



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Children's active physical learning is as effective and goal-targeted as adults'

Citation for published version:

Bramley, NR & Ruggeri, A 2022, 'Children's active physical learning is as effective and goal-targeted as adults', *Developmental Psychology*. <https://doi.org/10.1037/dev0001435>

Digital Object Identifier (DOI):

[10.1037/dev0001435](https://doi.org/10.1037/dev0001435)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Developmental Psychology

Publisher Rights Statement:

This paper has been accepted to *Developmental Psychology*.

©American Psychological Association, 2022. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. The final article is available, upon publication, at: <https://doi.org/10.1037/dev0001435>

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Children's active physical learning is as effective and goal-targeted as adults'

Neil R. Bramley (neil.bramley@ed.ac.uk)

Department of Psychology, University of Edinburgh, Scotland

Azzurra Ruggeri

MPRG iSearch, Max Planck Institute for Human Development, Berlin, Germany & School
of Education, Technical University Munich

This paper has been accepted to *Developmental Psychology*.
©American Psychological Association, 2022. This paper is not the copy of
record and may not exactly replicate the authoritative document published in
the APA journal. The final article is available, upon publication, at: [https://
psycnet.apa.org/PsycARTICLES/journal/dev/](https://psycnet.apa.org/PsycARTICLES/journal/dev/)

Author Note

Data, code and movies of all individual trials can be found at
<https://osf.io/v9fk2/> (Bramley & Ruggeri, 2022). The study was not preregistered.

Abstract

We explore how children and adults actively experiment within the physical world to achieve different epistemic goals. In our experiment, 101 4–10-year-old children and 24 adults either passively observed or used a touchscreen interface to actively interact with objects in a dynamic physical microworld with the goal of inferring one of two latent physical properties: relative object masses or local forces of attraction and repulsion. We find an age improvement in judgments as well as an advantage for active over passive learning. With the help of Bayesian statistics and a computational modeling framework for the quantitative analysis of participants' actions, we show that children's and adults' actions are equally successful in targeting their goal-relevant uncertainty, but that adults and older children are better able to use this information to respond correctly. We further unpack children's and adults' experimental strategies qualitatively, finding adults more likely to use a “deconfounding” strategy to isolate properties of interest, potentially creating evidence less susceptible to cognitive and perceptual errors.

Keywords: active learning; intuitive physics; Bayesian statistics; mental simulation; cognitive development; action

Children's active physical learning is as effective and goal-targeted as adults'

"The secrets of nature reveal themselves more readily under vexations of the art than when they go their own way." — Francis Bacon (1620)

Introduction

The word 'learning' may conjure images of textbooks and bored students slumped behind desks, yet in its most elemental form, human learning is self-directed, active and interactive, taking place outside the classroom in the real and wild physical world. Children take control of their experiences almost from birth, spending much of their early years pushing, pulling, prodding, chewing and grasping the objects around them, but also actively deciding what is interesting enough to be pushed and grasped, what is worth being chewed. It would be very surprising if these early behaviors were not playing an important role in development. One idea is that these actions are kinds of proto-experiments (Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018; Brewer & Samarapungavan, 1991) that help reveal the deep causal structure and hidden properties of the physical world that are rarely or never revealed by passive observation (Bacon, 1620/1878; Gopnik et al., 2004; Pearl, 2000). In this way, our physical actions may serve the general epistemic goal of building a causal world model that accurately reflects natural laws (Hohwy, 2013) and empowers us to predict, plan and pursue future goals (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017). As adults, we probe task-relevant physical properties of objects almost unconsciously — rock a table to gauge its stability, slide a glass across its surface to gauge its friction, or waft a metal object near a radiator to gauge its magnetism. These actions seem to combine an intuitive understanding of how the physical world works, with expertise in ways to exaggerate, isolate, or bring into sharper relief familiar properties of novel objects (Bramley et al., 2018). However, when this expertise emerges, how it develops, and what qualitative differences there could be in how children and adults probe the world are all open questions.

The goal of this paper is to explore the development of active physical inference by directly comparing children's and adults' behavior when learning about objects in a simulated physical "microworld" setting. To foreshadow, we find that both children and adults produce actions that provide information specific to their learning goals, while minimizing confounding evidence about other non-goal properties. However, older children and adults are more likely to make accurate judgments on the basis of the resultant evidence. Adults also show clearer hallmarks of controlled experimentation (Kuhn & Brannock, 1977), performing more actions that minimize the confounding influence of distractors compared to children.

Despite its complexity, even young children seem to be able to navigate and interact with the physical world far more successfully than cutting-edge AI technology. One line of work argues that this competence stems from the development of a generative model — or ‘intuitive theory’ — of everyday physics (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021). The idea is that learners bootstrap via a general theory of everyday physics, which allows them to work back from observations to infer the particular latent properties and phenomena needed to make mental simulations match specific observations (Ullman, Spelke, Battaglia, & Tenenbaum, 2017), although there is debate over how much intuitive theory is learned versus innate (Stahl & Feigenson, 2015). Such an approach would allow cognizers to use mental simulation to make predictions, imagine hypothetical and counterfactual situations and pursue arbitrary goals (Battaglia, Hamrick, & Tenenbaum, 2013; Smith, de Peres, Vul, & Tenenbaum, 2017). Another line of work has emphasized human limitations in physical reasoning and argued that judgments often reflect application of context-specific rules and heuristics rather than online simulations (Ludwin-Peery, Davis, Bramley, & Gureckis, 2021; McCloskey, 1983; Smith et al., 2017).

Active physical learning is one particular domain in which we might expect cached or heuristic solutions to be important. Active learning research studies how people use their actions to gather evidence and shape their learning (Coenen, Nelson, & Gureckis, 2018). An important subset of this field studies how learners probe the causal structure of the environment through actions or *interventions* (Pearl, 2000) that take control of variables of interest and may reveal the underlying causal structure (Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; Coenen, Ruggeri, Bramley, & Gureckis, 2019). Calculating the most informative action to take to resolve one’s uncertainty is generally prohibitively expensive outside of toy experimental settings, and this is compounded when actions must be chosen, performed, and potentially adjusted in real time (Davis, Bramley, & Rehder, 2020). However, the laws of nature are broadly universal, meaning there is a good degree of stability in what behaviors are likely to be effective for a generic physical enquiry goal. For instance, if an action such as “lifting” reveals the mass of one object, is also likely to reveal the mass of another, and so on, making the caching of general heuristics for identifying particular properties a computationally sensible idea (cf. Gershman, Horvitz, & Tenenbaum, 2015). Crucially, the relevant evidence for learning about physics is not the state of the world at a particular moment in time, but rather how its state *evolves over time*.

Ullman, Stuhlmüller, Goodman, and Tenenbaum (2018) explored adults’ inferences about relative masses, local (magnet-like) and global (gravity-like) forces, and friction from video clips of simulated 2D physics. They found that participants struggled to identify

masses, and were better at detecting local attraction than repulsion. These patterns were partly captured by the evidence available from an idealized simulation-based inference model. Until colliding with another dynamic object, an object's motion provided no evidence about its mass, and even collisions revealed only the *relative* masses of the objects. For the local (magnet-like) forces, objects that repelled one another would rarely stay long enough close together to exhibit strong evidence of their repulsion, while attracting objects would rapidly approach one another and stick together offering stronger extended displays of their attraction.

