Hypothesis generation through active inductive inference in children and adults

Neil R. Bramley*
Department of Psychology, University of Edinburgh, Scotland

Fei Xu
Psychology Department, University of California, Berkeley, USA

Author Note

Corresponding author*: neil.bramley@ed.ac.uk.

Developmental data was collected under IRB protocol (Ref No: 2019-10-12687). Adult data was collected under ethical approval granted by the Edinburgh University Psychology Research Ethics Committee (Ref No: 3231819/1). Supplementary material including all data and code is available at `https://github.com/bramleyccslab/computational_constructivism`. This study was not preregistered. Thanks to Jan-Philipp Fränken for help with coding free text responses. This research was supported by an EPSRC New Investigator Grant (EP/T033967/1) to N.R. Bramley and an NSF Award SMA-1640816 to F. Xu.

Abstract

A defining aspect of being human is an ability to reason about the world by generating and adapting ideas and hypotheses. Here we explore how this ability develops by comparing children's and adults' active search and explicit hypothesis generation patterns in a task that mimics the open-ended process of scientific induction. In our experiment, 54 children (aged 8.97 ± 1.11) and 50 adults performed inductive inferences about a series of causal rules through active testing. Children generated substantially more complex guesses about the hidden rules and were more elaborate in their testing behavior. We take a 'computational constructivism' perspective to explaining these patterns, arguing that these inferences are driven by a combination of thinking (recombining and modifying existing symbolic concepts) and exploring (actively investigating and discovering patterns in the physical world). We show how our approach and rich new dataset speak to questions about developmental differences in hypothesis generation, active learning and inductive generalization.

Hypothesis generation through active inductive inference in children and adults

"We think we understand the rules when we become adults but what we really experience is a narrowing of the imagination." —— David Lynch

A central question in the study of both human development and reasoning is how learners come up with new ideas and hypotheses to explain the world around them. Children excel at forming new categories, concepts, and causal theories (Carey, 2009) and by maturity, this coalesces into a capacity for intelligent thought characterized by its domain generality and occasional moments of insight and innovation. Constructivism is an influential perspective in developmental psychology (Carey, 2009; Piaget, 2013; Xu, 2019) and philosophy of science (Fedyk & Xu, 2018; Quine, 1969) that posits learners actively construct new ideas through a mixture of thinking—recombining and modifying ideas—and play—exploring and discovering patterns in the world (Bruner, Jolly, & Sylva, 1976; Piaget & Valsiner, 1930; Xu, 2019). In this way, constructivism views higher level cognition as symbolic and at least somewhat language-like (Fodor, 1975) in its ability to make "infinite use of finite means" (von Humboldt, 1863/1988). While the tenets and promise of constructivist accounts are appealing, it has until recently lacked a formalization capable of producing testable predictions in psychology experiments or detailed insights into human cognition. In this paper we draw on recent methodological advances to formalize a constructivist modeling framework and use it to analyze children and adults' behavior in an open-ended inductive learning task. Our account allows us to capture a wide range of responses and behaviors and closely examine developmental differences in hypothesis generation and active learning. To foreshadow, we argue that both children's hypothesis generation and active learning is based on a "flatter" generation process than adults, producing greater diversity of ideas and behaviors but that is weaker in evaluative precision and systematic coverage of the possibility space.

Concept learning

Classic work in experimental psychology suggests that some form of symbol manipulation is required to synthesize the human capacity for reasoning and problem solving (Bruner, Goodnow, & Austin, 1956; Johnson-Laird, 1983; Wason, 1968). However, early symbolic accounts faced challenges in explaining how such representations could be learned from sparse, noisy evidence, as well as how they could be effectively applied to reasoning under uncertainty (Oaksford & Chater, 2007; Posner & Keele, 1968). Meanwhile, various statistical accounts of concept learning have been developed that model concepts as based in "family resemblance" within a feature space — for instance, centered around a prototypical example or set of exemplars (Kruschke, 1992; Love, Medin, & Gureckis, 2004;

Medin & Schaffer, 1978; Shepard & Chang, 1963). Such accounts have helped explain how people impute discrete categories from finite continuous evidence and how they then use these to generalize effectively to novel stimuli (Shepard, 1987). However, the core representations implied by this approach lack symbolic structure so are not able to explain the human capacity for generating conceptual novelty (Komatsu, 1992).

More recently, Bayesian models have also played a major role in recent study of concept learning, providing a principled way of modeling probabilistic inference over both sub-symbolic and symbolic hypothesis spaces (Howson & Urbach, 2006). On the symbolic side this includes inferences about particular causal structures (Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; Gopnik et al., 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) as well as more general causal theories (Goodman, Ullman, & Tenenbaum, 2011; Griffiths & Tenenbaum, 2009; Kemp & Tenenbaum, 2009; Lucas & Griffiths, 2010). A key strength of the Bayesian framework is its account of how hierarchically structured representations support "learning to learn"—capturing the human capacity to make accurate few-shot inferences about familiar domains (Griffiths & Tenenbaum, 2009; Kemp, Goodman, & Tenenbaum, 2010). However, since Bayesian accounts, by definition, describe learning within a predefined hypothesis space, they are silent about how a learner might explore or generate possibilities within an infinite latent space. For example, causal structure learning accounts using the Causal Bayesian Network framework have typically assumed a fixed and finite—albeit sometimes very large—set of possible structural hypotheses (M. Jones & Love, 2011; Perfors, Tenenbaum, Griffiths, & Xu, 2011).[1] That is, hierarchical Bayesian accounts of induction are generally cast at Marr's computational level (Marr, 1982). They show that people behave roughly *as if* they consider and average exhaustively over a weighted taxonomic hierarchy of possibilities and parameters, despite this being intractable in practice in nontrivial settings.

Alongside Bayesian models, information theory has also featured frequently in cognitive science as a metric of idealized information acquisition (Gureckis & Markant, 2012), including choice of interventions and experiments that reveal causal structure (Bramley, Dayan, Griffiths, & Lagnado, 2017; Bramley et al., 2015; Coenen et al., 2015; Steyvers et al., 2003). However, information-theoretic analyses also presuppose the Bayesian notion that learners have the relevant possibilities in mind and act to discriminate between them, rather than to support the task of constructing or discovering better ones. Thus, while probabilities and information theory provide a jumping off point for top-down analysis of cognition, we should take their limitations seriously when seeking

[1] Although see (Buchanan, Tenenbaum, & Sobel, 2010) for an exception.

to reverse engineer humanlike inferential processing (Van Rooij, Blokpoel, Kwisthout, & Wareham, 2019).

The central goal of this paper is to develop a computational model of constructivist inference and use it as a lens to examine children's and adults' learning in a rich open-ended task where the space of potential hypotheses and behaviors is effectively unbounded. To achieve this, we focus on recent work in cognitive science that has attempted to marry symbolic and statistical perspectives. This work characterizes computational principles driving both human development and intelligence as resting on a capacity to flexibly generate, adapt, combine and re-purpose symbolic representations when learning and reasoning, but crucially doing so in ways that respect probabilistic principles of inference under uncertainty (Bramley, Dayan, et al., 2017; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, 2021; Piantadosi, Tenenbaum, & Goodman, 2016).

Constructivism

Fundamentally, we take the constructivist account to depart from computational-level Bayesian accounts because it presumes representational *incompleteness*. By this, we mean that the constructivist learner has not, and normally could not, consider and weigh all the possibilities in play when learning. Instead, the constructivist is able to *search* the hypothesis space and compare generated hypotheses to one another (A. N. Sanborn & Chater, 2016; Stewart, Chater, & Brown, 2006). For example, this might occur via processes that stochastically combine elements from an underlying concept grammar to produce new variants or to adapt existing hypotheses in the light of new data (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Lewis, Perez, & Tenenbaum, 2014; Nosofsky & Palmeri, 1998; Nosofsky, Palmeri, & McKinley, 1994). Outside of narrow experimental settings, this modal incompleteness seems completely normal. A simple illustration is the characteristic gap between generation of and evaluation of hypotheses or examples (Gettys & Fisher, 1979). People have been shown to spontaneously generate only a few explanations for a surprising event, such as a car engine's failure, and fewer when put under time pressure. However, they will endorse many more if a list of possibilities is provided to them. Similarly, it takes a lot of time and effort to generate a list of category members (such as all listing all fifty US States) while most Americans would find it straightforward to identify the states among a list of state and non-states. Inference about any area of active scientific inquiry, such as is reported in scientific journals like this one, will typically involve an enormous latent space of potential explanatory theories only a fraction of which have ever been articulated or tested. It is generally accepted that the ground truth is unlikely to be among the set of theories already on the table (Box, 1976).

Fortunately, the idea that inductive inference involves symbolic search demystifies a number of behavioral patterns that look like biases from the computational-level perspective. These include order effects, anchoring, probability matching and confirmation bias. Order effects are pervasive in human learning. If new hypotheses are arrived at through a limited local search starting from a previous hypothesis then we should expect patterns of dependency and autocorrelation between a single learner's hypotheses over time (Bramley, Dayan, et al., 2017; Dasgupta, Schulz, & Gershman, 2016; Thaker, Tenenbaum, & Gershman, 2017). Similarly, anchoring effects in estimation can be modeled as the result of limited local adjustment from a salient starting point (Griffiths, Lieder, & Goodman, 2015; Lieder, Griffiths, Huys, & Goodman, 2017). Probability matching is another phenomenon that is natural under a constructivist perspective. In experiments, participants often choose options in proportion to their probability of being correct or optimal rather than reliably selecting the best action, as we might expect if they had the full posterior to hand (Shanks, Tunney, & McCarthy, 2002). However, probability matching is actually often a *best case* scenario for a learner limited to using the the endpoint of a local search as their guess (Bramley, Dayan, et al., 2017). Furthermore, it has been shown that under a variety of plausible everyday utilities, a single-sample–based decision can represent the best computation–accuracy tradeoff to a resource limited learner (Vul, Goodman, Griffiths, & Tenenbaum, 2009). Confirmation bias is also pervasive in human reasoning and active learning (Klayman & Ha, 1989) and hard to explain in purely Bayesian terms. Wason (1960) famously asked participants to test and identify a hidden rule and initially simply told them that the sequence 2–4–6 followed the rule. The intended true rule was simply "ascending numbers" but participants frequently guessed more complex rules such as "numbers increasing by two". Analysis of participants' tests revealed that they frequently generated only tests expected to be rule-following under their hypothesis (such as 6–8–12), so failing to adequately challenge and disconfirm their hypothesis. On a constructivist perspective, learners can only base their exploration on testing the hypotheses they have actually generated. To the extent that certain simple hypotheses like "ascending numbers" were less likely to be generated on the basis of the provided example (cf. Tenenbaum, 1999), it is not surprising that participants failed to exclude these possibilities with their tests.

In the computational cognitive science literature, symbolic search ideas manifest under the label of "learning as program induction". Such models have begun to be applied to synthesizing humanlike problem solving and planning and tool use (Allen, Smith, & Tenenbaum, 2020; Ellis et al., 2020; Lai & Gershman, 2021; Lake, Ullman, Tenenbaum, & Gershman, 2017; Ruis, Andreas, Baroni, Bouchacourt, & Lake, 2020; Rule, Schulz,

Piantadosi, & Tenenbaum, 2018). We build on these in developing our account.

Accounts of Development

The "child as scientist" (Carey, 1985; Gopnik, 1996) — or more recently, "child as hacker" (Rule, Tenenbaum, & Piantadosi, 2020) — perspective casts children's cognition as driven by broadly the same inferential processes as adults' but at an earlier stage in a journey of construction and discovery. While children have been shown to be capable active learners (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; Meng, Bramley, & Xu, 2018; Sobel & Kushnir, 2006), performing interventions or asking questions more informative or relevant to their learning than mere observations or random choices, there is also evidence that children's ability to learn effectively from active learning data is more fragile than adults'. For example, Sobel and Kushnir (2006) found children were much less accurate at causal structure identification in "yoked" conditions — where they had to use evidence generated by someone else to learn — while adults are less effected, sometimes able to learn about as well from others' data as their own (Lagnado & Sloman, 2006). This performance gap has been argued to stem from a mismatch between whatever idiosyncratic hypotheses are under consideration by the observer being different than those being tested by the active learner, making them less able to use the data to progress their theories (Markant & Gureckis, 2014). Relatedly, children's active learning behavior is often more repetitive than adults' (McCormack et al., 2016; Sim & Xu, 2017) and frequently more narrowly focused on testing a single hypothesis at a time (A. Jones, Bramley, Gureckis, & Ruggeri, in revision; Ruggeri & Lombrozo, 2014; Ruggeri, Lombrozo, Griffiths, & Xu, 2016). This might reflect a less developed working memory, restricting the number of hypotheses children can keep track of and compare to evidence. It has recently been shown that children's foraging behavior in large autocorrelated environments is driven more by directed exploration and less by generalization than adults' (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2017). An early emphasis on exploration has been argued to be an effective solution to a lifelong explore–exploit tradeoff, since earlier discoveries can be exploited for longer (Gopnik, 2020). Constructivism also provides a potential explanation for transitions between developmental "stages", characterized by occasional leaps forward in insight. For instance, Piantadosi, Tenenbaum, and Goodman (2012) demonstrate how a program induction account can reproduce a characteristic developmental phase transition in number understanding from grasping a few small digits to understanding the full real number line. Differences between childlike and adultlike behavior might also be captured by parameterizable differences in search, potentially reflecting principles of stochastic optimization (Lucas, Bridgers, Griffiths, & Gopnik, 2014). For instance, young children

have been found to be quick to make broad abductive generalizations from a small number of examples — e.g. readily imputing novel physical laws to explain surprising evidence (L. E. Schulz, Goodman, Tenenbaum, & Jenkins, 2008). Building on this finding, children's hypothesis generation and search has been framed as rationally "higher temperature" than adults' — producing more diversity of ideas at the cost of being noisier (Lucas et al., 2014). This is algorithmically sensible as optimization over high dimensional spaces is known to be more effective when proposals are initially large and decrease over time, as in *simulated annealing* (Van Laarhoven & Aarts, 1987). However, a high diversity of guesses might also simply reflect that children have a flatter latent prior than adults, with construction weights less tuned through experience, inherently producing more diverse hypotheses at the cost of producing high probability ones less frequently. A third possibility is that children's hypothesis generation might be driven more by *bottom-up* processing than adults'. With less established expectations to go on, children's hypotheses might more directly *describe* encountered patterns, while adults might rely more on their existing knowledge hierarchy to generate and constrain hypotheses in a *top-down* way (Clark, 2012). We will contrast children's and adults' hypothesis generation and active learning in a rich task setting that allows us to closely investigate and delineate between these ideas.

