figure 3 for each of the three model's generated predictions, compared to children's reported behavior from the original experiment).

A. Capturing Children's prior beliefs

Overall, the results of the first modeling stage resulted in 94 unique distributions, fit to the responses made by children during the pretest (see Supplemental material for individual fits per participant). To get a feel for the overall distribution of best fitting beliefs across children and conditions, we also looked at the max fitting belief model for each child. Overall, 21 children were best fit by the Size theory, 16 best fit by the Material theory, 32 were best fit to the Mass theory, and 25 best fit to the Random model, demonstrating significant variability of beliefs of the children at the start of the task. The distribution of best fitting prior theory (Size, Material, Mass, or Random) was similar at the start of the experiment (as would be expected given random assignment) between the experimental conditions of the original study, Prediction and Postdiction ($\chi2(3) = 2.13$, $p = 0.54$).

B. Quantifying transitions in belief states

Additional modeling and analyses were performed to further investigate individual differences among children's belief states during the training trials. Specifically, we looked at whether there were any differences (between conditions) or general trends (within the full dataset) regarding children's "belief state", trial-by-trial. First, to ensure that we were accounting for children's actual belief state in conjunction with our model's predictions, we calculate "choice-moderated priors". That is, for each child at each trial, we calculated $p(C_t|h_i)p(h_i)$, where C_t is the choice made by the child during the original experiment on trial t and h_i is each of the four

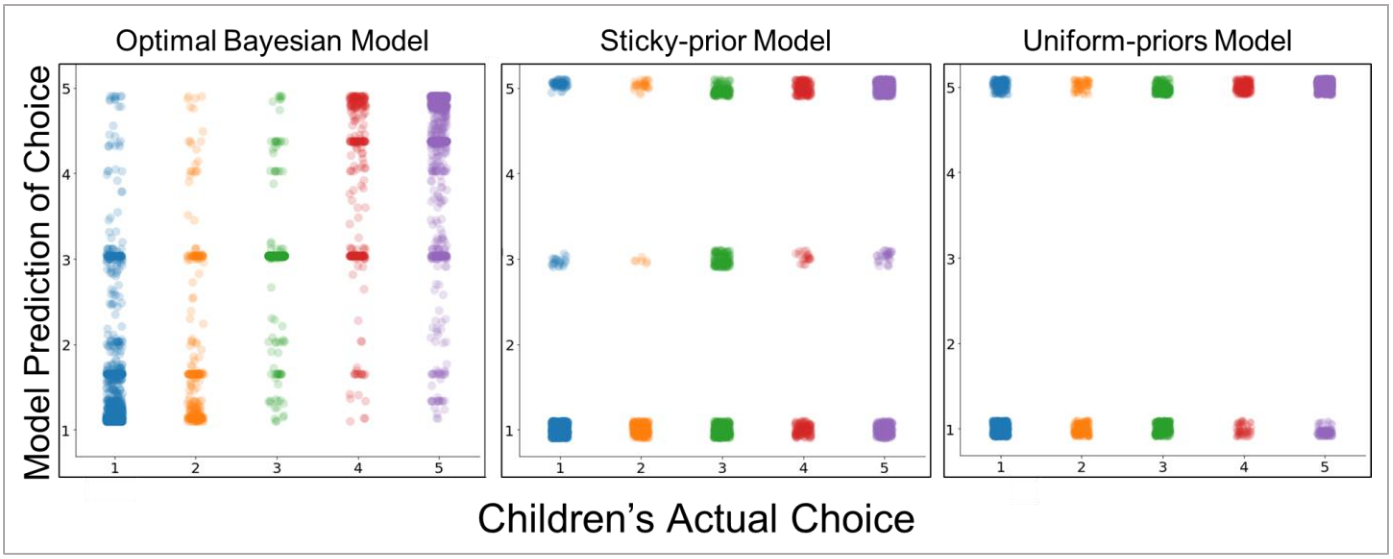


Fig. 3. Comparison of children's response behavior during the original experiment with each of the three competing model's prediction of children's choice behavior during the experiment, trial-by-trial. Children's actual choices from the Learning Phase of the experiment are plotted along the x-axis. Similarly, each of the competing model's prediction of children's behavior is plotted along the y-axis. Random jitter for all datapoints and color-coded options along the x-axis have been added to each figure for visualization purposes. By comparing the likelihoods, correlations, and accuracies of each model, we find that the Uniform-priors and the Sticky-prior models are outperformed by the Optimal Bayesian model when predicting children's behavior.

competing beliefs. Thus, we obtained a "choice-moderated" belief distribution for each child, trial-by-trial. Then, using these choice-moderated priors, we looked at the Bayes Posterior Odds to determine at which trial each child was considered to have converged onto the "correct" Size theory based on our model's predictions of their current distribution of beliefs on a given trial as affected by the child's actual choice behavior, trial-by-trial.

We calculated the ratio of the probability of the Size prior for each child based on their choices against the competing belief with the largest probability (Material, Mass, or Random Guessing), and determined the trial at which there was "substantial evidence" in support of the child's model holding a dominant Size belief (e.g., substantial evidence was set to when the mentioned ratio was greater than or equal to 3). We find no difference regarding what trial children seemed to have learned the Size theory between the Prediction (mean trial = 14.98) and Postdiction (mean trial = 13.37) conditions ($t(92) = 0.93$, $p = 0.35$). One possibility is the children who began the experiment with substantial evidence already in favor of the Size belief (e.g., they had the right model initially and would not learn further), influenced these results. So, we also ran this analysis excluding these children (excluded n = 19; with 8 in the Prediction condition and 11 in the Postdiction condition). However, there were no differences even after removing these "already-knowers" from the analysis (Prediction mean trial = 17.98) and Postdiction (mean trial = 17.57) condition differences ($t(73) = 0.37$, $p = 0.71$).