In the current paper we extend this work to the active setting, adapting the paradigm from Bramley et al. (2018) to investigate the developmental trajectory of active physics learning. In our task, participants can drag objects around using touch control on a tablet screen (see Figure 1a). While this is admittedly far simpler than real world control, such “billiard worlds” (Fragkiadaki, Agrawal, Levine, & Malik, 2015) have proven to be valuable for exploring intuitive judgments about physics (cf. Bramley et al., 2018; Smith et al., 2017). Action planning in physics learning tasks is particularly challenging (Li et al., 2019). Indeed, for learning to succeed, the right kind of dynamics have to be observed or brought about through control, and for this to happen, the learner must not only choose *where* (i.e., on which object), but also *how* and *when* to intervene (Gerstenberg et al., 2021).

Using a similar paradigm, Bramley et al. (2018) found that adults tailored their active intervention strategies to maximize the informativeness of the actions performed depending on their given learning goal. For example, participants with a goal of identifying the heavier of two objects would frequently knock them together or take turns shaking them from side to side. Likewise, participants with a goal of identifying the nature of a local force between two objects would frequently hold them close together.

Previous work on children's active learning indicates that even toddlers and preschoolers spontaneously make informative interventions to disambiguate the causal structure of a system, both in experimental settings and during spontaneous play (Cook, Goodman, & Schulz, 2011; Kushnir & Gopnik, 2005; L. E. Schulz & Bonawitz, 2007; Sim & Xu, 2017), and that the efficiency of these interventions increases with age (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016). In causal learning, younger learners (4-year-olds) are more flexible than older learners (6-year-olds; Gopnik & Bonawitz, 2015) and even adults in correctly drawing inferences about unusual causal relationships from observation (Lucas, Bridgers, Griffiths, & Gopnik, 2014). Moreover, preschoolers' causal learning performance bears hallmarks of Bayesian learning by age 4 (Sobel, Tenenbaum, & Gopnik, 2004), although children sometimes can perform informative interventions yet fail to integrate the evidence correctly (Meng, Bramley, & Xu, 2018). More recent work shows

that while even 3- and 4-year-olds can rely on different exploratory strategies depending on the statistical structure of a task, *selecting* the most efficient strategy from among the given options (Ruggeri, Swaboda, Sim, & Gopnik, 2019), only by 7 years of age do children start to be able to *generate* informative actions from scratch (Ruggeri & Lombrozo, 2015; Ruggeri, Lombrozo, Griffiths, & Xu, 2016). Together, these findings suggest that children, just like adults, may be able to tailor their actions to provide information specific to their learning goals and make accurate judgments on the basis of the observed evidence. In fact, it is plausible that because physical active learning is more developmentally basic compared to the tasks used in much of the previous literature on active learning — involving the objects of perception more directly and depending less on maintenance and application of linguistic concepts — children would excel.

In the current work, we contrast passive learning participants who observe simulated natural dynamics with active learning participants can additionally interact with the simulated objects. As recent developmental studies suggest that the advantage of active control over the learning experience are fairly stable across the lifespan (see Ruggeri, Markant, et al., 2019), we predict a similar performance boost for active over passive learning as was found in adults by Bramley et al. (2018).

Experiment

Methods

Participants

In total, we recruited 125 participants in museums around Berlin including 101 children (45 female, $M \pm SD$ age 7.15 ± 1.58 , range: 4.6–10.2 years, completion time 12.8 ± 0.90 minutes) and 24 adults (12 female, 37.7 ± 8.62 , range: 25.2–56.4 years, taking 11.9 ± 0.81 minutes).¹ Participants were predominantly white Europeans from diverse social classes and were native German speakers or fluent in German. IRB approval was obtained from the ethics committee of the Max Planck Institute for Human Development, Berlin (protocol: “Active Physics Learning”), and parents gave informed consent for their children to participate before the study. The study was not preregistered.

Some trial information was incorrectly stored for early participants resulting in slightly smaller sample of 105 participants for whom we could perform detailed action and

¹ Our sample size plan was originally based on a G*power calculation seeking $\geq 80\%$ power to detect a moderate (0.3) continuous effect of age, the variable we were most interested in, yielding a required sample size of 82 (Faul, Erdfelder, Buchner, & Lang, 2009). Eventually though, as this was just one of a number of data-dimensions we evaluated, we opted to use Bayesian statistics for all primary analyses, to better capture the relative strengths of our various conclusions.

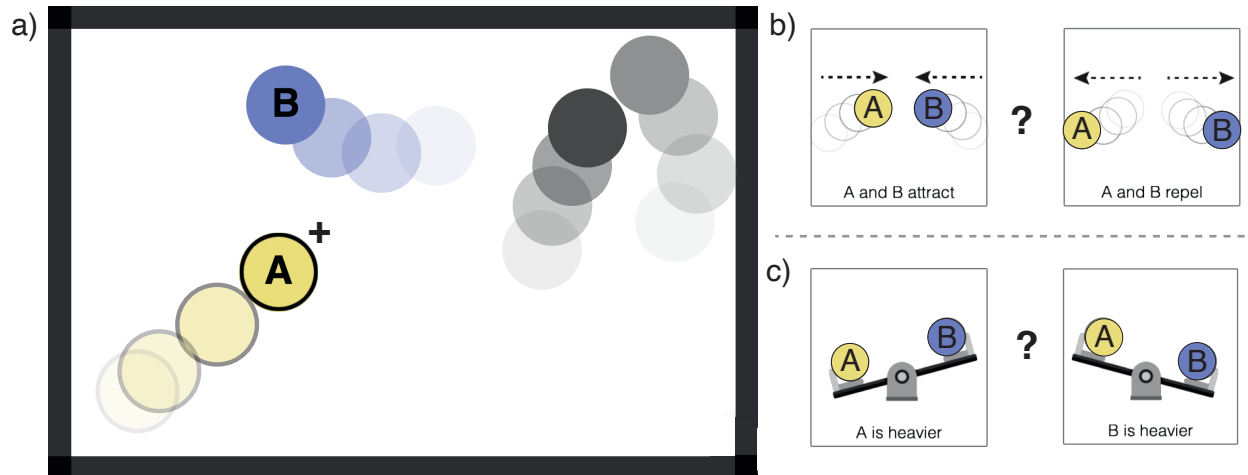


Figure 1

(a) Visualization of task: 2 colored “target” objects and 2 gray “distractor” objects move around, colliding and affecting one another with local (magnet-like) forces. A participant (Active condition) grabs object “A” (active condition only) by hold-pressing on it on touch screen and drags it upward and right (“+” symbol shows final position of touch control). (b–c) Response buttons displayed at end of the force-focused and mass-focused trials respectively.

information-based analyses. The smaller sample included 82 Children (37 female, 7.15 ± 1.56 , range: 5, oldest 10.2 years) and 23 adults (11 female, 37.5 ± 8.77 , range: 25.2–56.4 years, 12.6 ± 1.00 minutes).

Stimuli, Design and Procedure

The experiment was administered on a 10 inch Android tablet as a full screen web app, programmed in Javascript using a port of the Box2D physics game engine and optimized for tablet screens and touch control. Complete specification of the settings of the Box2D simulator is available in the Supplement and a source code for the experiment are available in the OSF Repository (<https://osf.io/v9fk2/>).

Participants were first introduced to the task. They were shown four objects of different colors moving around the screen bounded by solid walls with high elasticity (see Figure 1a), and were told that their task was to learn about physical properties of some of the objects. Participants were pseudo-randomly assigned to one of two experimental conditions, *passive* or *active*. In the passive condition, participants had to infer the objects’ physical properties by merely *observing* them moving around. In the active condition, participants could also interact with them on the screen using their finger to grab and drag them around, so creating different and potentially more or less informative dynamics. Participants in both conditions completed a practice trial to familiarize themselves with

the game procedure either observing four unlabeled objects moving around for 45 seconds (passive condition), or moving the objects with their finger (active condition) for 45 seconds before moving to the experimental session.