Task

In order to study hypothesis generation, we use a rich open-ended task that extends on Wason (1960) and the logical rule-induction tasks studied by Nosofsky et al. (1994), Lewis et al. (2014), Goodman et al. (2008), and Piantadosi et al. (2016). Akin to the blicket-detector paradigm in developmental causal cognition (Gopnik et al., 2004; Lucas et al., 2014), our task has a causal framing, probing inductive inferences about what conditions make an effect occur in a minimally contextualized domain. However, departing from Blicket detector tasks, we include a large and physically rich set of features that learners can draw on in their inferences allowing test scenes to vary in the number, nature and arrangement of objects. Our task is inspired by a tabletop game of scientific induction called "Zendo" (Heath, 2004) and builds on a pilot task examined in (Bramley, Rothe, Tenenbaum, Xu, & Gureckis, 2018). In it, learners both observe and create *scenes*, which are arrangements of 2D triangular objects called *cones* (Figure 1) and test them to see if they produce a causal effect (which arrangements of blocks "make stars come out" in our minimal framing). The goal is to both predict which of a set of new scenes will produce the effect and describe the hidden rule that determines the general set of circumstances produce the effect (try it here). Scenes could contain between 1 and 9 cones. Each cone has two immutable properties: size$\in$ {small, medium, large} and color$\in$ {red, green, blue} and

continuous scene-specific x∈(0,8), y∈(0,6) positions and orientations∈(0,2π). In addition to cones' individual properties, scenes also admit many relational properties arising from the relative features and arrangement of different cones. For instance, subsets of cones might share a feature value (i.e., be the same color, or have the same orientation) or be ordered on another (i.e., be larger than, or above) and pairs of cones might have relational properties like pointing at one another or touching. This results in an extremely rich implicit space of potential concepts.

We note that, by design, the dimensionality of this task makes it very difficult. As with Wason's 2-4-6 example, and genuine questions of scientific induction, the hard part of this task is not evaluating whether a candidate hypothesis can explain the data but rather generating the right hypothesis from among the infinitely many possibilities in the first place. As with the 2-4-6 task, there are always infinite data-consistent possibilities and many of these may be more salient than the ground truth. Without carefully chosen active testing, a learner will frequently fail to rule out simple possibilities that more parsimoniously capture the data than the ground truth, essentially being left with evidence that would not lead even an unbounded Bayesian agent to the correct answer.[2]

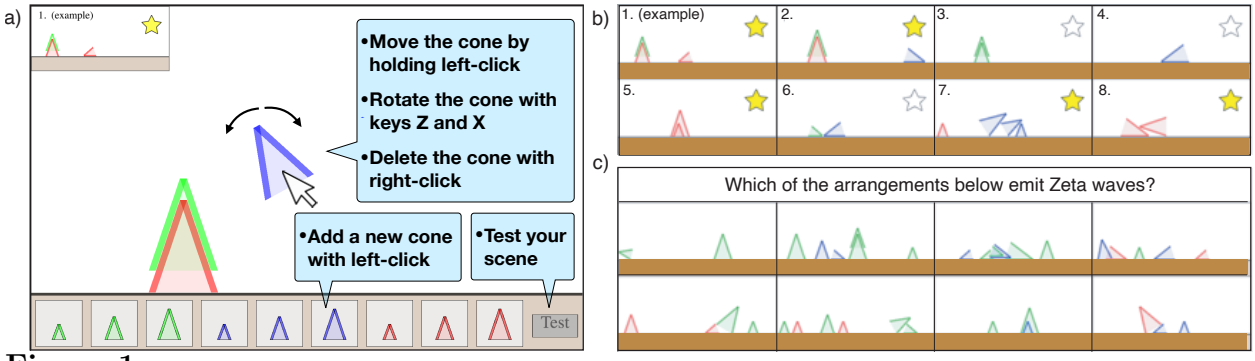


Figure 1

The experimental task: a) Active learning phase. b) An example sequence of 8 tests, the first is provided to all participants, and subsequent tests are constructed by the learner using the interface in (a). Yellow stars indicate those that follow the hidden rule. c) Generalization phase: Participants select which of a set of new scenes are rule following by clicking on them.

We use mixed-methods (Johnson, Onwuegbuzie, & Turner, 2007), analyzing both qualitative data in the form of freely generated guesses about the symbolic rules and

[2] In tabletop game form, Zendo typically takes dozens of rounds of tests and incorrect guesses by multiple guessers, as well as leading examples and clues from the rule-setter for even simple hidden rules to be identified. An online community on Reddit play a binary sequence version of Zendo, often taking hundreds of guesses before the answer is found if it is at all (for example here).

quantitative data in the form of forced choice generalizations. Concretely, we adopt an expressive concept grammar inspired by constructivist ideas in developmental psychology and formalized using program induction ideas from machine learning. We assume the latent space of possible concepts in our task are those expressible in first order logic combined with lambda abstraction and full knowledge of the potentially relevant features of the scene (see Appendix Table A-1 for the grammatical primitives we assume). Table 1 shows the five ground truth rules we used in our experiment expressed in natural language and in lambda calculus along with the initial rule-following example scene we provided to participants.

Given the inherent difficulty of this type of task we expect absolute accuracy to be fairly low for both children and adults (and for our models). However, we expect that many participants will be able to make guesses that are consistent with most if not all of the evidence they have produced. Since we expect the evaluation of evidence–hypothesis consistency to be weaker in children, we expect adults' guesses to be more strictly consistent with their evidence. While naively one might expect children to produce simpler guesses, our framework tends to predict the opposite. With a less trained construction probabilities, our models predict greater diversity and more elaborate structure in children's guesses relative to adults'. While we will not attempt to model participants' scene generation process in detail, we expect similar principles to hold, with children's scene generation being more baroque and elaborate than adults while simultaneously less closely tied to testing their hypotheses. Finally, there is the question of relative dominance of bottom-up and top-down processing in children's and adults guesses. To explore this, we consider two models that vary in this dimension.

Context-free hypothesis generation

In accounting for children's and adults' inferences, we entertain two related constructivist algorithms. The first takes a fully "top down" approach to inference, utilizing a probabilistic context-free grammar (PCFG) to define a latent prior over concepts expressible in first order logic. A PCFG is a collection of "productions" that stochastically build expressions in an underlying grammar (Ginsburg, 1966). A PCFG can be used to generate a prior sample of hypotheses that can then be weighted by their likelihoods of producing observations—here, their ability to reproduce the labels for the scenes that the participant has tested. The model's best guess about the hidden rule is then the *maximum a posteriori* hypothesis in the sample. The hypotheses make predictions about new scenes which can be weighted by their posterior probability and marginalized over to make generalizations. Because parts of this production process and underlying grammar involve branching—e.g., "and" and "or"—hypotheses can become arbitrarily long and complex,

involving multiple Boolean functions and complex relationships between an unlimited number of bound variables. In this way, an infinite latent space (in our case first order logic + lambda abstraction) is covered in the limit of infinite PCFG sampling (see Figure 2a).

The probabilities for each production in a PCFG can be fit to maximize correspondence with human judgments. Different PCFGs, containing different primitives and expansions, can be compared against human behavior. In this way, recent work has attempted to infer the "logical primitives of thought" (Goodman et al., 2008; Piantadosi et al., 2016). In the current work we consider a single expressive PCFG architecture but examine its behavior under uniform production weights but also with weights engineered to produce "childlike' and "adultlike" symbolic guesses. Crucially, under all three weighting schemes, our PCFG embodies the principle of parsimony: Simpler concepts—composed of fewer grammatical parts (Feldman, 2000)—have a higher probability of being produced and so are favored over more complex ones equally able to explain the data.

What PCFG approaches have in common is a generative mechanism for sampling from an infinite latent prior, here over possible logical concepts. However, sampled "guesses" must then be tested against data. Unfortunately, most samples are likely to be inconsistent with whatever data a learner has already encountered.[3] For this reason, the procedure is inherently inefficient, and requires a very large numbers of samples in order to reliably identify non-trivial rules. Thus, we also consider an alternative that provide a more computationally plausible generation mechanism.

Context-based hypothesis generation

Instance Driven Generation (IDG) (Bramley et al., 2018) is a recent proposal related to the PCFG but with one key difference. Rather than generating hypotheses *a priori*, it generates ideas *inspired* by encountered examples (cf. Michalski, 1969), thus blending top-down generation with bottom-up reactivity to evidence. An IDG learner starts by observing the features of objects in a scene and uses these to back out a true logical statement about the scene in a stochastic but truth-preserving way. If the scene is rule following, this statement constitutes a positive hypothesis about the hidden rule. Otherwise, it constitutes a negative hypothesis, i.e. about what must *not* be present. Thus, IDG does not generate uniformly from all possible concepts, but directly from a restricted space consistent with a focal observation. Figure 2b illustrates this approach. While the

[3] Many here are simply tautological (i.e., "All cones are red or not red"), contradictory (i.e., "There is a cone that is red and not red"), physically impossible ("Two distinct objects have the same position") Indeed, around 20% of the hypotheses generated by our PCFGs are tautologies, and 15% are contradictions. Many others combine a meaningful hypothesis with a tautological corollary (i.e., "There is a large red object that is larger than all medium sized objects").

PCFG starts at the outside and works inward, the IDG starts from the central content (drawn from an observation) and works outward out to a quantified statement, ensuring at each step that it is true of the scene. As with the PCFG, we consider a uniform variant as well as variants that include productions reverse engineered to match the summary statistics of guesses generated by children and by adults.

Active learning

Children have long been seen as primarily active learners, using "play" to explore their environment and test their hypotheses (Bruner et al., 1976; Cook, Goodman, & Schulz, 2011; Piaget & Valsiner, 1930). Information theory is used to benchmark active learning (Nelson, 2005; Shannon, 1951) but assumes the learner knows the relevant possibilities and acts to discriminate rather than to constructing or discover better ideas. It has been argued that this may explain why behavioral alignment with information maximization is mixed and task dependent (Coenen, Nelson, & Gureckis, 2018). In parallel to rational accounts, the developmental literature has emphasized the utility of "control of variables" heuristic (Chen & Klahr, 1999; Klahr, Fay, & Dunbar, 1993; Klahr, Zimmerman, & Jirout, 2011) — this amounts to manipulating one design variable relative to a previous test at a time to avoid confounded evidence, such that any changes in the outcome can be unambiguously attributed to the change in the input. While not always the most efficient strategy from a normative perspective (Coenen, Ruggeri, Bramley, & Gureckis, 2019; A. Jones et al., in revision), a control of variables strategy intuitively minimizes the cognitive costs of both testing hypotheses and of coming up with new scenarios (Gershman & Niv, 2010). Past research has only focused on restricted settings with a few simple variables and our task is much more complex. Nevertheless, in exploring the active learning in our task, we will look for the empirical signature of control of variables in the form of incremental and systematic testing patterns.