Finally, we also investigated whether learning the correct beliefs of water displacement during the Learning Phase was continuous. We looked at whether children in the experiment ever reverted to a belief distribution that was not dominated by the Size belief after the initial point of Size-belief convergence. For each child, starting from the trial at which their Bayes Factor was greater than 3 in support of the Size belief up until the final trial of the Learning Phase, we looked at how often children shifted away from this Size-dominant distribution, trial-by-trial. Here, we found that 34 of the 94 children did "flip back" to an incorrect belief at least once. However, only 8 of the 94 children flipped back and forth twice, and only 1 child did so three times. Further analysis comparing whether there was a difference in the frequency of how many children "flipped" between experimental conditions found no difference between the Prediction (17 of 48 flipped back) and Postdiction (17 of 46 flipped back) conditions ($\chi^2(3) = 0.02$, $p = 0.88$)

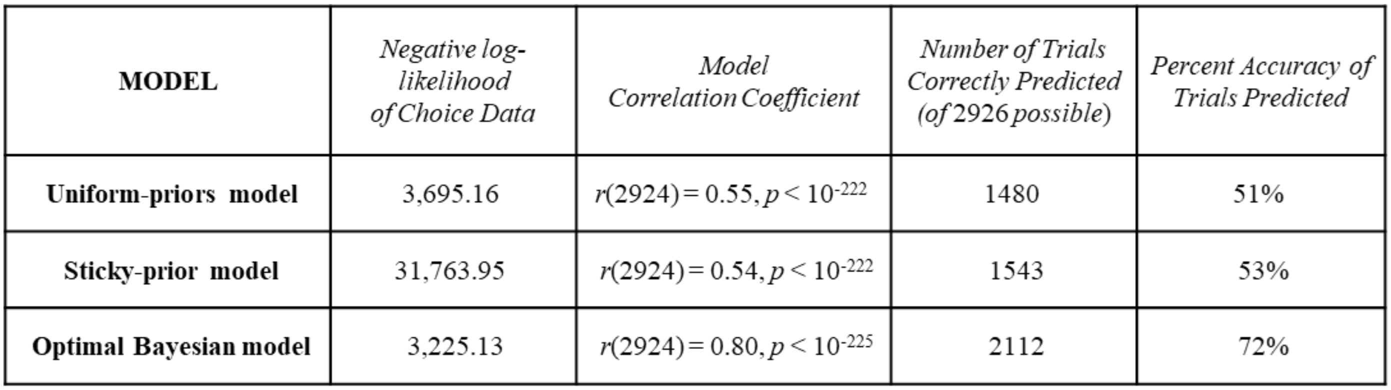
C. Model comparison of trial-by-trial learning

In addition to their mention within related subsections, specific statistical results for each of the three metrics used during model comparison (log-likelihoods, correlation coefficients, and accuracy scores) have additionally been compiled in Table I below for ease of comparison.

1) Model fits to children's responses

To analyze the performance of the Uniform-priors, Sticky-prior, and Optimal Bayesian models to the children's responses in the main training, we first calculated log-likelihood scores of the model fits to the children's responses. Thus, for each child (k), on each trial (t), the log of the probability of actual choice (c) made by the child (given the model being inspected) was calculated: $\log(p_{t,k}(c)p_{t,k}(c|\ h))$. Then, across all collected trials, the sum of all log-likelihoods was taken as the respective model's "score". We present the negative logs for ease of comparison, such that models with a lower score are considered "more likely" per the model's predictions, and therefore better in performance at predicting actual human responses. Here, for the three competing models, this negative log-likelihood score revealed that the Optimal Bayesian model (score = 3,225.13) fit the child data better than the Sticky-prior (score = 31,763.95) and Uniform-priors (score = 3,695.16) models. Thus, we find that the Optimal Bayesian model assigns the highest probability

TABLE I
STATISTICAL RESULTS OF MODEL COMPARISON TO BEHAVIORAL DATA



MODEL	*Negative log-likelihood of Choice Data*	*Model Correlation Coefficient*	*Number of Trials Correctly Predicted (of 2926 possible)*	*Percent Accuracy of Trials Predicted*
Uniform-priors model	3,695.16	$r(2924) = 0.55, p < 10^{-222}$	1480	51%
Sticky-prior model	31,763.95	$r(2924) = 0.54, p < 10^{-222}$	1543	53%
Optimal Bayesian model	3,225.13	$r(2924) = 0.80, p < 10^{-225}$	2112	72%

Table I. Comparing the three models (Uniform-priors, Sticky-prior, and Optimal Bayesian) to the behavioral results revealed better fits and significantly better performance in simulating children's behavior by the Optimal Bayesian model.

to the children's behavioral data, followed by the Uniform-priors model and the Sticky-prior model.