The experimental session then consisted two blocks, presented in pseudo-random order: (1) A *force*-focused block, in which participants were told they had to find out whether two “target” objects attracted or repelled each other, and (2) A *mass*-focused block, in which participants had to find out which of the target objects was heavier. Each block included 4 trials in random order in which the target properties were varied systematically so as to counterbalance the goal with two possible settings for each target property (see Table 1). Targets were labelled as “A” and “B”, while distractor objects were unlabeled (see Figure 1a). A fixed sequence of color pairs were used for the Target objects across the 8 trials. These were matched for saturation and lightness and spread pseudorandomly around the hue wheel. This was done to ensure participants differentiated clearly between different target objects both within and across trials. Distractor objects were always light and dark gray. At the beginning of each trial the learning goal was displayed and read out by the experimenter. At the end of each trial, the objects froze in position and the response options appeared on top (see Figure 1b–c). The experimenter read out the question again and, if necessary, explained to the participant that they should select the option they thought was true of the environment they had just observed or interacted with.

The target objects always either attracted or repelled one another, and one was always heavier than the other.² Additionally, the specific behavior of the objects differed substantially in every trial because we independently drew distractor forces for the other five combinations of target and distractor objects (uniformly from $\{attract, none, repel\}$) and independently randomized the initial locations and velocities of all objects. For active participants, the dynamics they observed also critically depended on whatever control they exerted using their finger on the touch screen. Concretely, taking control of an object made it temporarily elastically attracted to the position of the participant’s finger on the touch screen. In this sense, the object was moved by the finger, while being able to participate realistically in reciprocal physical interactions with the other objects, such as in collisions. Each trial lasted 45 seconds during which the physics simulator updated the positions of the objects 60 times per second for a total of 2700 frames of evidence. Video replays of all participants and trials are available in the OSF Repository (<https://osf.io/v9fk2/>).

At the end of the experimental session, the interface showed the participant which

² Pairwise forces were ± 3 Newtons. The heavier object weighed 2kg while the other objects weighed 1kg.

and how many of the 8 trials they had answered correctly and had the experimenter enter their basic demographics. Children were rewarded with one sticker per correct trial.

Table 1

Experiment Design

Block	1.				2.			
Goal:	Identify force				Identify mass			
Trial	1.	2.	3.	4.	1.	2.	3.	4.
True force:	attract	attract	repel	repel	attract	attract	repel	repel
True mass:	A heavy	B heavy	A heavy	B heavy	A heavy	B heavy	A heavy	B heavy

Results

We first analyze participants' performance by age-group, continuous age, condition and block. We then turn to analysis of the actions of participants in the active learning condition. Our *Performance* level analyses use our full sample ($n = 125$), but for the more detailed *Information* and *Actions* analyses we use the smaller sample for which we have complete records ($n = 105$, see *Participants*). We analyze the data primarily with Bayesian mixed-effects regressions. We include 95% posterior credible intervals for each parameter. Note that whether the interval for a given parameter includes the null value of zero is a common statistical decision criterion, and perhaps the closest analog to frequentist decisions about whether one should rejecting the null of no effect (Kruschke, 2013). We also use prior and posterior samples from each model to compute Bayes Factors (BF) for each parameter of interest, allowing us to assess the strength of the evidence favoring either the existence of the effect or the null. We additionally include a posterior probability of direction statistic for all primary results. This captures the proportion of the posterior density that lies on the favored side of the null, essentially measuring how confident we can be about the direction of the effect, conditional on the effect existing (Makowski, Ben-Shachar, Chen, & Lüdtke, 2019). In the Supplement, we detail our choice of priors, full result tables and repeat all analyses with a standard maximum likelihood mixed-effects analysis, demonstrating a close correspondence.

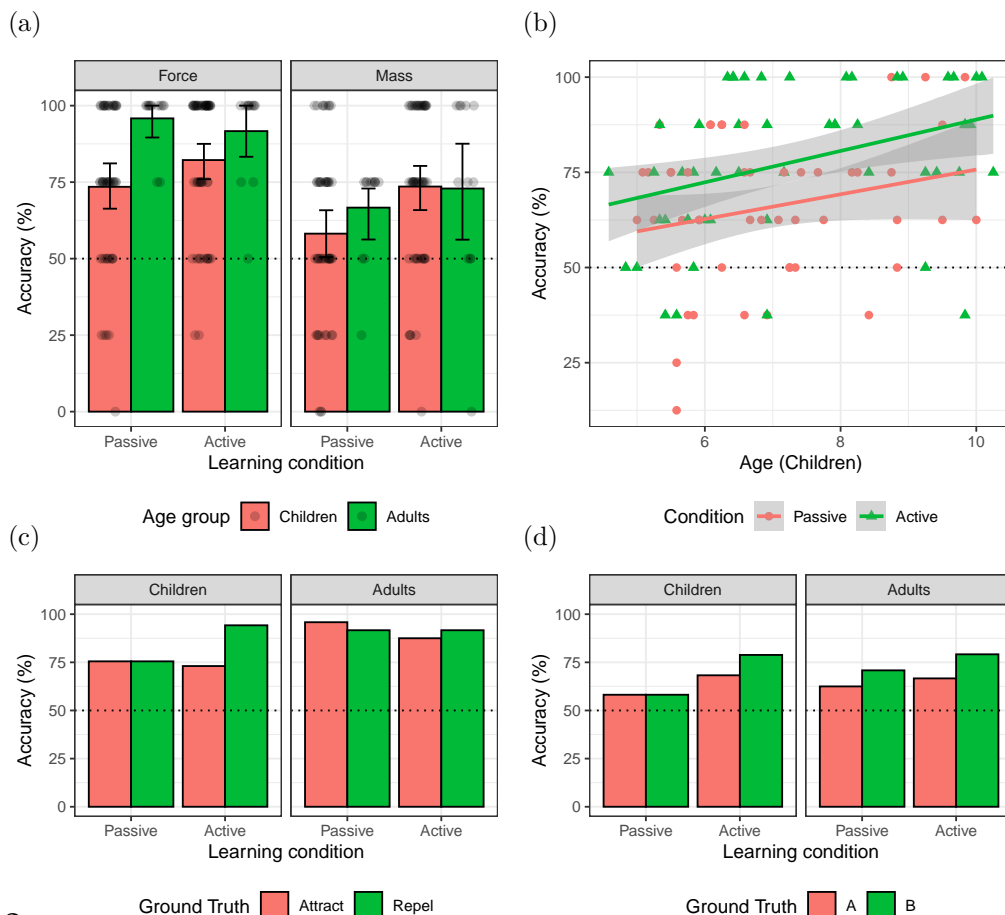
Performance

We first performed a Bayesian mixed-effects regression, with fixed effects of learning condition, age-group (between subject) and block (within subject) and a random intercepts for each participant, predicting percentage correct judgments. This revealed significant main effects of all three factors, with active learners making more accurate inferences