In sum, the core goal of this work is a close investigation of developmental differences in active open-ended hypothesis generation examined through the lens of a constructivist modeling approach that emphasizes the role of stochastic generation and search. To foreshadow, we find children's and adults' guesses reflect a partially bottom-up process of compositional concept formation. Children create more complex learning data than adults but do so less systematically. They then go on to make more complex guesses about the hidden rule that are only a marginally worse fit to evidence. Our constructivist framework suggests this behavior is a natural result of "flat" idea and action generation mechanisms. Crucially, our modelling then shows that both children's and adults' symbolic guesses causally drive their generalizations, rather than these being driven by similarity

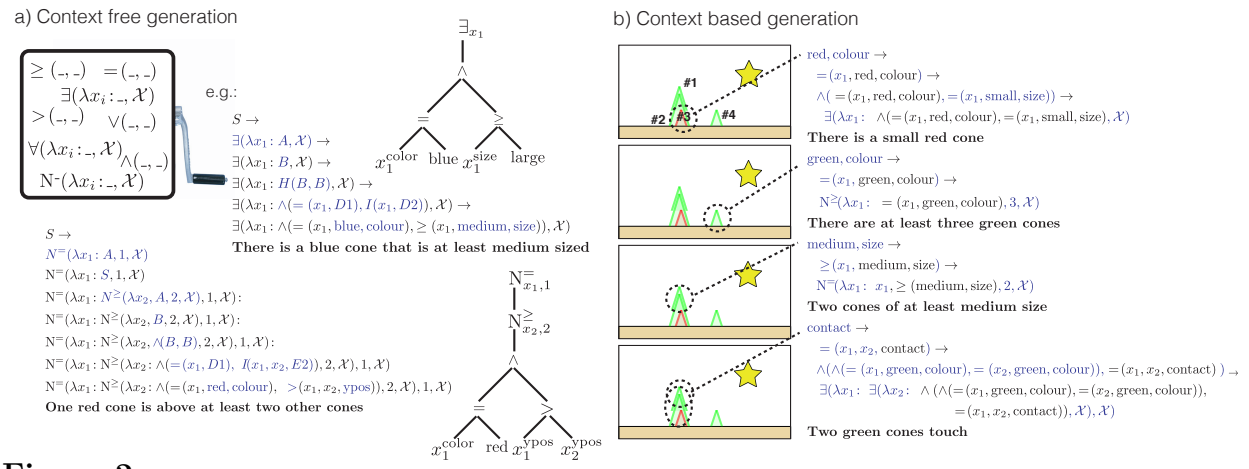


Figure 2

a) Example generation of hypotheses using the PCFG. b) Examples of IDG hypothesis generation based on an observation of a scene that follows the rule. New additions on each line are marked in blue. Full details in Appendix A.

and family resemblance.

Experiment

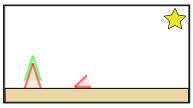
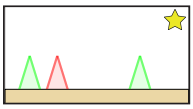
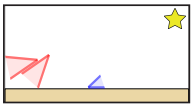
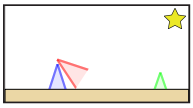
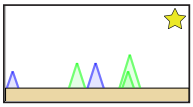
Methods

Participants

We recruited 54 children in the lab (23 female, aged 8.97 ± 1.11) and 50 adults online (22 female, aged 38.6 ± 10.2). Forty children completed all five trials and the remaining 14 completed 2.71 ± 1.07 trials before indicating that they had had enough. For these children we simply include the trials that they completed. We collected participants until we reached our intended sample size of 50 per agegroup after exclusions. We chose this sample size simply to exceed our 2018 (N=30) pilot with adults.[4] Ten additional adult participants completed the task but were excluded before analysis for providing nonsensical or copy-pasted text responses. Adult participants were paid $1.50 and a performance related bonus of up to $4 ($1.96±0.75). Children's sessions lasted between 30 minutes and an hour. For adults, the task took 27.49±12.09 minutes of which 9.8±7.9 was spent on instructions. The children's and adults' versions of the task are available to try here `https://github.com/bramleyccslab/computational_constructivism`.

[4] While we note that 104 is not a large sample by modern standards, our focus is on modeling inferences at the individual level. Each participant produces an exceptionally rich dataset and our analyses have unusually large storage and compute requirements making a larger sample infeasible to analyze.

Table 1
Rules Tested in Experiment

Rule	Initial Example
1. There's a red $\exists(\lambda x_1 : \ =(x_1, \text{red}, \text{color}), \mathcal{X})$	
2. They're all the same size $\forall(\lambda x_1 : \forall(\lambda x_2 : \ =(x_1, x_2, \text{size}), \mathcal{X}), \mathcal{X})$	
3. Nothing is upright $\forall(\lambda x_1 : \ \neg(=(x_1, \text{upright}, \text{orientation})), \mathcal{X})$	
4. There is exactly 1 blue $N^{=}(\lambda x_1 : \ =(x_1, \text{blue}, \text{color}), 1, \mathcal{X})$	
5. There's something blue and small $\exists(\lambda x_1 : \ \wedge(=(x_1, \text{blue}, \text{color}), =(x_1, 1, \text{size}), \mathcal{X})$	

Design

All participants faced the same five learning problems in an independently randomized order (see Table 1). For each learning problem participants were given an initial positive example, as shown in the table, and then performed self tests of their own before making generalizations and free guesses as to the hidden rule.

Materials and Procedure

Child sample.

Instructions. Participants sat in front of a laptop with a mouse attached, with the experimenter sitting next to them and interacted with the task through the browser.

The experimenter read out the instructions for the participant. These explained how the game worked and showed the participant five examples of possible rules the blocks could have (relating to color, size, proximity, angle, or relation). The instructions also included videos showing the participant how to manipulate the blocks using the mouse and keyboard. After the instructions, the participant was given a comprehension check of five true or false questions. If they did not get them all right on their first try, the experimenter read through the instructions again and asked them again. All participants passed the comprehension check the second time.

Learning Phase. The participant was then introduced to an initial example of a block type ("Here are some blocks called [name]s. We're going to click test to see if stars

will come out of the [name]s."). The initial example of each block type (i.e., each rule) was constant across participants. Since every initial example of a block type was a positive example, a star animation played when the "Test" button was clicked. The participant was encouraged to use either the trackpad or the mouse to click the "Test" button, whichever was comfortable for them.

After the initial positive example, the participant was shown a blank scene with blocks available to add to it, and was asked to test the blocks seven more times (Figure 1a). The scene creation interface was subject to simulated gravity, meaning there were physical constraints on how the objects can be arranged. The experimenter told them they could now play with the blocks like they saw in the instructional video. The experimenter also reminded the participant of how to add, remove, move, and rotate blocks on the screen using the mouse and keyboard. Participants were encouraged to ask for help with moving the blocks if needed. If they seemed to be having trouble, the experimenter would ask if they needed help with setting up the blocks. The participants were told that when they had finished moving the blocks around, they should press the "Test" button to see if stars came out of them. For positive tests, the experimenter would neutrally say: "Stars did come out of the [name]s that time" and for negative tests: "Stars did not come out of the [name]s that time."

Question Phase. After testing the blocks a total of eight times (Figure 1b), participants were shown a selection of eight more pre-determined scenes containing blocks (Figure 1c). The experimenter asked them to click on which pictures they thought the stars would come out of, reminding them that they could pick as many as they wanted, but they had to pick at least one. Unknown to participants, half of these scenes were always rule following but their positions on screen were independently counterbalanced.

Free Responses. Participants were then presented with a blank text box and asked, "What do you think the rule is for how the [name]s work?" The experimenter typed into the text box the participant's verbal answer verbatim, or as close as possible.

The Testing, Question, and Free Response phases were repeated identically for each of the five block types. After the five trials were completed, the participant was shown the results including each true rule and how well they did on each problem and was thanked for playing the game. As compensation, participants were allowed to pick a small toy out of a prize box, and parents were given a paper "diploma" to commemorate their child's visit.

Adult sample. We recruited our adult sample from Amazon Mechanical Turk and adults completed the task on their own computers. They completed the same instructions as the children with an additional section about bonuses and had to successfully answer comprehension questions, including an additional two about the

bonuses, before starting the main task. Specifically, adults were bonused 5 cents for each correct generalization (up to a possible 40 cents for each of the five trials) and an additional 40 cents for a correct guess as to the hidden rule, again for each of the five trials. Aside from having no experimenter in the room, and filling out the text fields themselves, the procedure was identical to the children's task. Full materials including experiment demos, data and code are available at the Online Repository.

Results

We first look at the qualitative characteristics of children's and adults' explicit rule guesses then assess relative accuracy of participants' rules and generalizations about new scenes. We compare children's accuracy to adults' and both to our constructivist learning algorithms: Fully top down context-free generation from an expressive latent prior — Probabilistic Context Free Generation (PCFG) — and a partially bottom-up generation — Instance Driven Generation (IDG) (Bramley et al., 2018). We then turn to analysis of the scenes produced by adults and children and finally evaluate a set of formal models' ability to produce both participants' free guesses and their generalizations.

Rule complexity and constituents

We had human coders translate participants' free text guesses about the hidden rule wherever possible into an equivalent logical lambda expression using the grammatical elements available to our learning models. We were able to do this for 86% (n=205) of children's trials and 88% (n=219) of adults' trials. For example, if the participant wrote *"There must be one big red block"* this was converted into $N^{=}(\lambda x_1 : \wedge(=(x_1, \text{large}, \text{size}), =(x_1, \text{red}, \text{color})), 1, \mathcal{X})$. This logical version can be automatically evaluated on the scenes and can be read literally as asserting *"There exists exactly one x_1 in the set of objects $\mathcal{X}$ such that x_1 has the size 'large' and the color 'red'"*. We had a primary coder, blind to the experimental hypotheses code all responses, and a second coder blind spot check 15% of these (64). The two coders agreed in 95% of cases. We provide further details about the coding in Appendix B and full coding resources and full coding data in the Online Repository.

To explore structural differences in children's versus adults' hypotheses, we first break down these encoded rule guesses into their logical parts. This reveals that children's encoded rules were substantially *more complex* than those generated by adults and that both were more complex than the ground truth rules. Children's and adults' rules also differed in terms of the prevalence of particular elements and features (see Figure 3). As an example, one child's rule for problem 1 was *"You must have two reds and one blue"* which

was translated to $N^{=}(\lambda x_1\colon\ N^{=}(\lambda x_2\colon\ (\wedge(=(x_1, \text{red}, \text{color}), =(x_2, \text{blue}, \text{color})), 1, \mathcal{X}), 2, \mathcal{X})$, requiring two quantifiers ($N^{=}$), one boolean ($\wedge$), 2 equalities ($=()$), and two references to the feature color. The typical child-generated-rule used 2.25 quantifiers (3b), 2.06 booleans (3c), 1.55 equalities and inequalities (3d), referred to 1.39 different primary features (color, size, orientation, x- or y-position, groundedness, 3e) and 0.37 relational features (contact, stackedness, pointing, or insideness, 3f). In contrast, the average adult generated rule required just 1.84 quantifiers, 1.20 booleans, 1.47 equalities and inequalities, and referred to 1.44 primary features but only 0.16 relational features. When children posited that an "at least", "at most" or "exactly" a certain number of objects must have certain features, the number they chose were substantially higher than that for adults (2.36 compared to 1.58). In terms of features, adults strongly tended to posit rules relating to color (58% compared to 39% of children's rules), while children were more likely to refer to positional properties (26% compared to 18% of adults' rules) and relations (31% compared to 14% of adults' rules) between the objects.

Reverse engineering Childlike and Adultlike prior productions. Having encoded all the rule guesses from adults and children, we can work back from the distribution of rules to create a sets of productions that produce similar posterior samples. To do this, we work back from the observed counts for each rule element doing this separately for children and for adults. To roughly accommodate the fact that each guess is based on different learning data, we regularized these counts by including a prior pseudo-count of 5 on all productions. This value was not fit to the data, and simply serves to smooth the predictions a little. For example, children's rules involved $\exists$ 263 times, $\forall$ 108 times and N 297 times, so we assumed prior production weights of $\{263 + 5,\ 108 + 5,\ 297 + 5\}/(263 + 108 + 297 + 15) = \{.39, .17, .44\}$. A full set of fitted prior weights for both adults and children are visualized and detailed in Figure 3g&h. Strictly these are samples from a range of different participants' posteriors $P(r|d_{p,t})$ not from their prior $P(r)$, since judgments were always conditional on some evidence. However, since evidence differs greatly across the rules we considered and scenes participants created, and since the structural elements of the grammar (Booleans, Quantifiers etc) are not tightly tied to scene-specifics, we feel this still provides an informative and useful elucidation of differences in a common set of productions that can produce children's and adults' hypotheses. This analysis illustrates that children's construction process is "flatter" than adults' under a constructive account, with a greater average entropy over the various production steps of this process 1.28 ± 0.50bits compared to 1.03 ± 0.59bits, $t(13) = 3.2, p = 0.007$. To avoid double counting the data in modeling subjects' specific guesses, we also created separate agegroup-appropriate prior production weights for each

participant based on the guesses of the participants' from the same agegroup, holding out their own guesses, and used these to produce a unique held-out prior sample for each participant.