2) Correlation between model predictions and children's responses

The second comparison regarded the direct correlation between children's actual responses during the experiment, and each model's generated prediction of children's choices. All three models resulted in significant correlation between predictions and children's responses (Uniform-priors model: $r(2924) = 0.55$, $p < 10^{-222}$; Sticky-prior model: $r(2924) = 0.54$, $p < 10^{-222;}$ Optimal Bayesian model: $r(2924) = 0.80$, $p < 10^{-225}$). Performing pairwise comparisons between each model-pair (via Fisher's R-to-Z transformation), revealed that the correlation for the Optimal Bayesian model is significantly higher from its competitors (versus Uniform-priors model, $z = 18.36$, $p \sim 0$; versus Sticky-prior model, $z = 18.90$, $p \sim 0$), with no difference between the Sticky-prior and Uniform-priors models ($z = 0.54$, $p = 0.29$).

3) Evaluating accuracy of model predicted response

The third form of model comparison entailed scoring each model, trial-by-trial, on their ability to accurately predict the choices made by children. This allows us to further inspect each model's performance more closely in both a quantitative (via percent accuracy) and qualitative (via frequentist statistics) manner. Here, each model was scored trial-by-trial, receiving an accuracy score of "1" if generating an exact match to children's behavior, and a "0" otherwise.

First, for the Uniform-priors model, 1480 of the 2926 trials (51%) were correctly predicted. Next, for the Sticky-prior model, 1543 of the 2926 trials (53%) were correctly predicted. Finally, for the Optimal Bayesian model, 2112 of the 2926 trials (72%) were correctly predicted. Comparing the three models, we find that there is a significant difference in accuracy among the models (one-way ANOVA; $F(2, 2923) = 177.44$, $p < 10^{-75}$). Further pairwise comparisons between models find that the Optimal Bayesian model performed significantly better than the Uniform-priors ($t(5850) = 17.40$, $p < 10^{-65}$) and Sticky-prior models ($t(5850) = 15.68$, $p < 10^{-53}$), but that these latter competitors were not significantly different from one another ($t(5850) = 1.65$, $p = 0.09$).

4) Role of Prediction in optimal learning models

For the fourth modeling stage, an additional analysis was performed to investigate whether differences in learning activities (Prediction versus Postdiction) were revealed by the best performing Optimal Bayesian model. Here, we hypothesize that if Prediction engages model-based learning, the Optimal Bayesian model will perform better in approximating the behavior seen in the Prediction condition compared to that of the Postdiction condition. This would be shown via a better "fit" to the behavioral data of the children in the Prediction condition.

Comparing the fit of our model to each of the conditions separately, we performed correlations between the model's predictions of the children's choices and the actual responses made by children, per condition. As with the aggregate data analysis, we found that there were significant correlations between our model and the child data in both the Prediction condition ($r(1501) = 0.86$, $p \sim 0$) and the Postdiction condition ($r(1421) = 0.75$, $p < 10^{-260}$). Comparing these correlations (via Fisher's R-to-Z transformation), we find that children's responses in the Prediction condition correlated significantly better with the optimal model children's predictions in the Postdiction condition ($z = 8.65$, $p \sim 0$).

We also compared accuracy scores within conditions. For the Prediction condition, 1167 of the 1503 trials (78%) were correctly predicted. For the Postdiction, 945 of the 1423 trials (66%) were correctly predicted. Performing a pairwise comparison, we find that the Optimal Bayesian model performed significantly better at predicting the Prediction condition, compared to the Postdiction condition ($z = 6.78$, $p < 0.001$). We additionally looked at the sources of error in both models. In general, errors of fit occurred when the model predicted the "most certain" responses (1, 5) but children responded with some uncertainty (2, 4). Allowing for an uncertain, but still correct directional response improves model fits to 95.0% in the Prediction Condition and 92.5% in the Postdiction condition, with a significant difference between conditions remaining, $\chi 2(2892) = 6.82$, p = .009. Thus, per our model analysis, we find support for the finding that children in the Prediction condition may have been learning more "optimally" (in a Bayesian sense), compared to the Postdiction condition, though both conditions are well fit to the Optimal Bayesian model.

V. Discussion

The goal of this computational endeavor was to further understand theory-based belief revisions in contexts of prediction-based, perceptually noisy learning scenarios, such as during learning about what object properties determine the degree of water displacement. We approached this from a computational standpoint - combining Bayesian inference models with recent findings on the roles of prior theories, rapid incremental theory change, and active predictions. In short, we found support for the notion that prediction may bolster “optimal” learning by encouraging engagement with one’s prior theories (and potential misconceptions) when inferring potential outcomes and actively accumulating evidence via perceptual cues. Specifically, we found support for our hypotheses in investigating our three main questions.

First, we found support that we can best understand the models that young learners start with and leverage in learning by building computational models that consider priors. In line with past research on water displacement [46,47], participants in both conditions were classified with “best-fitting” beliefs rather evenly across the four possible models. We also found support for the importance of leveraging malleable priors that evolve throughout learning in the face of evidence. The Optimal Bayesian model, which allowed for the consideration of multiple competing beliefs before determining their best prediction, outperformed the Sticky-prior and Uniform-priors model by having a higher probabilistic fit, by having a stronger correlation, and by having a higher accuracy when predicting children’s behavioral data. This demonstrates the importance of taking individual differences in belief starting states into account.

Second, our model analysis indicates that the learning process during belief revision may be best explained as mainly unfolding continuously over time, with influence from new observations trial-by-trial. Our analysis also allowed us to consider the sometimes “non-linear” nature of learning, consistent with the “Sampling Hypothesis” which suggests that children may be sampling responses to approximate Bayesian models, but leading to occasional “regressions” [14,53,54]. Overall, children’s learning only had momentary lapses of this “belief regression”, with only 34 of the 94 children being modeled as briefly “flipping” back to an incorrect belief.