($M \pm SD = 78.7 \pm 16.6\%$) than passive learners ($68.9 \pm 18.9\%$; $M \pm SE \beta = 9.88 \pm 3.36$, 95% Credible Interval [95%CI]=[3.3,16.4]%, P direction [PD] = .998, Bayes Factor [BF]=4.05), adults making more accurate inferences ($81.8 \pm 16.1\%$) than children ($72.0 \pm 20.2\%$; $\beta = 9.86 \pm 4.24$, 95%CI=[1.52,18.2]%, PD = 0.99, BF=1.28), and force judgments ($81.0 \pm 23.6\%$) being more accurate than mass judgments ($66.8 \pm 27.2\%$; $\beta = 14.1 \pm 2.92$, 95%CI=[8.34,19.9]%, PD>.999, BF>1000; Figure 2a, Supplemental Table S2). Repeating the analysis including all potential interaction terms suggested the lack of any two- or three-way interactions between these factors (all posterior 95% credible intervals including zero, all Bayes factors between 0.23 and 0.46, indicating anecdotal to moderate evidence for the null Jeffreys, 1961). For the 101 children, we additionally ran a Bayesian mixed-effects regression predicting accuracy by continuous age, condition and block, again with a random intercept per participant. This revealed a linear age improvement $M \pm SE \beta_{\text{age}} = 3.8 \pm 1.2\%/year$, 95%CI= [1.45, 6.15]%/year, PD = .999, BF=3.0 alongside main effects for active over passive learning children $\beta_{\text{condition}} = 10.6 \pm 3.74\%$, 95%CI=[3.2,17.9]%, PD=.998, BF=4.1, and on the force over the mass block $\beta_{\text{block}} = -11.8 \pm 3.41\%$, 95%CI=[-18.5,-5.04]%, PD>.999, BF=17.8 (see Figure 2b, Supplemental Table S3). Repeating the analysis including interaction terms found support for nulls of no two- or three-way interactions (all 95% posterior CIs including zero, all BFs between 0.062 and 0.58), the data also supported the null of no quadratic effect of age (-0.82 ± 0.86 , 95%CI=[-2.52,0.88], PD = 0.83, BF = 0.027).

Finally, we used a logistic mixed-effect regression to assess accuracy differences depending on the ground truth in force and mass blocks respectively (Figure 2c–d, Supplemental Tables S3&S4). For force responses, as predicted, this suggested an interaction with learning condition, such that participants in the active condition were more accurate on repel trials (log odds ratio = 1.41 ± 0.46 , 95%CI=[.5,2.33], PD =0.999, BF=53). Response type did not appear to interact with age-group in predicting accuracy on the force questions (log odds ratio = 0.432 ± 0.692 , 95%CI=[-0.898,1.82], PD = 0.73, BF=0.825). For the mass question there is no reason to expect a difference between “A heavy” and “B heavy” trials, since the goals are qualitatively identical. Accordingly, the data supported the null of no main effect of ground truth nor interaction with condition nor age-group (all CIs include zero, all BFs between 0.28 and 1.01).

In sum, we found that children's accuracy improved with age, that adults were more accurate than children overall, and that active learners were more accurate than passive. As predicted, this was driven by an active learning accuracy boost on the mass block and on repulsion trials on the force block. However, perhaps surprisingly given mixed active learning efficiency in past developmental work, these active learning effects did not differ in

**Figure 2**

(a) Mean accuracy (\pm bootstrapped SE) by block (panels Force goal vs. Mass goal), learning condition and age-group. Points (jittered in y axis) show individual participants' averages. (b) Accuracy by age and learning condition for children. Lines show linear best fit and shaded areas show 95% confidence intervals. Proportion correct by ground truth on (c) Force trials and (d) Mass trials. Dotted black lines in all plots indicate chance performance.

magnitude between 4-10-year-olds and adults.

Information

We now use an Ideal Observer (IO) model to better understand why active participants generally outperformed passive participants, and whether differences in active learning can explain children's lower accuracy relative to adults. Our IO model assumes learners begin each trial maximally uncertain about the target and distractor properties within their support³ and update their beliefs based on the evidence available throughout

³ For the two target objects followed by the two distractor objects, these are $\text{mass} \in \{[2, 1, 1, 1], [1, 2, 1, 1]\}$, $\text{force} \in \{\text{attract}, \text{repel}\}$ between target objects and $\in \{\text{attract}, \text{repel}, \text{none}\}$ for the other five pairwise combinations of target and distract objects, leading to a nominal hypothesis space of 972 microworld

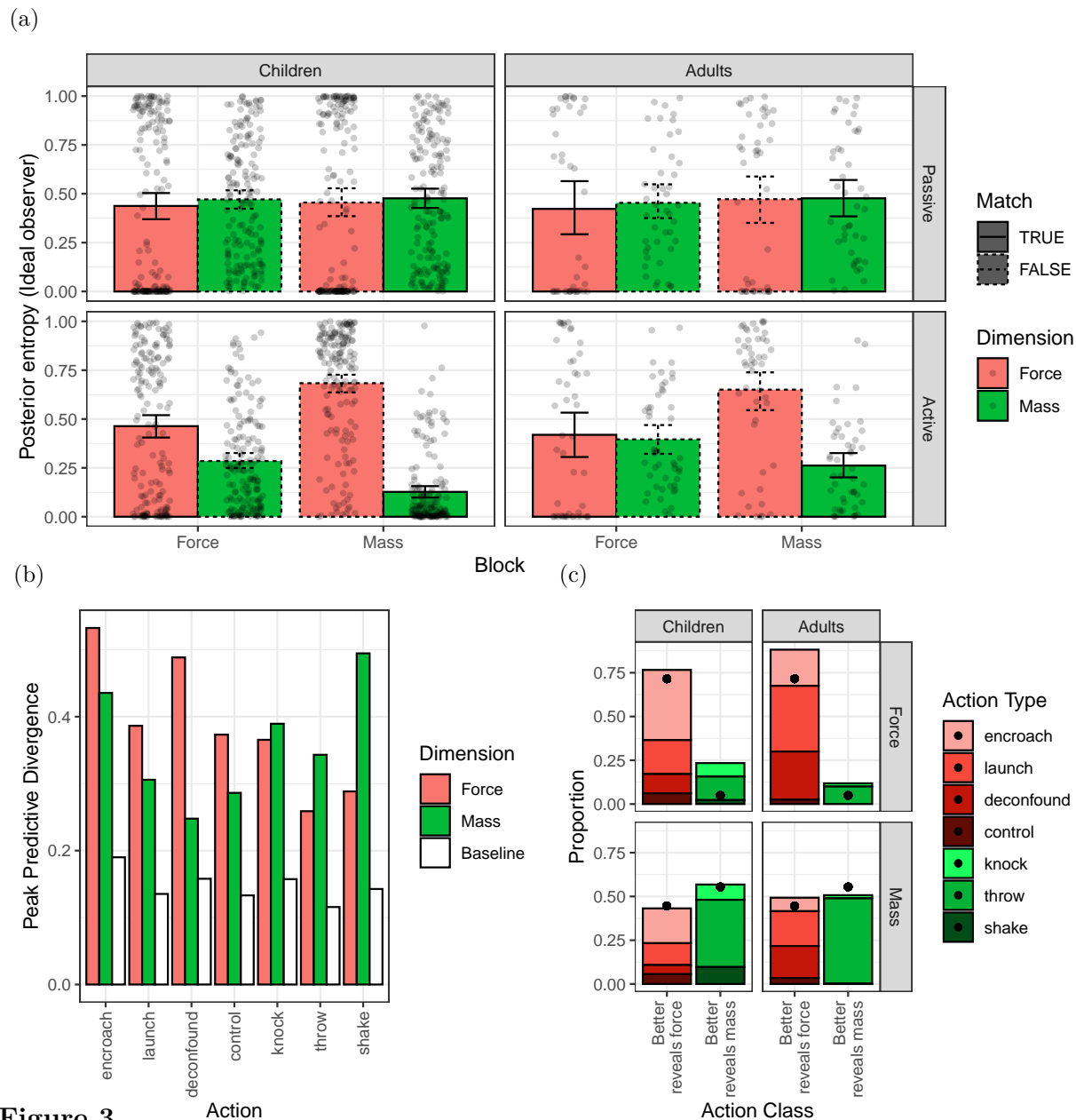


Figure 3

(a) Mean posterior entropy by age-group and condition (\pm bootstrapped SE) according to Ideal Observer account. Lower means more useful information was gathered. Points (jittered in y axis) show individual trials. (b) Peak predictive divergence for the averaged profiles of each action type (see Table 3). Higher is more informative. (c) Coded action type frequencies by age-group and condition.