Accuracy

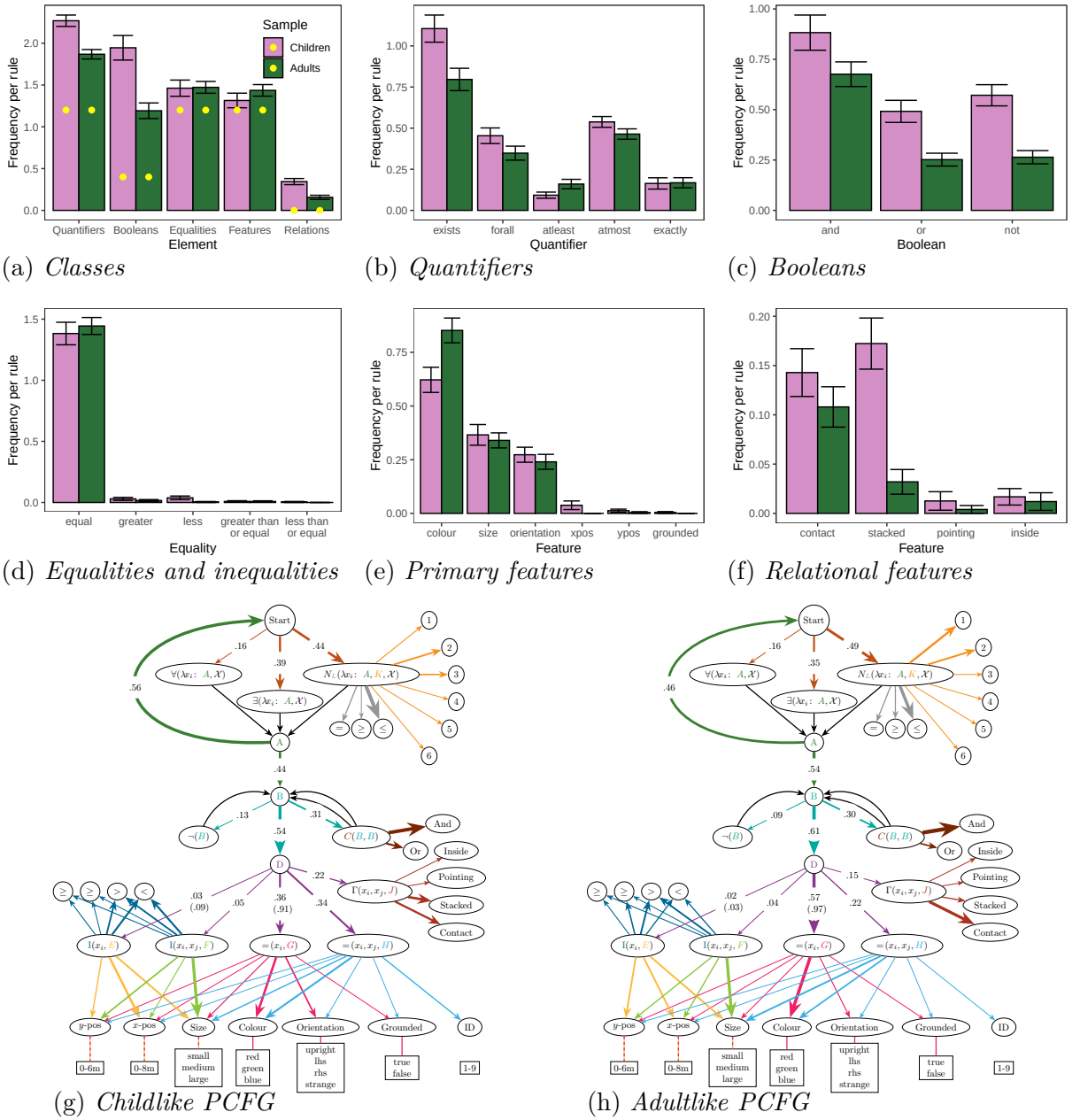
Having observed systematic differences in the content of children's and adults' hypotheses, we now ask if these manifest in children's and adults' inferential success; their ability to identify the ground truth and make accurate generalizations.

Rule guesses. Both children and adults were occasionally able to guess exactly the correct rules, doing so a respective 11% and 28% of trials. Adults produced the correct rule more frequently than children $t(102) = 4.0, p < .001$ and were more likely then children to guess correctly (at a corrected significance level of 0.01) for the "All are the same size", "One is blue" and "There is a small blue" rules (see Figure 4a). The plot reveals that no child identified rule 4 exactly "One is blue" and only one identified rule 5 "There is a small blue", while a slightly greater proportion of children than adults identified the positional "Nothing is upright" rule. Note that chance level baseline for these free guesses is essentially 0%. There are an unlimited number of wrong guesses and a small set of semantically correct guesses. It is also the nature of this inductive problem that there are an infinite number of wrong yet perfectly evidence-consistent rules for any evidence and often there is a simpler evidence-consistent rule available than the ground truth.[5] Thus, it is instructive to ask whether participants' rules, where not exactly correct, are nevertheless consistent with the evidence they have generated.

While, a completely random rule would only be consistent with all 8 scenes around $0.5^8 \times 100 = 0.4\%$ of the time, children's explicit rule guesses were perfectly consistent with the labels of the 8 training scenes 30% of the time and Adult's guesses were fully consistent 54% of the time. There was a moderate difference in average proportion of the learning data explained by children's compared to adults' rules $71\% \pm 27\%$ vs $87\% \pm 17\%$ $t(98) = 5.6, p < .001$. Similarly there was a difference the proportion of the participants' generalizations that were consistent with their rule guess $72\% \pm 21\%$ vs $84\% \pm 16\%$, $t(98) = 4.1, p < .001$ (see Figure 4c for a by-rule breakdown).

We now compare this to simulated *context free* (PCFG) and *context based* (IDG) hypothesis generation algorithms provided with the active learning data generated by the human participants. For each participant p— and separately for each learning task t in the case of the IDG — we generated 10,000 possible rules using *held-out*, age-consistent prior

[5] Although as more evidence arrives the ground truth is increasingly likely to be among "simplest" rules in a posterior sample.



(a) *Classes* (b) *Quantifiers* (c) *Booleans*

(d) *Equalities and inequalities* (e) *Primary features* (f) *Relational features*

(g) *Childlike PCFG* (h) *Adultlike PCFG*

Figure 3

(a–f) Relative frequency of rule elements in Children's and Adults' rule guesses. Yellow points in a show ground truth frequency. (g&h) Visualization of the childlike and adultlike PCFGs (top down rule generation process), reverse engineered to produce rules with empirical frequencies matched to children's and adults' guesses. A rule is produced by following arrows from "Start" according to their probabilities (line weights and annotation), replacing the capital letters with the syntax fragment at the arrow's target and repeating until termination.

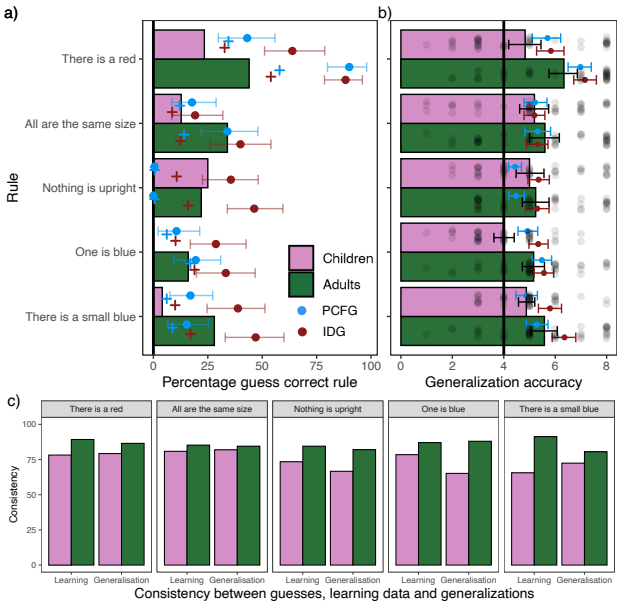


Figure 4

a) Percentage of participants guessing exactly the correct rule. Blue and red circles show performance of PCFG and IDG simulations guessing the maximum a posteriori rule, using age-group consistent prior production weights ± bootstrapped 95% confidence intervals. "+" symbols shows the mean accuracy of a single posterior sample. b) Generalization performance. Bars show mean ± bootstrapped 95% confidence intervals for children (pink) and adults (green). Model simulations select the maximum a posteriori *label after marginalizing over the posterior. Black vertical lines denote chance performance. c) Consistency between subjects' rule guess and their (self-generated) learning data, and generalization judgments.*

productions derived above $\hat{R}^{p}_{\text{PCFGh}}$ and $\hat{R}^{p,t}_{\text{IDGh}}$ that have statistics matched to those in Figure 3a–f. We also generated an additional sample of 10,000 rules based on uniform production weights $\hat{R}_{\text{PCFGu}}$, and similarly for the IDG generated a sample based on uniform productions for each task $\hat{R}^{p,t}_{\text{IDGu}}$. The PCFG samples act as an approximation to the infinite latent prior over rules $P(r)$ before seeing any data, while the IDG samples are always generated conditional on one of the learning scenes. We split the IDG sample evenly across these such that 1250 were "inspired" by each learning scene, necessarily repeating this procedure for each trial for each participant since each generates different evidence. In order to approximate a posterior over rules given self-generated learning scenes $\mathbf{d}$, we then weight these samples by their likelihood of producing all eight scene

labels observed during the learning phase

$$P(r|\mathbf{d}) \propto P(\mathbf{d}|r)P(r) \tag{1}$$

$$\approx P(\mathbf{d}|r) \sum_{\hat{r} \in \hat{R}} \mathbb{I}(r = \mathrm{r}) \tag{2}$$

and combine this with their prior weight given by counting how often they appear in the prior sample, with indicator function $\mathbb{I}(.)$ denoting exact or semantic equivalence. To test for semantic equivalence, we computed predictions for the first 500 participant-generated scenes for each rule and clustered together those that made identical predictions. We rounded positional features to one decimal place in evaluating rules to accommodate perceptual uncertainty. Concretely, we assumed the following likelihood function

$$P(\mathbf{d}|r) = \exp(-b \times N_{\text{outliers}}) \tag{3}$$

embodying the idea that: the more learning scene labels a rule cannot explain, the less likely it is to have produced them. For a large b, the likelihood function approaches the true deterministic behavior of the rules. However, in our analyses we simply assume a $b = 2$ to allow for some noise while maintaining computational tractability. This corresponds to a likelihood function that decays rapidly from 1 for rules that predict all 8 scenes' labels to .13 for a single misprediction, and .02 for 2 mispredictions and so on.

To generate IDG predictions, we merged the production probabilities from the PCFG into the Instance Driven Generation procedure detailed in the Appendix A. For scenes that did not follow the rule we followed the same procedure as for scenes that did, but wrapped the rule in a negation. For example, observing a non-rule-following scene in which there are objects in contact might inspire the rule that "no cones are touching".

Model rule guess accuracies are detailed in Table 2. Taking the *maximum a posteriori* (MAP) estimate (guessing in the event of ties) under all models leads to guessing the correct hypothesis with slightly higher frequency than participants (15–37% based on children's active learning and 20–51% based on adults'). For instance, under a uniform-weighted prior sample, the PCFG MAP is correct on 15%±35% of children's trials and 20%±40% of adults' trials. Note that since these simulations use the same prior sample, the small differences we see are due to the different learning data generated by children and adults. However, accuracy improves substantially and better reproduces the empirical child–adult accuracy difference if we use samples based on reverse engineered weights that reproduce the qualitative properties of children's and adults' rules (see Appendix A and Figures 3g&h). For age-appropriate prior samples, the PCFG guesses

correctly on 18%±38% of children's trials and 32%±46% of adults' trials. Using an age-inappropriate "flipped" prior sample (i.e. child-like samples for adults and adult-like samples for children's tasks) obliterates this difference 23%±42% for children's trials and 22%±41% for adults' trials. We see a similar pattern for the IDG algorithm, but higher accuracy across the board. The IDG has the better accuracy on both children's and adults' trials, guessing over half of the hidden rules correctly (51%±50%) in the case of adults' trials. However, achieving this accuracy requires maximizing over the full sample, while we have argued that process level accounts are more likely to yield behavior closer to posterior sampling (Table 2, right hand columns). Indeed posterior samples provide a visually closer fit to the by-rule guess rates (Figure 4a). To check what provides the best account of participants accuracy patterns we fit logistic mixed-effect regression model using each algorithm × prior combination to predict participants' by-task probability of guessing correctly, including random effects for both rule type and participant. The "Fit" columns of Table 2 shows the log likelihood for each of these models, revealing that participants' correct judgments were best predicted by posterior sampling under Instance Driven Generation with an age-appropriate prior (log likelihood = 211.5, $\beta = 5.44 \pm 1.74, Z = 5.99, p < .001$) improving over a baseline of -234.3 for a model with only intercept and random effects.

Table 2
Accuracy of Rule Guesses by Simulation Models

		Acc Max Posterior (%)			Acc Sample Posterior (%)		
Algorithm	Prior	Children	Adults	Fit	Children	Adults	Fit
PCFG	Uniform	15±35	20±40	-231	9±12	12±14	-227
PCFG	Agegroup	18±38	32±46	-233	12±19	20±25	-228
PCFG	Flipped	23±42	22±41	-235	16±21	15±22	-231
IDG	Uniform	27±44	39±48	-228	9±12	14±5	-218
IDG	**Agegroup**	37±48	51±50	-229	**14 ± 16**	**24 ± 22**	**-216**
IDG	Flipped	26±44	52±50	-234	14±20	23±22	-227

"Children" and "Adults" columns show the $M \pm SD\%$ correct accuracy of the requisite algorithm based on the learning data from that agegroup. "Fit" shows the log likelihood for a logistic mixed-effects regression using this model to predict participants' correct responses.

Generalizations. We now use the models to account for the quantitative response data constituted by forced choice generalizations about which of 8 new scenes will produce stars (i.e. follow the hidden rule). Across the five tasks, both children and adults guessed more accurately than chance (50%): *children* mean±SD $59\% \pm 11\%, t(53) = 5.9, p < .001$; *adults* $70\% \pm 14\%, t(49) = 10.3, p < .001$. Adults' generalizations were significantly more

accurate than children's $t(102) = 4.6, p < .001$ and children's accuracy improved significantly with age $F(1, 52) = 6.2, \eta^2 = .11, p = 0.015$. Indeed, adults' generalization accuracy was above a Bonferroni-corrected chance level of $p \leq 0.01$ for all five rules and children were similarly above chance except for rules 1. "There is a red" ($t(46) = 2.5, p = .015$) and 4. "One is blue" ($t(46) = .1, p = .915$; see Figure 4b).