Third, our models revealed that prediction may be one of the circumstances that require learners to engage their models of the world to support learning in a top-down fashion, compared to passively receiving information and evaluating past events. Applying the best-fitting Optimal Bayesian model to children’s behavior in each condition revealed a closer alignment between the model’s and children’s responses in the Prediction condition than responses in the Postdiction condition. The act of predicting may have encouraged children to leverage their prior theories more deeply, compared to children in the Postdiction condition that simply viewed events occurring passively. This suggests that children who are encouraged to make a prediction and, thereby, to engage their prior models, may be engaging in learning processes that are more “optimal” in a Bayesian sense.

By exploring the questions posed here, we can better understand the mechanisms that support prediction, perception, and learning in humans and robots more generally. One open topic for investigation is the role of perceptual processes in prediction, theory change, and science learning. Here, one component of perceptual processing comes into play via the noise incorporated into the various theory models. Learners were modeled as having uncertainty about edges and processes - coming into play within our learning models utilizing Weber’s Law [50-52] when performing visual discrimination during the task modeled. This was an important component of noise in learning and feedback, and thus a key feature of the probabilistic modeling process described in this paper. This raises additional questions regarding perceptual noise over the course of development. How might perceptual processes be incorporated into models capturing learning earlier in development, and over longer time periods? Might children learn to weigh these processes more or less as their processes sharpen with age? It also raises questions regarding noisier circumstances where discrimination of visualized objects may include a larger number of items to evaluate, or the learning of other scientific concepts. How might these processes unfold as scene complexity and ambiguity increase?

A. *Understanding Divergences from the Model*

Our results revealed that the Optimal Bayesian model would typically choose the more certain option across the trials (e.g., mainly, if not only, choosing 1, 3, or 5). In contrast, children occasionally provided responses that reflected greater uncertainty, choosing options 2 and 4 on occasion. There are multiple, non-mutually exclusive possibilities for this. One possible reason for this difference is that the model provides a computational “optimal” prediction following a “max-rule” for selecting responses, but children may instead by selecting responses following the “sampling hypothesis” (e.g., see [53]), leading to responses that are occasionally less likely. A second possibility is that our models must allow for greater perceptual noise (especially in mass judgments), which may lead to less certainty across the models. For example, children might explain away unexpected results (e.g., a large wood ball dispersing more water than a small metal ball) by updating beliefs about mass types generally (e.g., that metal may weigh less than originally assumed, or wood may weigh more) allowing children to preserve uncertainty in mass and material beliefs following these “incongruent” trials. Finally, children may have had uncertainty for other reasons, due to possible gaps in attention, unfamiliarity with the set-up, or memory limitations that lead to a preference to select “less confident” options on occasion. Future work could explore these potential sources of uncertainty and how they might best be captured by our framework.

B. *Future Work*

Prediction seems to be a remarkable tool by which learners engage models to support learning, but there exist many open questions regarding what mechanisms guide prediction and theory revision in tandem. Several different possibilities exist, requiring both further empirical investigation and computational implementations. Both endeavors are necessary as we approach the intertwined goals of understanding human learning in perceptually noisy environments and building human-like robots that interact in them.

There are several potential influences on learning that warrant future investigation. One involves the role of the emotional-physiological response of surprise. A second is implementing cognitive limitations, such as executive function measures, as prerequisites or precursors for stronger instances of prediction and theory revision. A third regards deeper insights into the role of engagement, specifically self-agency and self-directed choice. There may also be interest in further investigating the role of specific contexts, including how alterations to the environment or items within it (e.g., more complex objects with unique features or functions) may affect the difficulty of learning concepts. Finally, we discuss contemporary interests in stubborn misconceptions (e.g., that the Earth is flat), where folk psychology may be negatively affected by social pressures.

1) Potential Role of surprise

Recent work points to a beneficial role of asking learners to generate a prediction before presenting them new information that conflicts with their intuitions (for a review, see [28]). If the outcome following a prediction is different than what was expected, awareness of this conflict may increase the subjective value of the outcome's informativeness, and increase the perceived expectancy-violation - resulting in enhanced surprise [45]. Indeed, several studies have reported enhanced surprise as a result of predicting [35,44,55]. Therefore, future work should consider two potential avenues.

First, surprise responses, such as the physiological response of pupil dilation when surprised [55], should be considered as a parameter in Bayesian models. Based on the cited literature, we hypothesize that careful inclusion of surprise as an additional parameter in Bayesian models may result in a stronger correlation between model predictions and human behavior if surprise directly mediates learning. Second, it becomes important to also quantify and estimate a "model-based surprise" in a predictable way - calculated based on the model's assigned probabilities of outcomes potentially occurring in different scenarios. Here, surprise may be considered as a metric of how "informative" incoming information may be, given an individual's current understanding (e.g., their prior beliefs) [56-58]. Comparing such "model surprise" to measures of surprise collected during experiments will yield notable implications – particularly in interpreting potential correlations as representing surprise as computationally "rational", where the valuation of incoming information is bolstered and may aid in belief revision processes. Overall, such computational endeavors would help us understand the circumstances that influence surprise-based learning at the individual- and trial-level.