the clip, using a simulation-based inference procedure detailed in Bramley et al. (2018) to approximate the likelihood of the evidence under different parameter settings. The model

settings.

takes the dynamics of the scene on a frame-by-frame basis as its input, but also uses its own physics simulator to generate many hypothetical forward simulations, using discrepancies between these simulations and future observations to drive inferences about the true environment settings. This procedure results in a joint posterior over the unknown properties of the objects in the trial based on all the evidence revealed by the object dynamics on each trial. This can then be marginalized over to produce specific posteriors for the target mass, target force (or equally for any of the distractor force properties). The entropy of these posteriors gives a measure of the total evidence available to the participant (Shannon, 1951) where lower values indicate more information was produced. The accuracy of such an ideal observer depends on the degree of sensory precision we assume and consequently the quality and quantity of the dynamics it would need to observe to form a reliable preference for the true property. Therefore, we are not interested in comparing participants and models in terms of *absolute* uncertainty but rather, in assessing the extent to which the model's notion of *relative* strength of evidence lines up with differences in children's and adults' judgments.

Our IO analysis reveals that the amount of evidence participants gathered about the target properties is dynamically related to their learning condition and block but that it does not depend significantly on age-group (see Figure 3a). Critically, posterior force uncertainty was lower for active participants on force-focused trials ($M \pm SD$ $.45 \pm .38$ bits) than mass-focused trials (0.68 ± 0.31 bits) while mass uncertainty was likewise lower for active participants' mass-focused trials (0.16 ± 0.21 bits) than force-focused trials (0.31 ± 0.26). This was true for both Children (Force: 0.46 ± 0.37 vs. 0.68 ± 0.30 , Mass: 0.13 ± 0.19 vs. 0.29 ± 0.26) and Adults (Force: 0.42 ± 0.40 vs. 0.65 ± 0.34 , Mass: 0.26 ± 0.23 vs. 0.40 ± 0.26) taken separately. This shows that Active participants took actions that dynamically and successfully targeted their learning goals. Somewhat unexpectedly, force uncertainty was lower on average for passive than for active force-focused trials (0.447 ± 0.06 and 0.56 ± 0.13 bits respectively). However, mass uncertainty was lower for active ($.23 \pm 0.15$ bits) than passive mass-focused trials ($0.47 \pm .11$ bits). To unpack the full pattern of results, we ran a Bayesian mixed-effects regression predicting posterior entropy by condition, age-group (between subject) and dimension plus whether the learners' goal matched that dimension (within subject), including random intercept terms per participant as above. This is summarized in Table 2.

This confirms that target-specific IO uncertainty was substantially lower in the active condition (row 2), lower for mass than force (row 3) but critically also lower on the dimension matching the block goal (row 4). There was also a condition \times dimension interaction (row 7) indicating that the participants in the active condition saw more

advantage on the mass than the force dimension, condition \times match interaction (row 9) indicating that condition differences were driven by the active condition. Strikingly, this analysis reveals “very strong evidence” ($BF < \frac{1}{30}$, Jeffreys, 1961) *against* a direct effect of age-group (row 1), and “strong” evidence against the existence of an interaction between age-group and any other design dimension ($BFs < 0.1$) in shaping the evidence participants produced with their actions.

Table 2

Bayesian Mixed-Effects Regression Predicting Posterior Entropy

	Effect	$M \pm SE$	95%CI	PD	BF
<i>Main effects only model</i>					
1.	Age-group (B.S. <i>Children</i> =.425, <i>Adults</i> =.444)	0.020 ± 0.021	[-0.02, 0.06]	.824	0.033
2.	Condition (B.S. <i>Passive</i> =.459, <i>Active</i> =.399)	-0.06 ± 0.018	[-0.095, -0.026]*	>.999	5.54
3.	Dimension (W.S. <i>Force</i> =.506, <i>Mass</i> =.352)	-0.15 ± 0.017	[-0.19, -0.12]*	>.999	>1000
4.	Match (W.S. <i>True</i> =.380, <i>False</i> =.478)	0.098 ± 0.018	[0.06, 0.13]*	>.999	>1000
<i>Full interaction model</i>					
5.	Age-group:Condition	-0.028 ± 0.079	[-0.18, 0.13]	.64	0.0839
6.	Age-group:Dimension	0.017 ± 0.080	[-0.15, 0.18]	.582	0.0806
7.	Condition:Dimension	-0.37 ± 0.053	[-0.48, -0.27]*	>.999	>1000
8.	Age-group:Match	0.032 ± 0.080	[-0.13, 0.19]	.655	0.0853
9.	Condition:Match	0.20 ± 0.052	[0.10, 0.31]*	>.999	127
10.	Dimension:Match	-0.023 ± 0.052	[-0.13, 0.08]	0.667	0.0579
11.	Age-group:Condition:Dimension	0.16 ± 0.11	[-0.059, 0.38]	.924	0.319
12.	Age-group:Condition:Match	-0.022 ± 0.11	[-0.24, 0.19]	.579	0.111
13.	Age-group:Dimension:Match	-0.051 ± 0.11	[-0.27, 0.17]	.674	0.123
14.	Condition:Dimension:Match	-0.039 ± 0.075	[-0.19, 0.11]	.703	0.0853
15.	Age-group:Condition:Dimension:Match	0.017 ± 0.16	[-0.29, 0.32]	.544	0.154

Note: Marginal means provided for factor levels in Effect column. Match: True = Dimension for which uncertainty is calculated matches the learners' goal. Intercepts and main effect terms for interaction model omitted from table for brevity. * indicates the 95%CI excludes zero (no effect). See Supplement for comparison with maximum likelihood mixed-effects fit.

Extending this, across the 41 children in the active condition for whom we have complete action data, we found evidence in support of the null of no relationship between continuous age and posterior goal-specific entropy ($\beta = 0.00174 \pm 0.0066$, 95%CI=[-0.011,0.014], PD = .61, BF = 0.0067, Supplemental Table S7).

Finally, we can ask whether these evidence quality patterns are directly associated with accuracy at the trial level. Bayesian logistic mixed-effects model predicting $P(\text{Correct})$ by entropy and including age-group, condition and block as covariates, suggests this is not the case ($\beta_{Entropy} = -0.17 \pm 0.24$, 95%CI=[-0.65,0.30], PD = .75, BF=0.32, Supplemental Table S8).