We compare this pattern against simulated constructivist PCFG and IDG learner benchmarks. To do this, we use the requisite predictive distribution to model generalizations to the set of test scenes $\mathbf{d}^*$

$$P(\mathbf{d}^*|\mathbf{d}) = \int_R P(\mathbf{d}^*|R)P(R|\mathbf{d})\ dR \tag{4}$$

$$\approx \sum_{r \in \hat{R}} P(\mathbf{d}^*|r)P(r|\mathbf{d}) \tag{5}$$

Provided with the active learning data generated by the human participants, both performed in the human range at generalization. Using uniform production weights and taking the marginally most likely generalization labels over a posterior weighted sample of PCFG-generated rules based on the participants active learning data yielded accuracies of 61.4%±19.6% for children's and 63.5%±20% for adults' data. Using agegroup-appropriate priors, PCFG model performs a little better and reproduces the empirical difference between children's and adults' accuracy: 62.8±19.8% for children's and 68.8±20.9% for adults' PCFG weights. The uniform IDG, again, performed slightly better than the PCFG, generalizing at 65.2%±19.3% from children's active learning data and 69.0%±21.0% from adults'. Again using agegroup-appropriate prior productions, the IDG models' performance increased slightly for both adults' and children's data and better reproduced the difference between children's and adults' accuracy (68.8±20.1% and 74.2%±21.7%).

The stronger generalizations of the IDG compared to the PCFG replicates the findings of Bramley et al. (2018) and extends this to children as well as adults. Intuitively, this is because the bottom-up mechanism ties the hypotheses generated to features of the learning cases, effectively narrowing in on plausible hypotheses more efficiently. More broadly, these simulation results underscore the inherent difficulty of this task in particular and open-ended inductive inference in general. The PCFG and IDG were not statistically better or worse than participants at any rule inference under after Bonferroni correction with the exception that the IDG outperformed children on rule 4 $t(96) = 4.7, p < .0001$. Thus strikingly, even in this "small world" with known and fully observed features, and even allowing simulations to sample and maximize over implausibly large numbers of hypotheses, we could not robustly outperform human adults in this task. Additionally, we saw that building in human inductive biases boosted performance and that adults' stronger

inductive biases go some way to explain differences in generalizations.

Interim discussion

Children were only moderately less able to guess rules that fit the evidence than adults and there were only moderate differences in the compatibility between children's and adults' rules and their generalizations. However, children appeared to overfit the evidence more, essentially producing more complex and naïve characterizations of the rule-following scenes than did adults. This can be seen in the larger number of quantifiers and relations mentioned in children's rules than in adults', essentially referring to complex properties of the learning scenes that were irrelevant to their label.

A complicating factor is that children generated different learning data to adults. However, our PCFG and IDG simulations suggest exposure to different data cannot explain most of the accuracy differences between children and adults. Using identical production weights and the scenes generated by adults and children led to only small differences in accuracy for the PCFG and moderate for the IDG, while using a "flatter" set of productions fit to match childlike rules, and a more "peaked" set fit to adults' rules, better reproduces the accuracy differences. We take this to suggest hypothesis construction differences drive a large portion of the differences in children's and adult's inductive inferences.

We now turn to analyze active learning (scene generation) behavior. We first characterize the differences between the scenes generated by children and by adults and then ask whether these can be attributed to differences in hypothesis generation.

Search behavior

As well as generating more complex rules, children also tended to create more complex scenes than adults during the learning phase. The average child-generated scene contained 3.7±0.88 objects (about the same as were present on average in the initial examples) compared to 2.8±0.57 objects for adults ($t(102) = 5.8, p < .001$). The complexity of test scenes was inversely related to performance overall ($F(1, 102) = 39.0, \beta = -0.08, \eta^2 = .28, p < .001$) and also within both the child sample ($F(1, 52) =, \beta = -0.056, \eta^2 = .20, p < .001$) and adult sample ($F(1, 49) = 9.1, \beta = -0.096, \eta^2 = .16, p < .001$) taken individually (see Figure 5a). Within the child sample, age was inversely associated with scene complexity with an average of 0.35 fewer objects per scene for each additional year of age $F(1, 52) = 12.6, \eta^2 = .19, p < .001$. Aside from this difference, we can also assess whether children's or adults' scenes bear the hallmarks of a local "search" across possible scene dimensions.

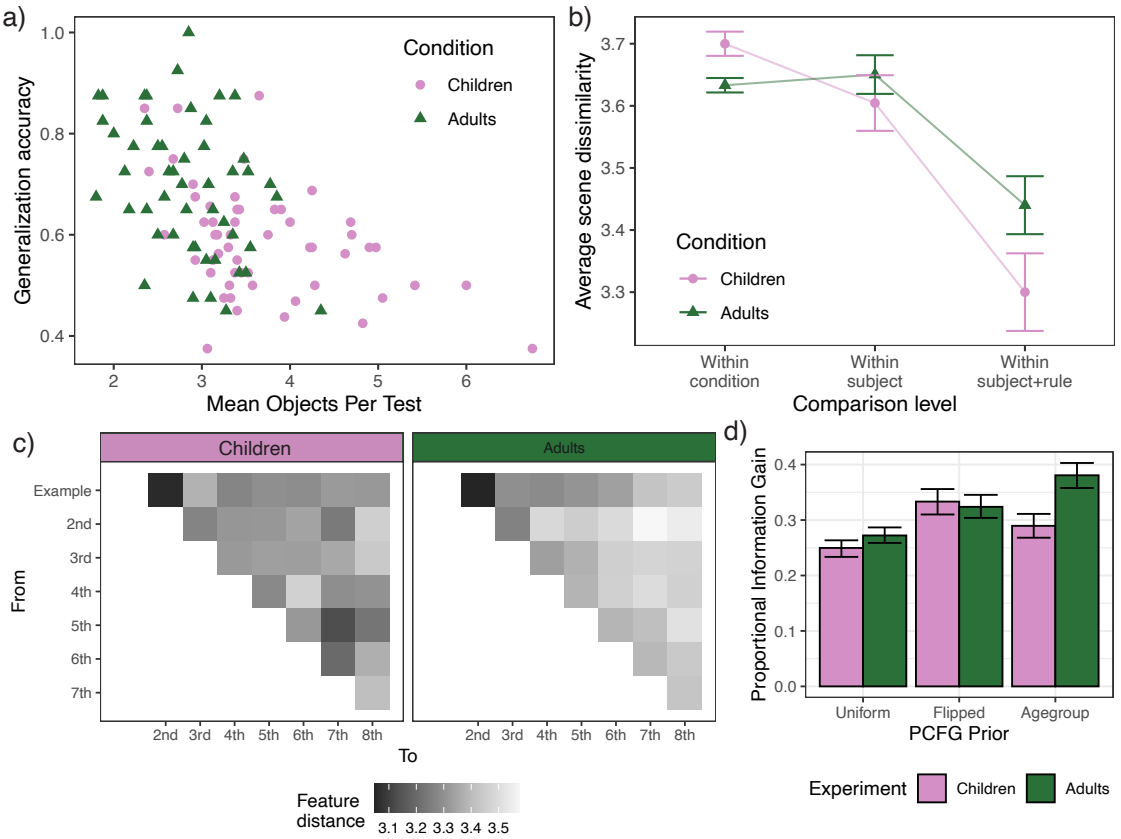


Figure 5

(a) Generalization accuracy by number of objects per test scene. (b) Average dissimilarity between self-generated scenes at different levels of aggregation. Error bars show standard errors for subject means. (c) Average similarity matrices between initial example and self generated scenes 2 to 8. See Appendix C for detailed procedure and similarity matrices separated by component. d) Information gain under different priors.

Scene sequences and similarity. While we do not yet have a model of scene creation process, we hypothesized that *control of variables* (Kuhn & Brannock, 1977) is a reasonable marker of systematic active learning. In the current setting, this manifests as a tendency to generate new evidence that comes close to recreating a previous scene (i.e. whose labels is already known) but making some change to it. This allows a learner to isolate boundary conditions for membership of the rule-following or non-rule-following scenes, and so potentially fine tune a focal hypothesis, or choose among a small set of similar alternatives (Bramley, Dayan, et al., 2017). Additionally, we speculated that reuse in general is likely to reduce cognitive load (Gershman & Niv, 2010).

If this is the case, we should expect the scenes generated by participants to be more similar to the initial example than to a random scene or a scene drawn from a different learning problem. To explore this, we constructed a distance metric that we used to

measure the featural dissimilarity between any pair of scenes. The metric is based on edit distance, encoding how much and how many of the features (positions, colors, shapes) of the objects in one scene would have to be changed to reproduce the other scene. This involved z-scoring and combining a "minimal-edit set" of feature differences and incorporating a proportional cost for additional or omitted objects and scaling by the number of objects in the scenes. We provide a detailed procedure and example of how we computed these edit distances and break them down into their separate components in the Appendix C. The mean distance between any randomly selected pair of participant-generated scenes was M±SD= 3.67 ± 0.94. Taken as a whole, the scenes generated by children were more diverse than adults' with average dissimilarity of 3.70 ± 0.14 compared to 3.63 ± 0.08, $t(102) = 2.9, p = 0.0048$.

However, this diversity seems to be primarily *between* rather than *within* subject for children's choices. Within subject but across trials, the average inter-scene dissimilarity for children was $3.60 \pm .33$ similar to that for adults' $3.65 \pm .22$, $t(102) = .83, p = .4$. Focusing more narrowly, within the scenes produced by an individual subject while learning about a single rule, we see a reversal of the aggregate pattern. That is, within a learning task, children's scenes are marginally *less* diverse on average than adults' (children: 3.30±0.459, adults: 3.44±0.33, $t(102) = 1.77, p = 0.08$, Figure 5b&c).

Figure 5c breaks down the within-trial scene dissimilarity by test position for the two age groups. Adults' scenes are clearly anchored to the initial example (right hand facet) — shown by the dark shading in the top row indicating high similarity decreasing from left to right for later tests — Adults' scenes are also substantially sequentially self-similar — shown by the relatively darker shading along the diagonal compared to the off-diagonal. In contrast, children's similarity patterns look more uniform. However, for both adults and children, the first self-generated scene is more similar to the Initial example than any other scene.

Children's and adults' scene generation patterns manifest in differences in the quality of the total evidence generated according to an information gain analysis. For example, using the unweighted PCFG sample, prior entropy is 7.74 bits and children's evidence produces an information gain (reduction in uncertainty) of 1.93±0.45 bits while adults' data average an information gain of 2.11±0.38 bits $t(102) = 2.12, p = 0.035$ (see Figure 5d). Relative to the fitted PCFG priors, the difference in information gains is rather larger, with children's scenes leading to information gain at 2.28±0.66 bits (prior entropy 7.87±0.05), and adults' at 2.96±0.64 (prior entropy 7.77±0.04) $t(102) = 5.3, p < .0001$. Under the flipped priors — that is, taking the adultlike PCFG prior for children and childlike PCFG prior for adults — children's tests look more informative than under their

own prior, generating 2.58±0.68 bits, and adults' tests slightly less informative than under their own prior 2.55±0.57 bits, eliminating the statistical difference $t(102) = 0.24, p = 0.81$. On the face of it, this is evidence against the idea that children's more elaborate hypothesis generation and concomitantly flatter construction weights are driving them rationally toward more elaborate testing choices. However, we see this information-theoretic analyses as limited in what reveals. This is because is predicated on an implausibly complete representation of uncertainty that we approximated by using a large sample of prior hypotheses, while we might expect constructivist search behavior to be driven by more focal testing of a smaller number of possibilities. Nevertheless, we present these information scores as norms for completeness and comparison with other active learning tasks.

Experiment Discussion

Our analysis through the constructivist lens suggests we cannot attribute the differences in children and adults' hypothesis generation to their differences in active learning, nor can we attribute their differences in active learning behavior directly to differences in hypothesis generation. That is, assuming the same hypothesis generation process for children and adults' data does not reproduce the differences in rule guesses and accuracy. At the same time, assuming children and adults' scene creation is driven by distinguishing among *different* sets of hypotheses, sampled from a childlike or adultlike latent prior, does not explain the developmental differences in the complexity and systematicity of the scene creation. Rather, these data support the idea that developmental shift in hypothesis generation and active search are manifestations of a gradual sharpening of the constructivist generative mechanisms that produce relevant ideas but also novelty in cognition.

We finally turn to a model-based analysis of the free responses and generalizations. To foreshadow, we find both children's and adults' guesses are better accounted for by our partially bottom-up IDG account of hypothesis generation than by our fully top-down PCFG norm. We then find that both children's and adults' generalizations cannot be explained by a non-symbolic "family resemblance" model but are well predicted by their individual symbolic rule guesses.

Model comparison

Guesses

To evaluate our constructivist PCFG and IDG models' ability to explain participants' free response guesses, we estimated probability of each approach generating

exactly the participant's encoded guess based on their active learning data.

By definition, all 87% of participants' rules that we were able to encode in our concept grammar have nonzero support under a PCFG prior, and due to the stochasticity we assumed in our likelihood function, all possibilities should also nonzero have posterior probability. However, in practice it is impossible to cover an infinite space of discrete possibilities with a finite set of samples, meaning there are a substantial number of cases in which we did not generate the participants' guess. The proportion of Children's and Adults' rules that were generated at least once in 10,000 samples was 45% for children's and 50% for adults' rules from the uniform prior and 42% for children's and 53% for adults' guesses from the age-appropriate prior. The IDG samples included children's rules 49% of the time and adults' 62% of the time using uniform weights and using fitted weights this increased to 69% of children's and 76% of adults' rules.