2) Potential Role of executive functions

Modeling the relationship among prediction, theory change, and executive function skills (such as inhibition or cognitive flexibility [59,60] may provide further insight into other relevant mechanisms that support learning. This is important to do because executive function skills are considered to be especially relevant to learning in academic contexts [61-65]. In particular, recent studies have found significant relationships between executive function and theory revision in multiple domains [55,66,67]. For example, a recent experiment finds that in some cases, only children that reach a certain threshold of executive function capabilities are able to efficiently utilize conflicting information to update their beliefs [46 - Brod et al., 2020]. Here, it is hypothesized that perhaps as children's executive function capacities develop, then children may be more able to handle more complex theories when needed for learning [68,69].

Thus, executive function measures can be incorporated into computational models of theory revision as additional parameters that mediate learning. Such modeling would allow us to investigate a number of questions: would executive function affect model performance straightforwardly, where higher executive function scores predict better performance in prediction and theory change? Or, perhaps, does there exist a threshold where a minimum executive function score is required for meaningful theory change?

Such questions regarding belief revision and executive function can be investigated in multiple ways. One such way is investigating whether children with stronger executive function capacities are more efficient in their learning. Here, this may be found if children that perform better during executive function tasks are better simulated by ideal Bayesian learning models (per metrics like those described in this manuscript). Another computational step may entail the design of a Bayesian model that accounts for executive function skills – such as the ability to inhibit incorrect prior beliefs, or, flexibly switch focus towards updated, "more correct" theories. Such future investigations of these described computational approaches may shed insight into how executive function capacities may be necessary (at some sufficient level) for belief revision.

3) Potential Role of agency

In a follow-up experiment, we are working on evaluating the relevance of agency and choice in mechanisms of prediction and theory change. Recognizing one's agency in their actions and perceiving control over actions taken are important factors in driving our motivation and behavior [70,71], is found to reduce speed-accuracy tradeoffs [72], and promote persistence and adaptation [73-75].Thus, like the comparisons made in this manuscript (and in the dataset's original experiment) between the active engagement of making a prediction versus the more passive act of post-hoc evaluating, we can compare scenarios where learners are either actively making predictions versus passively viewing another person make predictions, themselves.

Comparison of these two contexts varying on a learner's level of "agency" may highlight the importance of "activating" and leveraging one's own beliefs to facilitate learning. Specifically, as found here and in the original study, if engagement with one's priors bolsters belief revision, then learning in the new "passive viewer" scenario may be less "optimal" per Bayesian principles. Further, applying computational models, like those described in this manuscript, to this problem will help us to more precisely characterize the role of agency as an additional mediator of learning.

4) Investigations of Distinct Object Features

Several factors may influence the revision of individual priors and posteriors, including contextual factors that additionally alter the environment or objects being interacted with. For example, the idiosyncrasies of different materials – such as a sponge and its absorption – might make learning concepts such as water displacement more difficult by including additional relevant features (e.g., the material's absorption rate). By including additional factors based on more-specific object features within our models, we may be able to further understand how scientific concepts are learned more broadly. For example, further experiments may carefully manipulate the presence or strength of various peculiar features, such as the described absorption rate. Here, the added complexity of compared objects and their features may affect the learning process – perhaps reducing the "speed" or "efficiency" at which the correct size-theory of water displacement is learned. Furthermore, recent studies [e.g., 76] and review [77] highlights the impact of irrelevant, but salient variables during concept learning, suggesting that learning of the correct concepts may be hindered by the interference of the increased salience of salient, but incorrect features. From this discussed example, inclusion of the newly included, more complex feature of absorbency, (that is still irrelevant in regard to outcomes of water displacement) may find that learning about water displacement may become more difficult –perhaps requiring stronger inhibition to avoid this additional influence during learning. Importantly, while this does provide a clear target for extensions in future investigations, past research has not yet investigated children's theories of absorption empirically, thus it is outside the scope of the current work.

5) Investigations of Social Pressures & Misinformation

Finally, an additional contextual effect in evidence-based learning that can be captured in future models is whether the evidence being evaluated is credible. Recent work investigating pseudoscientific or conspiratorial beliefs suggests that the prevalence of incorrect beliefs about the world may be related to the frequency that a learner encounters evidence of the false information (e.g., the "illusory truth effect" [78]) and may be bolstered by social cues that elicit inferences that frequently shared information is truthful [79]. These social pressures – where a shared incorrect theory may be influential or foundational within a social group (e.g., beliefs about the Earth being flat) – may mislead a learner into believing that this frequently noted evidence is true.

Following recent work that finds that when appraising information provided by another social partner, such sampled evidence may be considered "representative" – highlighting features that may be more important to be focused on (sometimes even if such features are technically irrelevant), and subsequently affecting future decisions (the "intentional selection assumption" [80,81]). Various experiments may look into such social pressures as affecting belief revision. For example, one experiment may investigate scenarios where a social partner provides learners with evidence that appears intentionally selected, but potentially incorrect. Then, behavior during this social-scenario may be compared against contexts where information is acquired more "plainly" – such as in the modeled experiment. Investigation of variance in framing scenarios like this may highlight what enables processes that influence a learner's acceptance of misinformation – perhaps still following principles of "ideal" Bayesian learning, despite being towards the goal of an incorrect theory. Thus, investigation of these regressed or naive theories may potentially highlight specific social or motivational aspects of evidence-based learning.