In sum, we found that children and adults were similarly able to use touch control to gather evidence targeted to their learning goal. As expected and consistent with the

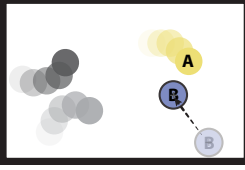
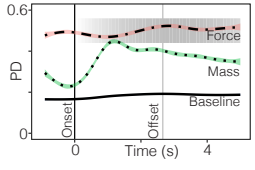
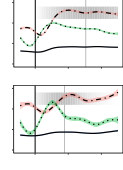
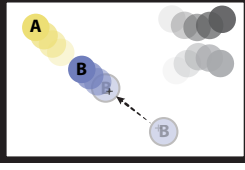
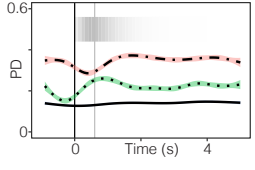
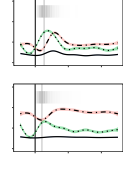
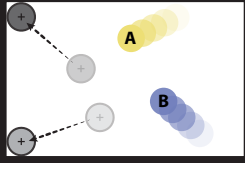
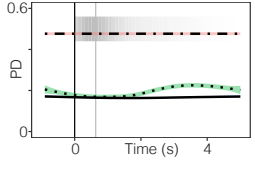
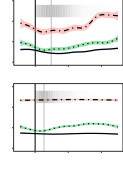
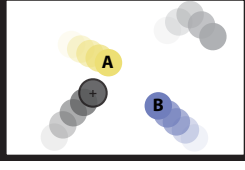
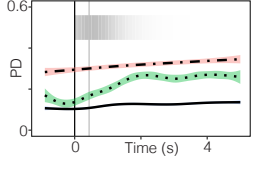
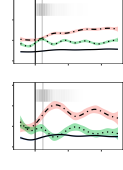
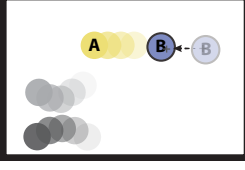
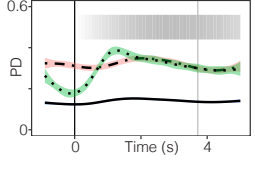
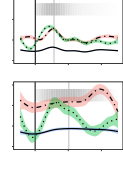
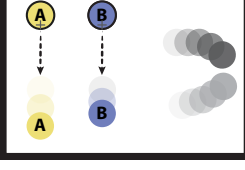
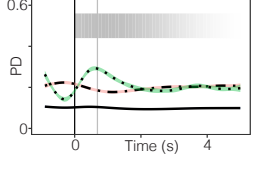
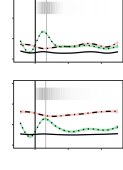
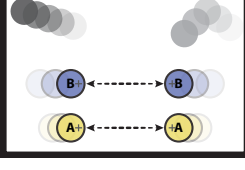
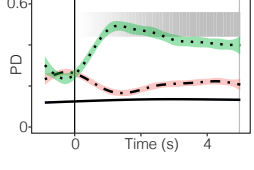
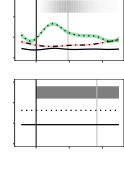
accuracy patterns, active learners substantially increased their evidence about masses, while, for force, there was an effect of learning goal in the expected direction but an overall decrease in the strength of evidence relative to passive observation. Most interestingly, children's actions were as informative and reactive to the learning goal as adults' and appeared stable across the 4-10 age range. Within the child sample, informativeness of actions did not improve across the age range we tested despite the fact that older children went on to make more accurate judgments about the properties. Thus, in this setting, and distinct from some classic findings in higher-level cognition settings, 4-10-year-olds appear to be at least as capable as adults at probing latent physical properties in goal directed ways, yet less reliable than adults at using this evidence in explicit judgments.

Actions

Since children's actions are as informative as adults', a remaining question is whether there are systematic differences in the specific types of actions children and adults performed in service of learning. To explore this question, we now classify each of the actions that children and adults performed using their touch screen control. We used the seven qualitative categories of "micro strategy" detailed in Table 3, using the protocol developed in (Bramley et al., 2018). We then analyze how frequently these categories of action were used depending on age-group and learning goal. We recorded 4412 instances in which a participant in the active condition took control of an object for at least $1/6$ th of a second.⁴ We had two independent coders, blind to our experimental hypotheses and condition, watch video replays of each of each trial, pausing at each action and classifying them according to our coding scheme using video coding software DataVyu (Datavyu Team, 2014). We were able to identify a primary category for 84% of participants actions (see Supplementary materials for full coding resources, checks and merging). Taking these classifications as labels, we then created average informativeness "profiles" for each type of action using the *Predictive Divergence* quantity developed by (Bramley et al., 2018). Conceptually, Predictive Divergence measures how strongly informative the instantaneous dynamics are about each physical property, providing an online "timeline" measure of strength of evidence. For instance, an action that brings about a situation in which the subsequent direction and velocities of the objects is expected to be very different if there is a repulsive rather than an attractive relationship has a high Predictive Force Divergence and similarly for actions whose outcomes depend a lot on mass have high Predictive Mass Divergence. We compare these against a Predictive

⁴ We excluded 341 extremely short actions on the basis that they are likely to have been accidental and are too short to meaningfully categorize.

Table 3*Strategies observed in Experiment 1.*

Strategy	Schematic	Profiles: Children	Adults
a) Encroaching: Grabbing one target object and moving it close to the other.			
b) Launching: Grabbing one of the target objects and “throwing it” against the other target object.			
c) Deconfounding: Grabbing a distractor object and moving it away from the target objects (e.g. into a corner).			
d) Controlling: Briefly grabbing a fast moving object and releasing it to slow it down.			
e) Knocking: Grabbing one of the target objects and knocking it against the other (without letting it go)			
f) Throwing: Grabbing a target object and throwing it, avoiding collision with any of the other objects.			
g) Shaking: Grabbing a target object and rapidly shaking it from side to side.			

Schematic: *Target objects* are labeled A or B, and *distractor objects* are unlabeled. **Profiles:** Predictive divergence profiles for coded actions smoothed using a GAM (Hastie & Tibshirani, 1990), with fills showing 99% confidence intervals. Black vertical lines mark onset of control. Shaded horizontal fills and gray vertical lines respectively indicate range and median time after onset at which the participant let go of the object. Children (N=44; actions=3115). For comparison, final column shows action profiles from Bramley et al. (2018) (top; N=40; actions=1829); and for Adults (bottom; N=12; actions=1297).

Baseline Divergence which is averaged across all six varied target and non-target properties of the worlds. These smoothed information profiles are shown in Table 3, separated by age-group and also compared to the Adult sample in (Bramley et al., 2018). This analysis reveals that all of these actions generally produce more information about target than non-target properties (colored lines are clearly above the black baseline) but that some are better at revealing force than mass and visa-versa. Figure 3b shows the peak PD for each strategy profile. On the basis of this we classified “Encroaching”, “Launching”, “Deconfounding”, “Controlling” as primarily force-revealing, and “Knocking”, “Throwing” and “Shaking” as primarily mass-revealing. Figure 3c then shows the frequency of of the different action labels by age-group and learning goal. A Bayesian mixed-effects beta regression predicting proportion of force-revealing actions by block and age-group, including random effects of participant ID confirms that participants performed far more mass-revealing actions in the mass block (log odds ratio 1.12 ± 0.16 , 95%CI=[0.813, 0.155], PD>.999, BF>1000), and that children performed a slightly greater proportion of mass revealing actions than adults (log odds ratio, = -0.428 ± 0.22 , 95%CI=[-0.85,-0.007], PD = 0.977, BF=1.49). There was anecdotal evidence also for an interaction between block and age-group (log odds ratio 0.585 ± 0.339 , 95%CI=[-0.0754, 1.25], PD = 0.959, BF=1.49). Finally, to compare individual strategy prevalences, we performed Bayesian mixed-effect Poisson regressions predicting the count for each strategy type by block and age-group, with an without interaction term, again including participant ID as a random effect. As above, we include full results and equivalent maximum likelihood regression with Bonferroni correction in the Supplement (Table S9). Frequency of encroaching, launching, deconfounding, throwing and shaking all differed substantially between blocks in the expected directions (BFs > 29.1). Deconfounding and throwing were more common in adults (log count ratio for deconfounding = 1.43 ± 0.44 , 95%CI=[0.57,2.3], PD = .999, BF=63.6 and throwing = 0.94 ± 0.34 [0.27,1.59], PD = 0.995, BF = 11.5) and shaking was more common in children (-2.25 ± 0.57 , 95%CI=[-3.4,-1.16], PD >.999, BF>1000).