Table 3
Models Chance of Producing Participants' Rule Guesses

		Children		Adults	
Algorithm	Prior	Mean (%)	N best	Mean (%)	N best
PCFG	Uniform	3.3 ± 5.0	13	7.2 ± 7.2	10
PCFG	Agegroup	4.3 ± 7.4	13	12.5 ± 12.0	15
IDG	Uniform	3.4 ± 5.1	10	8.7 ± 8.6	2
IDG	**Agegroup**	$\mathbf{4.5 \pm 7.1}$	**15**	$\mathbf{14.1 \pm 13.6}$	**22**

Table 3 details model fits to participants' guesses. This suggests the IDG is the stronger hypothesis generation candidate, assigning higher probabilities on average to the rules that participants guessed. As expected, the variants of the PCFG and IDG with fitted production weights are better aligned with participants' guesses than variants with uniform (or mismatched) weights. However, all models produced adults' guesses with a much higher probability than children's guesses.

Figure 6a additionally visualizes participants guesses in terms of their posterior probability under PCFG and IDG sampling and compares this to what we would expect if guesses are samples from the posterior (black line), the result of finding the maximum a posteriori guess of the 10,000 considered hypotheses (dashed line) or else are simply samples from the prior (dotted line). This visualization shows that, under all the models we consider, adults' guesses are distributionally more consistent with posterior sampling than posterior maximization, while children's appear somewhere between prior and posterior sampling.

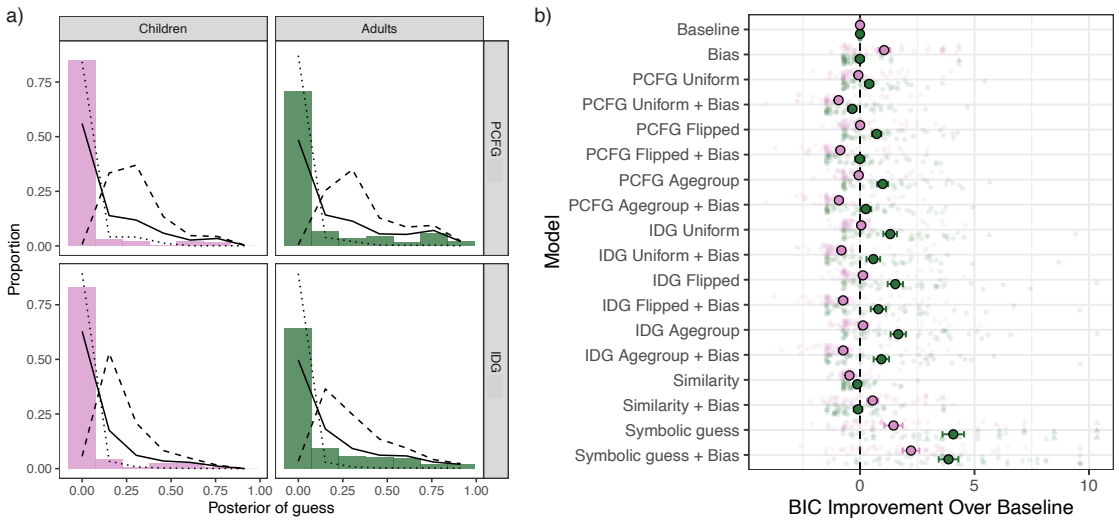


Figure 6

a) Posterior probability of participants' guesses under PCFG and IDG samples. Full black line compares with posterior samples, dashed line with selection of the posterior maximum a posteriori hypothesis (or sampling from them if there are more than one), dotted line compares with samples from the prior. b) Individual generalization model fits showing BIC improvement over baseline per trial (higher is better). Opaque points show mean±SE, faint points show individual fits, with triangles used to mark where the model is the best fit (of all 18 tested) for that participant.

Generalizations

A standard benchmark for models of concept learning is a fit with participants' generalizations to new exemplars. Thus, we complete our analyses by comparing a range of models' ability to account for participant's specific generalizations. The set of models we consider allows us to test our core claims that children's and adults' induced representations are symbolic and compositional, as opposed to statistical and similarity-driven.

We fit a total of 18 models to the data. All models have between 0 and 2 parameters. For each model, we fit the parameter(s) by maximizing the model's likelihood of producing the participant data, using R's `optim` function. We compare models using the Bayesian Information Criterion (Schwarz, 1978) to accommodate their different numbers of fitted parameters.

The models we fit were:

1. **Baseline**. Simply assigns a likelihood of .5 to each generalization ∈ {rule following, not rule following} for each of the 8 generalization probes for each of the 5 learning trials.

2. **Bias**. Acts a stronger baseline by allowing participants to have an overall bias

toward or against selecting generalization scenes as rule following. For this model, b = 1 if >50% of generalizations predict the scene is rule following and 0 otherwise. The model is fit using a mixture parameter λ to mix this modal prediction with the baseline prediction of .5 $P(\text{choice}) = \lambda b + (1 - \lambda).5$.

3-8. PCFG {Uniform, Flipped, Agegroup} {No Bias, Bias}. These models base their generalizations on the marginal likelihood that each generalization scene is rule following under the Probabilistic Context Free Generation (PCFG) posterior $r = P_{\text{PCFG}}(\mathbf{d} * |\mathbf{d})$. "Uniform" uses a prior with uniform production weights. "Flipped" uses a prior generated with mismatched weights —- that is, adultlike weights for children's generalizations and childlike weights for adults' generalizations. "Agegroup" uses a sample based on weights derived from other participants in the same agegroup holding out the participants' own guesses. In each case, these predictions are then softmaxed using $P(\text{choice}) = \frac{e^{r/\tau}}{\sum_{r \in R} e^{r/\tau}}$,with temperature parameter $\tau \in (0, \infty)$ (Luce, 1959) optimized to maximize model likelihood. Large positive τ indicates random selection. $\tau \to 0$ indicates hard maximization Variants with including a bias term also mix this prediction with the subject's modal response b as in

$$P(\text{choice}) = \lambda b + (1 - \lambda)\frac{e^{r/\tau}}{\sum_{r \in R} e^{r/\tau}}. \tag{6}$$

9-14. IDG {Uniform, Flipped, Agegroup} {No Bias, Bias}. These models use the marginal likelihood of each generalization scene as rule following under the Instance Driven Generation based posteriors with variants as with the PCFG variants and again fit with softmax parameter $\tau \in (0, \infty)$.

15-16. Similarity {No Bias, Bias}. Inspired by Tversky's statistical and similarity based *contrast model of categorization* (cf., Tversky, 1977), we used the inter-scene similarity between each generalization scene and each training scene to compute the relative average similarity of each generalization case to the rule-following vs. the not rule-following training scenes. Similarities were computed using the same procedure used in the Active Learning section of the Results and detailed in Appendix C. We computed the mean difference between rule-following and not-rule following similarities as a $\Delta Similarity$ score for each participant×trial×item combination. Positive scores mean generalization item has a greater feature similarity to the rule following learning scenes than the not rule-following learning scenes. Negative scores mean the reverse. To convert these into choice probabilities, we take the inverse logit of these scores $r = \frac{\Delta\text{Similarity}}{\Delta\text{Similarity}+1}$ and

again fit these r values to maximize the likelihood of participants' choices using a softmax function with inverse temperature parameter $\tau \in (0, \infty)$. Intuitively, this model provides a non-symbolic alternative account of generalization behavior.

17-18. Symbolic Guess {No Bias, Bias}. This model takes participants' free guess of the hidden rule, coded in lambda abstraction, and uses these directly to generate a prediction vector $r \in R$:{*rule-following*=1, *not rule-following*=0} for each scene. For trials in which the participant does not provide an unambiguous rule, the model assigns a .5 likelihood to each generalization choice. These were again fit with a softmax parameter $\tau \in (0, \infty)$.

A good fit for *Symbolic Guess* would support our core claim that participants inductive generalizations are directly driven by their constructed symbolic ideas. Meanwhile, a better fit for *Similarity* would suggest that generalizations are rather based on sub-symbolic feature similarity, with participants guesses relegated to a supporting role as rough symbolic re-descriptions of an ultimately sub-symbolic representation (e.g., Dennett, 1991; Johansson, Hall, & Sikström, 2008). To the extent that our constructivist simulations reflect participants' inductive inference mechanisms we expect the end-to-end PFG and IDG mdoels to also fit generalization patterns even though they are blind to participants explicit guesses. This also acts as a sanity check for our approach for any readers skeptical about the validity of self-report data.

We fit all models to the children's and adults' data, and then separately to each individual participant. The full table of model fits is presented in the Appendix (Table A-3). Individual level results are highlighted in Figure 6b.

Results

In line with our core hypothesis, *Symbolic guess + Bias* is the best fitting model of both children's and adults' generalizations outperforming all the models we considered based just on only the learning data. For children's generalizations taken together, *Symbolic guess + Bias* has BIC 2149, improving 490 over Baseline with bias term mixture weight of $\lambda = .26$ and choice temperature parameter $\tau = 0.80$. For adults, this is BIC 1776 with a larger BIC improvement of 996 over Baseline, with a $\lambda = 0.08$ indicating less bias and temperature $\tau = 0.50$ indicating tighter alignment with the guessed-rule's predictions. Probing this bias, we see children undergeneralized substantially on average, selecting just $2.75 \pm 1.42/8$ scenes compared to adults' $3.42 \pm 1.03/8$ (unknown to the participants, there were always 4 rule following generalization scenes). Focusing on individual fits, the picture is mixed for children's generalizations, with 16/50 best fit by the *Bias* only model, followed

by 15 by the *Symbolic guess* model, 9 by the *Symbolic Guess + Bias* model and a further 7 by the fully random *Baseline*. No other model best fit more than 2 children. For adults, 32/52 were best fit by *Symbolic guess*, 6 by *Bias*, 4 by *Symbolic guess + Bias* and no other model best fit more than 2 participants. Overall, children's generalizations are much harder to predict than adults' with end-to-end constructivist accounts of their generalizations performing close to *Baseline*. This is partly to be expected since our child-like construction weights inherently produce a very diverse set of guesses and correspondingly diffuse set of generalization predictions. However, conditioning on Children's symbolic guesses we were able to predict their generalizations far better than by *Similarity*, *Bias* or any other model we considered. Adults' generalizations seem more straightforwardly driven by their symbolic guesses, with better individual fits on average using their guess directly without adjusting by any bias toward or against predicting scenes to be rule-following. This makes sense since, with a clear hypothesis in mind, there is little rationale to select more or fewer than than the generalization scenes consistent with that rule.

Thus, confirming our key constructivist hypothesis, there is a clear alignment between participant's symbolic rule guesses and their generalizations. Thus, these results are consistent with our computational constructivism framework, with generalization behavior best accounted for as driven by the logical structure symbolic hypotheses rather than family resemblance. Our account explains why inference patterns in this setting are highly idiosyncratic to different participants yet explicable as the result of a process of compositional of construction and search. As with the free rule guesses, the IDG was better aligned with participants' generalizations than the PCFG, particularly for adults, and particularly when using Agegroup-specific rather than Uniform or Flipped production weights. Thus, our model fitting also supports the idea that participants were inspired by patterns present in the learning data, such as the objects and relations in the initial positive example. However, this does not appear to be a developmental difference per se, with both children's and adults' judgments better accounted for by our IDG than our PCFG algorithm across all of our analyses.

Discussion

We explored children and adults' active hypothesis generation and inductive inference in an interactive task where the space of possibilities and actions is compositional, open and practically unbounded. Our results are rich and nuanced but broadly we found that:

1. A constructivist framework could explain the diversity and distribution of people's symbolic hypotheses in our inductive learning task, synthesizing both their sporadic

correct guesses but also their incorrect ideas and also offering an explanatory framework for exploring differences between children's and adults' generation mechanisms.

2. Children's guesses and active testing reflect "flatter" generation mechanisms than adults', producing more diversity but less predictability.

3. The logical form of both children and adults' symbolic guesses predicts their generalizations to new scenes far better than feature similarity.

4. We found evidence of probability matching, with the frequency adults made particular guesses more similar to posterior probability than the pattern expected under noisy selection of the most probable hypothesis.

5. Both children and adults' hypothesis generation was context sensitive, as shown by better fit throughout our analyses for a partially bottom-up Instance Driven generation account — sensitive to patterns in the learning data — over a fully top-down Context Free (PCFG) generation account — aware only of the relevant features and possible values.

We now discuss these results more broadly, first highlighting some limitations, then expanding on what we see as the implications of this work for theories of concepts and of development and finally pointing to some future directions.

Limitations

Experimental Control

While this task and new dataset provide an exceptionally rich window on developmental differences in inductive inference, some of what is gained in open-endedness is lost in experimental control. There is residual ambiguity about the extent that differences in active learning caused differences in hypothesis generation and visa versa. Partialing this out would require experiments that fix the evidence and probe the hypotheses generated, or that fix the hypotheses in play and probe what evidence is sought. However, we have argued that such constrained tasks run the risk of short-circuiting natural cognition. Learners may struggle to test hypotheses they did not conceive themselves, and are known to struggle to use data they have not generated to evaluate their hypotheses (Markant & Gureckis, 2014; Sobel & Kushnir, 2006). We take this to mean that sole quantitative focus on studies that fix one or other aspect of the problem may provide a misleading perspective on end-to-end active inference in the wild. We feel that our open ended task provides a valuable complementary perspective to more constrained tasks.