VI. Conclusion

We have provided a taste of the rich history of conceptual change, proposing just one specific computational modeling approach to address gaps between empirical findings and current computational implementations. We first formalized the prior beliefs that children may have to construct quantifiable representations of children's prior theories at the individual level, as informed by their past behavior. Second, we described the intuitive computations in which they are revised in light of new observations to approximate children's learning using an "optimal" Bayesian model - revealing dynamics of theory revision trial-by-trial. Third, we investigated the contexts under which this theory revision follows more or less "optimal" performance per Bayesian inference. Altogether, this work benefits the goal of implementing similar processes in robots and machines.

The work we present here aligns well with past work on Bayesian models of human learning (e.g., [18,37-39,82]) while also extending on these past implementations. Specifically, the present models contribute to the literature on learning models in three key ways. First, and most clearly, by being the first model of the target task - children's theories of water displacement. Second, we investigated whether differences occur due to variation in response modalities (predicting vs post-hoc evaluations) — even when information is kept consistent. Third, we performed a 'finer-grain' analysis by virtue of simulating the microgenetic study and assessing our models' performance trial-by-trial.

These are only the initial steps in this endeavor of building human-like machines that account for informed priors, model-based learning, and theory-based predictions. Importantly, we must keep in mind that even human learners sometimes forego these processes, and may still engage in model-free, associative learning processes when appropriate. Nonetheless, we hope to have convinced readers that prediction is a powerful cognitive tool that may promote learning by engaging one's prior theories - and thus a worthwhile avenue for future research in understanding the mechanisms of theory revision during science learning. By exploring the questions posed in terms of computational models, we can better understand the underlying structures and processes that support prediction, perception, and learning in humans and robots more generally. We predict that by expounding our theories of belief revision here, we will be better prepared to revise them in future work.

References

[1] J.K. Kruschke, "Bayesian approaches to associative learning: From passive to active learning," *Learning & behavior*, vol. 36, no. 3, pp. 210-226, 2008.

[2] S. Carey, "Are children fundamentally different kinds of thinkers and learners than adults?" *Thinking and learning skills*, vol. 2, pp. 485-517, 1985.

[3] A. Gopnik and A. N. Meltzoff, *Words, thoughts, and theories*, MIT Press, 1998.

[4] H.M. Wellman and S. A. Gelman, "Cognitive development: Foundational theories of core domains," *Annual review of psychology*, vol. 43, no. 1, pp. 337-375, 1992.

[5] E. Bonawitz, A. Fischer, and L. Schulz, L. "Teaching 3.5-year-olds to revise their beliefs given ambiguous evidence," *Journal of Cognition and Development*, vol. 13, no. 2, pp. 266-280, 2012.

[6] E.S. Spelke *et al.*, "Origins of knowledge," *Psychological review*, vol. 9, no. 4, pp. 605, 1992.

[7] R. Baillargeon, “Infants' physical world,” *Current directions in psychological science*, vol. 13, no. 3, pp. 89-94 ,2004.

[8] A.M. Leslie, and S. Keeble, “Do six-month-old infants perceive causality?,” *Cognition*, vol. 25, no. 3, pp.265-288, 1987.

[9] P. Muentner, and E. Bonawitz, “The development of causal reasoning,” in M. Waldman (Ed.) *Oxford Handbook of Causal Reasoning*. Oxford, United Kingdom: Elsevier Limited, 2017.

[10] M. Bullock, R. Gelman, and R. Baillargeon, “The development of causal reasoning,” in W. J. Friedman (Ed.), *The developmental psychology of time,* New York: Academic Press, 1982.

[11] A. Karmiloff-Smith, and B. Inhelder, “If you want to get ahead, get a theory,” *Cognition,* vol. 3, no. 3, pp. 195-212, 1974.

[12] R.S. Siegler, “Three aspects of cognitive development,” *Cognitive psychology,* vol. 8, no. 4, pp. 481-520, 1976.

[13] I. Bascandziev *et al.,* “The role of domain-general cognitive resources in children’s construction of a vitalist theory of biology,” *Cognitive Psychology,* vol. 104, pp.1-28, 2018.

[14] E. Bonawitz *et al.*, “Sticking to the evidence? A behavioral and computational case study of micro-theory change in the domain of magnetism,”, *Cognitive science,* vol. 43, no.8, 2019.

[15] P.V. Potvin *et al.*, “Persistence of the intuitive conception that heavier objects sink more: A reaction time study with different levels of interference,” *International Journal of Science and Mathematics Education,* vol. 13, pp. 21–43, 2015.

[16] L.M.B. Foisy *et al.*, “Inhibitory control and the understanding of buoyancy from childhood to adulthood,” *Journal of Experimental Child Psychology,* vol. 208, no.105155, 2021.

[17] R. Babai and A. Amsterdamer, “The persistence of solid and liquid naive conceptions: A reaction time study,” *Journal of Science Education and Technology,* vol. 17, pp. 553–559, 2018

[18] J.B. Tenenbaum, T.L. Griffiths, and C. Kemp, “Theory-based Bayesian models of inductive learning and reasoning,” *Trends in cognitive sciences,* vol. 10, no. 7, pp. 309-318, 2006.

[19] A. Gopnik and E. Bonawitz, “Bayesian models of child development,” *Wiley interdisciplinary reviews: cognitive science,* vol. 6, no. 2, pp. 75-86, 2015.