Thus, we see children and adults apply broadly similar strategies to active learning with the frequency of performing different actions clearly related to learning goal in ways that line up with their information content. This analysis revealed differences between children’s and adults’ actions, with adults more likely to perform deconfounding actions. Conceptually, these served to reduce the influence of distractor objects and provided clearer evidence about force particularly as the PD profiles show (Table 3), thus we might link this kind of behaviour with a kind of “control of variables” approach (Kuhn & Brannock, 1977). Shaking turned out to be one of the most effective ways to reveal an objects mass, yet curiously our sample of adults essentially never performed “shaking”

(1/12 in active condition) while most of the children did (31/41).

Discussion

Our results show that children's active physical learning is as effective and goal-adaptive as adults', and suggests that age differences in performance arise only in the ability to *use* the information generated to make accurate inferences. Although we expected children to perform meaningful interventions, it was striking to find 4-year-olds already exhibiting adult levels of sophistication and goal-specificity in probing the properties of the objects in our simulated physical microworlds. This suggests that mastery in active physical learning emerges earlier in development than in any other context studied so far, where adult levels of performance can be found only in late childhood or even later (e.g., question asking or exploration, see Ruggeri & Lombrozo, 2015; E. Schulz, Bertram, Hofer, & Nelson, 2019).

Our measures of information were based on tracking a set of objects' trajectories simultaneously and comparing these against counterfactual simulations that each resolve the multi-body interactions without fault (Ludwin-Peery et al., 2021). While this provides a valuable benchmark, examining more computationally frugal and fallible models of physical reasoning has the potential to add nuance to our notions of about what constitutes informative dynamics for bounded learners (cf. Bass, Smith, Bonawitz, & Ullman, 2021; Ullman et al., 2017). That is, it is interesting to consider how computational limitations might shape what constitutes a physically informative action for human learners. For instance, according to our Ideal Observer, participants produced more evidence about mass than about force, yet both children and adults were more accurate in their force than in their mass judgments. This discrepancy is surprising considering that people have well-calibrated intuitions about Michottean collisions (Vicovaro, 2018), and these provide one of the key sources of mass evidence for the Ideal Observer. On the other hand, the ways that dynamics reveal force and mass may be differentially robust to perceptual, attentional and computational limitations. Our predictive divergence measure shows that mass evidence was low most of the time, but spiked briefly whenever the target objects collided — because the angle of reflection of objects in collisions depends on their relative masses — and also, in the active condition, when target objects were moved rapidly under control—because heavier objects react more sluggishly to control than lighter objects. It will be familiar to anyone who has spent time playing bar billiards that the angle of exit from collisions between spherical objects is sensitive to their exact angle of impact — a few millimeters' imprecision will generally result in a missed pocket. This implies that even modest sensory imprecision in tracking such objects through collisions

will dramatically degrade the quality of predicted post-collision trajectories, concomitantly reducing the quality of the resulting inferences (cf. Smith et al., 2017). In contrast, characteristically force-related dynamics seem more easily recognizable under perceptual uncertainty, and presumably are also more robust to uncertainty about the other latent physical properties (including object masses). Most of the time, attractive objects will swerve toward one another while repulsive objects will swerve away from one another.⁵ The angle of the curvature depends on latent properties including objects' masses, but the type of force is recognizable even without accurately predicting the path. According to our predictive divergence measure, force evidence was more consistently present than mass evidence. This may have also made force judgments more robust to limited attention, making it less critical to be looking in the right place at the right moment.

These considerations suggest that process-level accounts of active physical inference might model people as learning to perform actions that they expect will produce dynamics that are robust to uncertainty imprecision but also depend strongly on the latent property of interest. One avenue for future investigation is to explore to what extent children's and adults' epistemic actions in the physical world are shaped by their perceptual and computational limitations (Gong, Gerstenberg, Mayrhofer, & Bramley, Submitted). In the current context, this might shift the balance away from curating highly sensitive "Michottean" multi-object interactions towards repeating more robust and stereotyped actions (such as shaking or lifting) in the case of mass identification. Along these lines we saw, anecdotally, that participants would often perform the same kind of action repeatedly, e.g. shaking or launching both targets with a similar motion (e.g. see video replays included in the OSF Repository (<https://osf.io/v9fk2/>)). We might understand these behaviors as proto-experimental, facilitating aggregation of summary-statistics (Ullman et al., 2018), using repetition to average out sources of inferential noise stemming from both perceptual and computational limitations. Interestingly, this appears to be as much a feature of children's active physics learning as adults'. However, adults were more likely than children to perform deconfounding actions: moving non-target objects out of the way, so reducing their influences on the dynamics. We expect both children and adults to struggle to simulate all the objects and forces simultaneously (Ludwin-Peery et al., 2021; Ullman et al., 2017). Thus, adults' tendency to deconfound may be indicative of a better awareness of the need to minimize influence of confounds and so reduce the chance of attribution errors.

It is worth noting that, while the simulated environments were more physically realistic compared to other active learning tasks from the previous literature, they are still

⁵ Provided the objects are free to move, not too far apart, and their force is not systematically counteracted by those of the distractor objects.

a substantial departure from the real world. One basic limitation of touchscreen control is that it lacks haptic feedback. For example, if interested in the mass of an object in the real world, one might test how difficult it is to move by monitoring the degree of pushback when attempting to move it, rather than monitoring how far it moves when given a fixed impulse. Despite this basic physical information channel being absent in this task, adults and children were still able to adjust and interact with the environments in epistemically valuable ways, quickly settling on goal-specific strategies and applying these consistently. We found no evidence for accuracy improvement or strategy change across trials in either age group. As another marked departure from the real world, our birds-eye “billiard world” scenarios eliminate the normal role of gravity. In everyday life, one can use gravity as a familiar constant to help investigate other properties. For example one might use the effort required to lift something off the ground as a generic way to estimate its mass, or use a hanging object’s deflection from vertical to measure its susceptibility to a magnet-like force. Indeed, everyday mechanisms, from scales to wind socks, use gravity as an indirect means to measure another property. It would be possible to probe the role of these remaining dimensions of real-world learning through increasingly immersive virtual reality learning tasks, e.g. where simulated objects move in three dimensions and haptic gloves simulate pushback from contact. Alternatively, one could record children or adults’ interactions with real environments using sensors (cf. Kretch & Adolph, 2017). However, crossing the “reality gap” too quickly runs the risk of losing the computational framework we are able to apply here to benchmark performance and closely examine the information produced by actions in a virtual setting.