Theoretical Expressivity

There are many ways we could have set up the primitives, parameters and productions of our PCFG and IDG. This makes for a dangerously expressive set of theories of cognition. We do not claim to have explored this space exhaustively here but rather that our modeling lends support to the idea that some symbolic and compositional process drives children and adults' active inductive inferences about the world. That is, we can explain the variability and productivity of human hypothesis belief formation in symbolic terms. Identifying the computational primitives of thought may not be a realistic empirical goal since a feature of constructivist accounts is their flexibility. Learners can grow their concept grammar over time, caching new primitives that prove useful (Piantadosi, 2021). Moreover, it is well known many different symbol systems can mimic one another (Turing, 1937), meaning that expressivity alone cannot distinguish between them. Since, we expect different learners to take different paths in an inherently stochastic learning trajectory, this limits universal claims about representational content.

Feature selection

We assumed our scenes had directly observable features and cued these to participants in our instructions. However, a number of recent models in machine learning combine neural network methods for feature extraction with compositional engines for symbolic inference, creating hybrid systems that can learn rules and solve problems from raw inputs like natural images (cf. Nye, Solar-Lezama, Tenenbaum, & Lake, 2020; Valkov, Chaudhari, Srivastava, Sutton, & Chaudhuri, 2018). We see these approaches as having promise to bridge the gap between subsymbolic and symbolic cognitive processing.

Sampling differences between children and adults

One potential concern is that the complexity of children's guesses relative to adults stems partly from their being collected verbally and in the presence of an experimenter rather than typed during an online experiment. Speaking carries different cognitive demands than typing and may lead to children simply responding in a more verbose way than adults. While we cannot rule this out, we do not think this is a major concern. Adults were well compensated for accuracy, meaning their motivation was primarily to be correct rather than brief. The semantic content of both children's and adults rules were extracted through our coding of them into lambda calculus meaning that surface differences in concise expression can be separated from logical complexity. Furthermore children's guesses were not the only thing that was more elaborate about their behavior. They were also more elaborate in their active testing choices, producing more complex

scenes despite having to create these in the same manner as adults. Since the testing interface was reset on each trial, this complexity took more effort, with children's scenes requiring substantially more clicks and more time to produce than adults'.

Use of verbal protocols

Another worry about our use of free responses is that they rely on a capacity for precise linguistic expression not to mention the assumption that learners have insight into the structure of their own concepts. It is known that children's vocabularies differ from adults', raising the concern that some of our results reflect language use rather than the concepts being articulated. While our artificial environment contains only simple objects and basic features that are familiar to even young children, there is evidence that children's speech does not distinguish as well among quantifier usage (e.g., all, each, every) until late in childhood (Brooks & Braine, 1996; Inhelder & Piaget, 1958). Thus, it could be that linguistic imprecision is behind some of the differences between children's and adults' guesses. For instance, this seems like a potential explanation for the lack of any exactly correct guesses from children about the quantifier-dependent rule 4 "exactly one is blue". However, a closer look at responses reveals that only 11/47 children guessed a rule that mentioned blue at all. Meanwhile 37/50 of adults' rules mentioned blue, but all but seven of these were wrong about the particulars of the quantification. In many cases other potential quantifications were not ruled out by adults' testing. For instance, several subjects never tried adding more than one blue object to a scene and later responded that *at least one* object must be blue. Thus, it seems that children's rules simply picked out different features of the scenes than adults. An interesting question is whether, in the cases where a child's guess is logically inconsistent with some of their learning data, this is because their representation itself is imprecise, or because their verbal description imprecisely describes their representation. Another possibility could be that adults are better introspectors than children, better able to "read out" the structure of their own representations (Morris, 2021). While these are intriguing possibilities our current experiment cannot fully resolve these explanations.

Implications for theories of concepts

As we noted at the outset, psychological theories of concepts have oscillated between symbolic accounts — that seek to explain conceptual productivity and creativity — and similarity accounts — that seek to explain how concepts drive probabilistic generalization. The constructivist framework is based in the symbolic camp, however it inherits many of the advantages of similarity accounts by maintaining a relationship with probabilistic

inference embodied by the stochastic mechanisms of generation and search. Thus, we see our findings as support for recent claims that higher level cognition utilizes some form of stochastic generative sampling to approximate rational inference (Bramley, Dayan, et al., 2017; A. Sanborn et al., 2021; Zhu, Sanborn, & Chater, 2020) and that this might also explain aspects of human cultural and technological development that take place over populations and multiple generations (Krafft, Shmueli, Griffiths, Tenenbaum, et al., 2021).

As it stands, neither our PCFG or IDG are plausible process models of concept formation. The PCFG is a framework for normative top-down inference in the limit of infinite sampling, and IDG is a hybrid that is less sample inefficient as a brute force approach to inference in situations where a learner already has some evidence. A process account needs to explain how a learner searches the latent posterior in either framework with limited memory and computation. Following a number of recent research lines (Bramley, Mayrhofer, Gerstenberg, & Lagnado, 2017; Dasgupta, Schulz, & Gershman, 2017; Fränken, Theodoropoulos, & Bramley, in revision; Ullman, Goodman, & Tenenbaum, 2012), we see incremental adaptation of one or a few focal hypotheses in the light of evidence as a promising approach. A learner might use an observation to generate an initial idea akin to our IDG, but then explore permutations to this to account for the rest of the evidence. While hypothesis search models like RULEX (Nosofsky & Palmeri, 1998; Nosofsky et al., 1994) provide candidate heuristics for achieving such a search, their long run behavior lacks a clear relationship with computational-level rationality (Navarro, 2005). However, if a learners' algorithmic adaptations approximate a valid approximation scheme, for instance accepting permutations with the Metropolis-Hastings probability $\max(1, \frac{P(r')}{P(r^t)})$ (Bramley, Dayan, et al., 2017; Dasgupta et al., 2016; Hastings, 1970; Thaker et al., 2017), we can explain why more probable hypotheses are discovered more often as well as explaining probability matching and order effects directly from our account of generation. Since the endpoint of an MCMC search approaches an independent posterior sample, we would expect a population of searchers to end up with a set of hypotheses that look like posterior samples. Moreover, since individual searchers have finite time to search, we would expect order effects and dependence in their ideas over time. To the extent that participants deviate from a probabilistically valid approximation scheme, for instance "hill climbing" or accepting only strictly better fitting ideas, we might also explain how they can get stuck in local optima and exhibit mal-adaptive order effects like garden paths (Gelpi, Prystawski, Lucas, & Buchsbaum, 2020).

Our modeling of generalizations revealed that there is no straightforward family resemblance between the features of rule-following training scenes (generated by the participant) and rule-following generalization scenes (as pre-selected for the experiment).

This resulted in our Similarity model performing at chance and uncorrelated with participants while all our symbolic model variants received some support. While this is far from an exhaustive test of sub-symbolic concept models, even a successful similarity-based account of generalizations would only account for half of the behavior in this task. As well as generalizing systematically, participants gave detailed natural language descriptions of their ideas. The majority of these we could convert into logical statements (86%) that predicted most generalizations (72%: children, 84%: adults) and were consistent with the majority of their learning data (71%: children, 87%: adults). Any subsymbolic account of concepts would essentially need to be paired with an explanation for how people generate these verbal descriptions of their non-symbolic concepts that nonetheless reflect their use (cf. Dennett, 1988). Arguably, this task is no easier than the one of generating a symbolic hypothesis about the nature of the world in the first place. Thus we feel that our results are more straightforwardly explained by our symbolic account whereby the logical structure of the hypotheses participants describe is actually the causal mechanism driving their generalizations rather than some form of computationally expensive but behaviorally impotent retrospective confabulation (cf. Johansson et al., 2008).

Implications for theories of development

Our analyses revealed a variety of developmental differences. Children's guesses were more complex than adults', and consequently we could capture them with a significantly "flatter" generation process that inherently produced a wider diversity of hypotheses. This is broadly normative: Having been exposed to less evidence, with less idea what conceptual compositions and fragments will be useful in understanding their environment, we would expect children's construction process to be less fine tuned. In other words, children are justified in entertaining a wider set of ideas than adults. However, we noted at the start that there are several potentially distinct algorithmic stories that could underpin this diversity: (1) children might simply have hypothesis generation mechanism that embodies a rationally flatter latent prior, (2) they might additionally explore more radically, over and above differences in the relative credibility their latent prior actually attaches to different possibilities (Gopnik, 2020; Lucas et al., 2014; Wu et al., 2017) or (3) we also considered that children's generation mechanisms might be more dominated by "bottom-up" processes. We take our comparison of PCFG and IDG to speak against option 3. Adults' hypotheses were, as far as we could tell, at least as anchored to idiosyncratic patterns of their learning data as children's. However these data do not distinguish clearly between options (1) and (2). To do this, one would need to collect multiple judgments from the same participant within a learning problem, and somehow also elicit judgments about the

relative quality of these guesses. If children's guesses shift within a problem in a way that is less sensitive to the relative subjective probabilities than adults, this would support the idea that children's hypothesis generation is more "high temperature" exploratory than adults' (Gopnik, 2020), over and above differences in the flatness of their latent prior.

As well as producing more complex guesses, children also produced more elaborate scenes during active learning than adults. One possible characterization is that children's active scene construction also depended on a "flatter" generative prior, resulting in more diversity of exploration approaches. Indeed, differences in active exploration are the other side of the coin of the high temperature search idea (Friston et al., 2016; Gopnik, 2020; Klahr & Dunbar, 1988; E. Schulz, Klenske, Bramley, & Speekenbrink, 2017). On the other hand, the problem of generating informative tests is not quite the same as that of finding the right hypothesis. It is important to avoid redundancy and produce tests that target uncertainty and, in combination, serve to test a wide variety of salient hypotheses. In this sense, adults' testing behavior was more systematic, better reducing global measures of entropy and potentially reflecting a more top-down, or strategic, *control of variables* approach to gathering evidence and updating beliefs (Kuhn & Brannock, 1977). Within each trial, children's testing was more repetitive than adults', suggesting that they made slower progress in exploring the problem space, or were generally less able to keep track of what they had done, perhaps mixing more random actions in with systematic ones (Meder, Wu, Schulz, & Ruggeri, 2021).

Children's guesses were also less consistent with their evidence than adults'. This might be because they were less able to extract common features across learning scenes (Ruggeri & Feufel, 2015; Ruggeri & Lombrozo, 2015). However, it could also be a consequence of a more generalized limitation in ability to generate, store and compare hypotheses. With a flatter prior and limited sampling, one has a lower chance generating a hypothesis that can explain all the evidence. Children also under-generalized, often selecting only 1 or 2 of the 8 test scenes (there was actually always 4) doing so even when their symbolic guesses predicted more should be selected. It could be that children found this part of the task overwhelming, perhaps tending to stop after identifying one or two hypothesis consistent scenes rather than evaluating all of them. On the face of it, this reflects Wu et al.'s (Wu et al., 2017) finding that children are weaker generalizers than adults.

Curiously, children were more likely to refer to relational and positional properties in their guesses, while adults were by most likely to make guesses that pertained to the primary object features (color and size). This is an independently interesting finding. Since relational features are structurally more complex than primitive features, we might have

predicted they would be more readily evoked by adults. It could be that children bought in more to the scientific reasoning cover story, treating mechanistic explanations, such as that objects must touch or be positioned in particular ways to produce stars, as credible (Gelman, 2004). Conversely, adults may have been more likely to expect Gricean considerations to apply, e.g. that experimenters would likely set simple rules using salient but abstract features like color over perceptually ambiguous properties like position (Szollosi & Newell, 2020). However, it could also be the case that there are deeper differences between the experiences of children and adults that render structural features more relevant to children and surface features more relevant to adults.

Conclusions

We analyzed an experiment combining rich qualitative and quantitative measures of children's and adults' inductive inference. We found a number of developmental differences and demonstrated that we can make sense of these through the computational constructivism lens. Our results add empirical support and theoretical detail to recent characterizations of children as more diverse thinkers and active learners than adults, and bring us closer to a computational understanding of human learning across the lifespan.

References

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

Bonawitz, E. B., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338.

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.

Bramley, N. R., Mayrhofer, R., Gerstenberg, T., & Lagnado, D. A. (2017). Causal learning from interventions and dynamics in continuous time. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Bramley, N. R., Rothe, A., Tenenbaum, J. B., Xu, F., & Gureckis, T. M. (2018). Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Brooks, P. J., & Braine, M. D. (1996). What do children know about the universal quantifiers all and each? *Cognition*, *60*(3), 235–268.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* Routledge.

Bruner, J. S., Jolly, A., & Sylva, K. (1976). *Play: Its role in development and evolution.* Penguin.

Buchanan, D. W., Tenenbaum, J. B., & Sobel, D. M. (2010). Edge replacement and nonindependence in causation. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 919–924). Austin, TX: Cognitive Science Society.

Carey, S. (1985). Are children fundamentally different kinds of thinkers and learners than adults. *Thinking and learning skills*, *2*, 485–517.

Carey, S. (2009). *The origin of concepts: Oxford series in cognitive development.* Oxford University Press, England.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the

control of variables strategy. *Child Development*, *70*(5), 1098–1120.