[20] S. Vosniadou and C. Ioannides, “From conceptual development to science education: A psychological point of view,” *International Journal of Science Education,* vol. 20, pp. 1213–1230, 1998

[21] A. Shtulman and J. Valcarcel, “Scientific knowledge suppresses but does not supplant earlier intuitions,” *Cognition,* vol. 124, pp. 209–215, 2012.

[22] A. Shtulman and C.H. Legare, “Competing explanations of competing explanations: Accounting for conflict between scientific and folk explanations,” *Topics in cognitive science,* vol. 12, no. 4, pp. 1337-1362, 2020.

[23] Y. Zhu *et al.*, “Event-Related Potential Evidence for Persistence of an Intuitive Misconception About Electricity,” *Mind, Brain, and Education,* vol. 13, no. 2, pp. 80-91, 2019.

[24] G. L. Murphy and D.L. Medin, “The role of theories in conceptual coherence,” *Psychological review,* vol. 92, no. 3, pp. 289, 1985.

[25] V.K. Mansinghka *et al.*, “Structured priors for structure learning,” in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence,* 2006.

[26] R.S. Siegler and Z. Chen, “Developmental differences in rule learning: A microgenetic analysis,” *Cognitive Psychology,* vol. 36, no. 3, pp. 273-310, 1998.

[27] E. Bonawitz, “Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference,” *Cognitive psychology,* vol. 74, pp. 35-65, 2014.

[28] G. Brod, “Predicting as a learning strategy,” *Psychonomic Bulletin & Review. Advance Online Publication*, 2021.

[29] M. Limón, “On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal,” *Learning and instruction,* vol. 11, no. 4-5, pp. 357-380, 2001.

[30] E. Bonawitz *et al.*, “Children balance theories and evidence in exploration, explanation, and learning,” *Cognitive Psychology,* vol. 64, pp. 215–234, 2012.

[31] K. Kimura and A. Gopnik, “Rational higher-order belief revision in young children,” *Child Development*, vol. 90, no. 1, pp. 91-97, 2019.

[32] G. Brod, M. Hasselhorn, and S. A. Bunge, “When generating a prediction boosts learning: The element of surprise,” *Learning and Instruction,* vol. 55, pp. 22-31, 2018.

[33] T.M. Gureckis and D. B. Markant, “Self-directed learning: A cognitive and computational perspective,” *Perspectives on Psychological Science*, vol. 7, no. 5, pp. 464-481, 2012.

[34] B.M. Lake *et al.*, “Building machines that learn and think like people,” *Behavioral and brain sciences,* vol. 40, 2017.

[35] M. Theobald and G. Brod, “Tackling Scientific Misconceptions: The Element of Surprise,” *Child Development*, 2021.

[36] C. Kemp *et al.*, “A probabilistic model of theory formation,” *Cognition,* vol. 114, no. 2, pp. 165-196, 2010.

[37] J.B. Tenenbaum *et al.*, “How to grow a mind: Statistics, structure, and abstraction,” *Science*, vol. 331, no. 6022, pp. 1279-1285, 2011.

[38] T.D. Ullman *et al.*, “Theory learning as stochastic search in the language of thought,” *Cognitive Development,* vol. 27, no. 4, pp. 455-480, 2012.

[39] N.D. Goodman *et al.*, “Learning a theory of causality,” *Psychological review*, vol. 118, no.1, pp.110, 2011.

[40] W.F. Brewer *et al.*, “Explanation in scientists and children,” *Minds and Machines,* vol. 8, no. 1, pp. 119-136, 1998.

[41] J. Piaget, “The child's conception of causality,” *Journal of Philosophical Studies,* vol. 5, no. 20, 1930

[42] L. Schauble, “Belief revision in children: The role of prior knowledge and strategies for generating evidence,” *Journal of Experimental Child Psychology,* vol. 49, pp. 31–57, 1990.

[43] G.J. Posner *et al.*, “Accommodation of a scientific conception: Toward a theory of conceptual change,” *Science education,* vol. 66, no. 2, pp. 211-227, 1982.

[44] J. Breitwieser and G. Brod, “Cognitive prerequisites for generative learning: Why some learning strategies are more effective than others,” *Child development,* vol. 92, no. 1, pp. 258-272, 2021.

[45] G. Brod and J. Breitwieser, “Lighting the wick in the candle of learning: generating a prediction stimulates curiosity,” *NPJ science of learning,* vol. 4, no. 1, pp. 1-7, 2019.

[46] J. Piaget and & B. Inhelder, “Le développement des quantités chez l'enfant,” 1941.

[47] N.C. Burbules and M.C. Linn, “Response to contradiction: Scientific reasoning during adolescence,” *Journal of Educational Psychology,* vol. 80, no. 1, pp. 67, 1988.

[48] C. Dawson and J. Rowell, “Displacement of water: Weight or volume? An examination of two conflict based teaching strategies,” *Research in Science Education,* vol. 14, no. 1, pp. 69-77, 1984.

[49] M.C. Linn and B.S. Eylon, “Knowledge integration and displaced volume, *Journal of Science Education and Technology,* vol. 9, no. 4, pp. 287-310, 2000.

[50] S. Dehaene, “Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation,” *Sensorimotor foundations of higher cognition,* vol. 22, pp. 527-574, 2007.

[51] S. Droit-Volet *et al.*, “Time, number and length: Similarities and differences in discrimination in adults and children,” *Quarterly journal of experimental psychology,* vol. 61, no. 12, pp. 1827-1846, 2008.

[52] S. Droit-Volet and P.S. Zélanti, “Development of time sensitivity and information processing speed,” *PloS one,* vol. 8, no. 8, 2013.