The fact that we were able to categorize the large majority of participants’ actions into a small set of proto-experiment strategies, and that these were distributed similarly for children and adults, speaks to a broader idea that generic investigative action schemata for probing the physical world coalesce surprisingly early in development. This is an extension of the familiar idea of *learning to learn* (Kemp, Goodman, & Tenenbaum, 2010), but goes beyond the application of domain priors for inference into the application of action priors for learning in familiar domains conditional on particular epistemic goals. It may be that both simulation-based and heuristic camps are half right about human physical reasoning. Identifying latent properties from dynamics *does* depend on comparing simulation-driven expectations against observations, but learning of generic action schemata with established expectations under different properties may go a long way to bootstrap and streamline this process, such that little new simulation needs to be done online (cf. Ludwin-Peery et al., 2021). In this sense, one of the most striking results of our experiment was participants’ flexibility. Both children and adults were able to adapt to a dynamic environment that,

while familiar in some respects, was deeply unreal other respects, lacking basic properties like a third dimension, gravity, and haptic feedback.

In conclusion, our major finding was that real-time active physical learning is mastered surprisingly early in development, with children as young as four interacting with simulated physical objects in ways that are just as goal-specific and informative as the actions of older children and adults. However, there were still important differences between children's and adults' behavior. Adults made more accurate judgments and were more likely to take actions that ostensibly made the dynamics simpler and easier to interpret, such as controlling for the confounding influence of distractor objects by moving them out of the way. This work provides new insight into the developmental roots of the human ability to interrogate the physical world and actively drive learning.

Supplemental Material

Additional supporting information is provided in the Supplementary Information. Data, code and movies of all individual trials can be found in the in the OSF Repository (<https://osf.io/v9fk2/>).

Acknowledgments

Thanks to Laura Ziemann for taking the lead with the data collection and to Paulina Weiss and Maya Buchholz for coding the action data. Thanks to Zachary Horne for guidance on the Bayesian statistics. Neil Bramley was supported by EPSRC New Investigator Grant (EP/T033967/1).

References

- Bacon, F. (1620/1878). *Novum organum*. Clarendon Press.
- Bass, I., Smith, K., Bonawitz, E., & Ullman, T. (2021). Partial mental simulation explains fallacies in physical reasoning.
- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. doi: 10.1073/pnas.1306572110
- Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *195*, 9–38. doi: 10.1016/j.cogpsych.2018.05.001
- Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through interventions. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *41*(3), 708–731. doi: 10.1037/xlm0000061
- Bramley, N. R., & Ruggeri, A. (2022). *Developmental differences in active physical experimentation*. OSF Repository [data, code, movies] https://osf.io/v9fk2/?view_only=e6dcfbbd834b4be282fbe65cc399a013.
- Brewer, W. F., & Samarapungavan, A. (1991). Children's theories vs. scientific theories: Differences in reasoning or differences in knowledge. *Cognition and symbolic processes: Applied and ecological perspectives*, 209–232.
- Coenen, A., Nelson, J., & Gureckis, M., Todd. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*, 1548–1587. doi: 10.3758/s13423-018-1470-5
- Coenen, A., Rehder, R., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133. doi: 10.1016/j.cogpsych.2015.02.004
- Coenen, A., Ruggeri, A., Bramley, N. R., & Gureckis, T. M. (2019). Testing one or multiple: How beliefs about sparsity affect causal experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: 10.1037/xlm0000680
- Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349. doi: 10.1016/j.cognition.2011.03.003
- Datavyu Team, T. (2014). *Datavyu: A video coding tool*. <http://datavyu.org>. Databrary Project, New York University.
- Davis, Z. J., Bramley, N. R., & Rehder, B. (2020). Causal structure learning in continuous

- systems. *Frontiers in Psychology*, *11*, 244. doi: 10.3389/fpsyg.2020.00244
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. doi: 10.3758/BRM.41.4.1149
- Fragkiadaki, K., Agrawal, P., Levine, S., & Malik, J. (2015). Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*.
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, *29*(1), 1–49. doi: 10.1162/NECO_a_00912
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. doi: 10.1126/science.aac6076
- Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2021). A counterfactual simulation model of causal judgment. *Psychological Review*. doi: 10.1037/rev0000281
- Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (Submitted). Active causal structure learning in continuous time.
- Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 75–86. doi: <https://doi.org/10.1002/wcs.1330>
- Gopnik, A., Glymour, C., Sobel, D., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1–31. doi: 10.1037/0033-295X.111.1.3
- Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Wiley Online Library.
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Jeffreys, H. (1961). *The theory of probability*. OUP, Oxford.
- Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–243. doi: 10.1111/j.1551-6709.2010.01128.x
- Kretch, K. S., & Adolph, K. E. (2017). The organization of exploratory behaviors in infant locomotor planning. *Developmental science*, *20*(4), e12421.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573. doi: 10.1037/a0029146
- Kuhn, D., & Brannock, J. (1977). Development of the isolation of variables scheme in experimental and “natural experiment” contexts. *Developmental Psychology*, *13*(1), 9.
- Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities

- and interventions. *Psychological Science*, *16*(9), 678–683. doi: 10.1111/j.1467-9280.2005.01595.x
- Li, S., Sun, Y., Liu, S., Wang, T., Gureckis, T., & Bramley, N. (2019). Active physical inference via reinforcement learning. Proceedings of the Annual Meeting of the Cognitive Science Society.
- Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299. doi: 10.1016/j.cognition.2013.12.010
- Ludwin-Peery, E., Davis, E., Bramley, N. R., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*. doi: 10.1016/j.cogpsych.2021.101396
- Makowski, D., Ben-Shachar, M. S., Chen, S., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *2767*. doi: 10.3389/fpsyg.2019.02767
- McCloskey, M. (1983). Naive theories of motion. *Mental models*, 299–324.
- McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. A. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22. doi: 10.1016/j.jecp.2015.06.017
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.
- Meng, Y., Bramley, N., & Xu, F. (2018). Children's causal interventions combine discrimination and confirmation. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.
- Pearl, J. (2000). *Causality*. New York: Cambridge University Press (2nd edition).
- Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216. doi: 10.1016/j.cognition.2015.07.004
- Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental Psychology*, *52*(12), 2159. doi: 10.1037/dev0000240
- Ruggeri, A., Markant, D. B., Gureckis, T. M., Xu, F., Bretzke, M., & Xu, F. (2019). Memory enhancements from active control of learning emerge across development. *Cognition*, *186*, 1–34. doi: 10.1016/j.cognition.2019.01.010
- Ruggeri, A., Swaboda, N., Sim, Z. L., & Gopnik, A. (2019). Shake it baby, but only when needed: Preschoolers adapt their exploratory strategies to the information structure of the task. *Cognition*, *193*, 104013.

- Schulz, E., Bertram, L., Hofer, M., & Nelson, J. D. (2019). Exploring the space of human exploration using entropy mastermind. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 2762–2768).
- Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045. doi: 10.1037/0012-1649.43.4.1045
- Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, *30*, 50–64.
- Sim, Z. L., & Xu, F. (2017). Learning higher-order generalizations through free play: Evidence from 2- and 3-year-old children. *Developmental Psychology*, *53*(4), 642. doi: 10.1037/dev0000278
- Smith, K. A., de Peres, F., Vul, E., & Tenenbaum, J. B. (2017). Thinking inside the box: Motion prediction in contained spaces uses simulation. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333. doi: 10.1027/1618-3169.56.1.27
- Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94. doi: 10.1126/science.aaa3799
- Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, sep). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. doi: 10.1016/j.tics.2017.05.012
- Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82. doi: 10.1016/j.cogpsych.2017.05.006
- Vicovaro, M. (2018). Causal reports: Context-dependent contributions of intuitive physics and visual impressions of launching. *Acta Psychologica*, *186*, 133–144. doi: 10.1016/j.actpsy.2018.04.015