[53] E. Bonawitz *et al.*, “Probabilistic models, learning algorithms, and response variability: sampling in cognitive development,” *Trends in cognitive sciences*, vol. 18, no. 10, pp.497-500, 2014.

[54] S. Denison *et al.*, “Rational variability in children’s causal inferences: The sampling hypothesis,” *Cognition*, vol. 126, no. 2, pp. 285-300, 2013.

[55] G. Brod *et al.*, “Being proven wrong elicits learning in children–but only in those with higher executive function skills,” *Developmental science*, vol. 23, no. 3, 2020.

[56] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal,* vol. 27, no. 3, pp. 379-423, 1948.

[57] Z. Rafi and S. Greenland, “Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise,” *BMC Medical Research Methodology,* vol. 20, no.1, pp. 1-13, 2020.

[58] S.R. Cole *et al.*, “Surprise!,” *American Journal of Epidemiology,* vol. 190, no.2, pp. 191-193, 2021.

[59] A. Miyake and N.P. Friedman, “The nature and organization of individual differences in executive functions: Four general conclusions,” *Current Directions in Psychological Science,* vol. 21, no. 1, pp. 8-14, 2012.

[60] A. Diamond, “Executive functions,” *Annual Review of Psychology,* vol. 64, no. 1, pp. 135–168, 2013.

[61] U. Müller *et al.*, “Executive function, school readiness, and school achievement,” in *Applied cognitive research in K–3 classrooms.* S. K. Thurman and C. A. Fiorello, Eds., Routledge/Taylor & Francis Group, 2008, pp. 41–83.

[62] J.R. Best *et al.*, “Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample,” *Learning And Individual Differences,* vol. 21, no.4, pp. 327-336, 2011.

[63] C.M. Roebers, “Executive function and metacognition: Towards a unifying framework of cognitive self-regulation,” *Developmental Review,* vol. 45, pp. 31–51, 2017.

[64] M.T. Willoughby *et al.*, “The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement,” *Psychological Assessment,* vol. 24, pp. 226–239, 2012.

[65] J.A. Welsh *et al.*, “The development of cognitive skills and gains in academic school readiness for children from low-income families,” *Journal Of Educational Psychology,* vol. 102, no. 1, pp. 43, 2010.

[66] D. Zaitchik *et al.*, “The effect of executive function on biological reasoning in young children: An individual differences study,” *Child development,* vol. 85, no. 1, pp. 160-175, 2014.

[67] I. Bascandziev *et al.*, “A role for executive functions in explanatory understanding of the physical world,” *Cognitive Development,* vol. 39, pp. 71-85, 2016.

[68] S. Marcovitch and P.D. Zelazo, “A hierarchical competing systems model of the emergence and early development of executive function,” *Developmental Science,* vol. 12, no.1, pp. 1–25, 2009.

[69] P.D. Zelazo, “The development of conscious control in childhood,” *Trends in Cognitive Sciences,* vol. 8, no. 1, pp. 12–17, 2004.

[70] E.L. Deci and R.M. Ryan, "The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior,” *Psychological inquiry,* vol. 11, no. 4, pp. 227-268, 2000.

[71] E.T. Higgins, “What reigns supreme: Value, control, or truth?,” *Motivation Science,* vol. 5, no. 3, pp. 185, 2019.

[72] B. Eitam *et al.*, “Motivation from control,” *Experimental brain research,* vol. 229, no. 3, pp. 475-484, 2013.

[73] J.P. Bhanji and M.R. Delgado, “Perceived control influences neural responses to setbacks and promotes persistence,” *Neuron,* vol. 83, no. 6, pp. 1369-1375, 2014.

[74] L.A. Leotti and M.R. Delgado, “The inherent reward of choice,” *Psychological science,* vol. 22, no. 10, pp. 1310-1318, 2011.

[75] V.P. Murty *et al.*, “The simple act of choosing influences declarative memory.” *Journal of Neuroscience,* vol. 35, no. 16, pp. 6255-6264, 2015.

[76] H. Galili *et al.*, “Intuitive interference in geometry: An eye-tracking study,” *Mind, Brain, and Education,* vol. 14, no. 2, pp. 155-166, 2020.

[77] G. Allaire-Duquette *et al.*, “Interventions aimed at overcoming intuitive interference: insights from brain-imaging and behavioral studies,” *Cognitive Processing,* vol. 20, no. 1, pp. 1-9, 2019.

[78] L. Hasher, D. Goldstein, and T. Toppino. "Frequency and the conference of referential validity," *Journal of verbal learning and verbal behavior,* vol. 16, no. 1, pp. 107-112, 1977.

[79] E. Orticio, L. Marti, and C Kidd. "Beliefs are most swayed by social prevalence under uncertainty," *In Proceedings of the Annual Meeting of the Cognitive Science Society,* vol. 43, no. 43. 2021.

[80] Colantonio, J. *et al.*, “The intentional selection assumption,” *Frontiers in Psychology,* vol. 12, 2021.

[81] P. Shafto and E. Bonawitz, “Choice from among intentionally selected options,” in *Psychology of learning and motivation,* vol. 63, B.H. Ross, Ed. Academic Press, 2015, pp. 115-139.

[82] T.L., Griffiths, and J.B. Tenenbaum. “Theory-based causal induction,” *Psychological review,* vol. 116, no. 4, pp. 661, 2009.