Clark, A. (2012). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral Brain Sciences*, 1–86.

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.

Coenen, A., Rehder, R., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

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*, *45*(11), 1923.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2016). Where do hypotheses come from? *Center for Brains, Minds and Machines (preprint)*.

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

Dennett, D. C. (1988). The intentional stance in theory and practice. In R. Byrne & A. Whiten (Eds.), *Machiavellian intelligence* (pp. 180–202). Oxford, UK: Oxford University Press.

Dennett, D. C. (1991). *Consciousness explained*. London, UK: Penguin.

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., ... Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*.

Fedyk, M., & Xu, F. (2018). The epistemology of rational constructivism. *Review of Philosophy and Psychology*, *9*(2), 343–362.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Fränken, J.-P., Theodoropoulos, N. C., & Bramley, N. R. (in revision). Algorithms for adaptation in inductive inference.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in cognitive sciences*, *8*(9), 404–409.

Gelpi, R., Prystawski, B., Lucas, C. G., & Buchsbaum, D. (2020). Incremental hypothesis

revision in causal reasoning across development.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251–256.

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational behavior and human performance*, *24*(1), 93–110.

Ginsburg, S. (1966). *The mathematical theory of context free languages*. McGraw-Hill Book Company.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–9.

Gopnik, A. (1996). The scientist as child. *Philosophy of science*, *63*(4), 485–514.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, *375*(1803), 20190502.

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.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.

Gureckis, T. M., & Markant, D. B. (2012, September). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

Heath, C. (2004). *Zendo–Design History*. Retrieved from `http://www.koryheath.com/zendo/design-history/`

Howson, C., & Urbach, P. (2006). *Scientific reasoning: the Bayesian approach*. Open Court Publishing.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures* (Vol. 22). Psychology Press.

Johansson, P., Hall, L., & Sikström, S. (2008). From change blindness to choice blindness. *Psychologia*, *51*(2), 142–155.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, *1*(2), 112–133.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press.

Jones, A., Bramley, N. R., Gureckis, T. M., & Ruggeri, A. (in revision). Changing many things at once sometimes makes for a good experiment, and children know that.

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and Brain Sciences*, *34*(4), 169–88.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–243.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*(1), 1–48.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*(1), 111–146.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science*, *333*(6045), 971–975.

Klayman, J., & Ha, Y.-w. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *15*(4), 596.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological bulletin*, *112*(3), 500.

Krafft, P. M., Shmueli, E., Griffiths, T. L., Tenenbaum, J. B., et al. (2021). Bayesian collective learning emerges from heuristic social learning. *Cognition*, *212*, 104469.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22.

Kuhn, D., & Brannock, J. (1977). Development of the isolation of variables scheme in experimental and "natural experiment" contexts. *Developmental psychology*, *13*(1), 9.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *32*(3), 451–60.

Lai, L., & Gershman, S. J. (2021). Policy compression: an information bottleneck in action selection.

1203 Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building
1204 machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.
1205 Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and
1206 reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).
1207 Lewis, O., Perez, S., & Tenenbaum, J. (2014). Error-driven stochastic search for theories
1208 and concepts. In *Proceedings of the annual meeting of the cognitive science society*
1209 (Vol. 36).
1210 Lieder, F., Griffiths, T., Huys, Q. J., & Goodman, N. D. (2017). The anchoring bias
1211 reflects rational use of cognitive resources.
1212 Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category
1213 learning. *Psychological Review*, *111*(2), 309.
1214 Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better
1215 (or at least more open-minded) learners than adults: Developmental differences in
1216 learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.
1217 Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using
1218 hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147.
1219 Luce, D. R. (1959). *Individual choice behavior*. New York: Wiley.
1220 Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? learning via
1221 active and passive hypothesis testing. *Journal of Experimental Psychology: General*,
1222 *143*(1), 94.
1223 Marr, D. (1982). *Vision*. New York: Freeman & Co.
1224 McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. A. (2016).
1225 Children's use of interventions to learn causal structure. *Journal of Experimental*
1226 *Child Psychology*, *141*, 1–22.
1227 Meder, B., Wu, C. M., Schulz, E., & Ruggeri, A. (2021). Development of directed and
1228 random exploration in children. *Developmental Science*, *24*(4), e13095.
1229 Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning.
1230 *Psychological Review*, *85*(3), 207.
1231 Meng, Y., Bramley, N., & Xu, F. (2018). Children's causal interventions combine
1232 discrimination and confirmation. In *Proceedings of the 40th annual conference of the*
1233 *cognitive science society*.
1234 Michalski, R. S. (1969). On the quasi-minimal solution of the general covering problem. In
1235 (Vol. A3, pp. 125–128).
1236 Morris, A. (2021). Invisible gorillas in the mind: Internal inattentional blindness and the
1237 prospect of introspection training.
1238 Navarro, D. J. (2005). Analyzing the rulex model of category learning. *Journal of*

Mathematical Psychology, *49*(4), 259–275.

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–99.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*(3), 345–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53.

Nye, M. I., Solar-Lezama, A., Tenenbaum, J. B., & Lake, B. M. (2020). Learning compositional rules via neural program synthesis. *arXiv preprint arXiv:2003.05562*.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to bayesian models of cognitive development. *Cognition*, *120*(3), 302–321.

Piaget, J. (2013). *The construction of reality in the child* (Vol. 82). Routledge.

Piaget, J., & Valsiner, J. (1930). *The child's conception of physical causality*. Transaction Pub.

Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines*, *31*(1), 1–58.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3p1), 353.

Quine, W. v. O. (1969). *Word and object*. MIT press.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2017). Question asking as program generation. *arXiv preprint arXiv:1711.06351*.

Ruggeri, A., & Feufel, M. (2015). How basic-level objects facilitate question-asking in a categorization task. *Frontiers in psychology*, *6*, 918.

Ruggeri, A., & Lombrozo, T. (2014). Learning by asking: How children ask questions to achieve efficient search. In *36th annual meeting of the cognitive science society* (pp. 1335–1340). Austin, TX: Cognitive Science Society.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient

search. *Cognition*, *143*, 203–216.

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.

Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., & Lake, B. M. (2020). A benchmark for systematic generalization in grounded language understanding. *arXiv preprint arXiv:2003.05161*.

Rule, J. S., Schulz, E., Piantadosi, S. T., & Tenenbaum, J. B. (2018). Learning list concepts through program induction. *BioRxiv*, 321505.

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in Cognitive Sciences*.

Sanborn, A., Zhu, J., Spicer, J., Sundh, J., León-Villagrá, P., & Chater, N. (2021). Sampling as the human approximation to probabilistic inference.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*.

Schulz, E., Klenske, E. D., Bramley, N. R., & Speekenbrink, M. (2017). Strategic exploration in human adaptive control. *bioRxiv*, 110486.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, A. C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211–223.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, *30*, 50–64.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., & Chang, J.-J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, *65*(1), 94.

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.

Sobel, D. M., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition*, *34*(2), 411–419.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive*

Psychology, *53*(1), 1–26.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Szollosi, A., & Newell, B. R. (2020). People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences*.

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Thaker, P., Tenenbaum, J. B., & Gershman, S. J. (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology*, *77*, 10–20.

Turing, A. M. (1937). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London mathematical society*, *2*(1), 230–265.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

Valkov, L., Chaudhari, D., Srivastava, A., Sutton, C., & Chaudhuri, S. (2018). Houdini: Lifelong learning as program synthesis. In *Advances in Neural Information Processing Systems* (pp. 8687–8698).

Van Laarhoven, P. J., & Aarts, E. H. (1987). Simulated annealing. In *Simulated annealing: Theory and applications* (pp. 7–15). Springer.

Van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press.

von Humboldt, W. (1863/1988). *On language*. New York: Cambridge University Press.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? optimal decisions from very few samples. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (Vol. 1, pp. 66–72). Austin, TX: Cognitive Science Society.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, *12*(3), 129–140.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2017). Exploration and generalization in vast spaces. *bioRxiv*. doi: 10.1101/171371

Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review*, *126*(6), 841.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological review*, *127*(5), 719.

Appendix A: Models

Generating context free (PCFG) model predictions

We created a grammar (specifically a *probabilistic context free grammar* or PCFG; Ginsburg, 1966) that can be used to produce any rule that can be expressed with first-order logic and lambda abstraction referring to the features participants referred to in our task. The grammatical primitives we assumed are detailed in Table A-1.

Table A-1
A Concept Grammar for the Task

Meaning	Expression
There exists an x_i such that...	$\exists(\lambda x_i :, \mathcal{X})$
For all x_i ...	$\forall(\lambda x_i : ., \mathcal{X})$
There exists {at least, at most, exactly} N objects in x_i such that...	$N_{\{<,>,=\}}(\lambda x_i : ., N, \mathcal{X})$
Feature f of x_i has value {larger, smaller, (or) equal} to v	$\{<, >, \leq, \geq, =\}(x_i, v, f)$
Feature f of x_i is {larger, smaller, (or) equal} to feature f of x_j	$\{<, >, \leq, \geq, =\}(x_i, x_j, f)$
Relation r between x_i and x_j holds	$\Gamma(x_i, x_j, r)$
Booleans {and,or,not}	$\{\wedge, \vee, \neq\}(x)$
Object feature	**Levels**
Color	{red, green,blue}
Size	{1:small, 2:medium, 3:large}
x-position	(0,8)
y-position	(0,8)
Orientation	{Upright, left hand side, right hand side, strange}
Grounded	true if touching the ground
Pairwise feature	**Condition**
Contact	true if x_1 touches x_2
Stacked	true if x_1 is above and touching x_2 and x_2 is grounded
Pointing	true if x_1 is orientated {left/right} and x_2 is to x_1s {left/right}
Inside	true if x_1 is smaller than x_2 + has same x and y position (± 0.3), false

Note that $\{<, >, \geq, \leq\}$ comparisons only apply to numeric features (e.g., size).

There are multiple ways to implement a PCFG. Here we adopt a common approach to set up a set of string-rewrite rules (Goodman et al., 2008). Thus, each hypothesis begins life as a string containing a single *non-terminal symbol* (here, S) that is replaced using

rewrite rules, or *productions*. These productions are repeatedly applied to the string, replacing non-terminal symbols with a mixture of other non-terminal symbols and terminal fragments of first order logic, until no non-terminal symbols remain. The productions are so designed that the resulting string is guaranteed to be a valid grammatical expression and all grammatical expressions have a nonzero chance of being produced. In addition, by having the productions tie the expression to bound variables and truth statements, our PCFG serves as an automatic concept generator. Table A-2 details the PCFG we used in the paper.

We use capital letters as non-terminal symbols and each rewrite is sampled from the available productions for a given symbol.[6] Because some of the productions involve branching (e.g., $B \rightarrow H(B, B)$), the resultant string can become arbitrarily long and complex, involving multiple boolean functions and complex relationships between bound variables.

We include a variant that samples uniformly from the set of possible replacements in each case, but we also reverse engineer a set of productions that produce exactly the statistics of the empirical samples, as described in the main text.

We used the process described in A-2 to produce a sample of 10,000 with a uniform generation prior and an additional 10,000 for each participant with a "held out" age-consistent prior based on the rule guesses of other participants in the requisite agegroup. For the flipped prior analyses, we used the sample generated for the chronologically first participant from the other agegroup.

Generating instance driven (IDG) model predictions

We used the algorithm proposed in Bramley et al. (2018) to produce a sample of 10,000 "grounded hypotheses" for each participant and trial, splitting these evenly across the 8 learning scenes that participant produced and tested. For each, we generated two sets: One using a uniform construction weights, and one with an age-appropriate "held out" set of weights based on the rule guesses of other participants in the requisite agegroup. For the flipped prior analyses, we used the weights from the chronologically first participant from the other agegroup to generate samples inspired by the current participants' evidence.

To generate hypotheses as candidates for the hidden rule, the model uses the following procedure with probabilities either set to uniform or drawn from the PCFG-fitted

[6] The grammar is not strictly context free because the bound variables (x_1, x_2, etc.) are automatically shared across contexts (e.g. x_1 is evoked twice in both expressions generated in Figure 2a). We also draw feature value pairs together and conditional on the type of function they inhabit, to make our process more concise, however the same sampling is achievable in a context free way by having a separate function for every feature value, i.e. "'isRed()" and sampling these directly (c.f. Rothe, Lake, & Gureckis, 2017).

Table A-2
Prior Production Process

Production	Symbol	Replacements→		
Start	$S \rightarrow$	$\exists(\lambda x_i\colon\ A, \mathcal{X})$	$\forall(\lambda x_i\colon\ A, \mathcal{X})$	$N_I(\lambda x_i\colon\ A, K, \mathcal{X})$
Bind additional	$A \rightarrow$	B	S	
Expand	$B \rightarrow$	C	$J(B, B)$	$\neg(B)$
Function	$C \rightarrow$	$=(x_i, D1)$	$I(x_i, D2)$	$=(x_i, x_j, E1)^{\mathbf{a}}$
		$I(x_i, x_j, E2)^{\mathbf{a}}$	$\Gamma(x_i, x_j, E3)^{\mathbf{a}}$	
Feature/value	$D1 \rightarrow$	value,	feature	
(numeric only)	$D2 \rightarrow$	value,	feature	
Feature	$E1 \rightarrow$	feature		
(numeric only)	$E2 \rightarrow$	feature		
(relational)	$E3 \rightarrow$	feature		
Boolean	$J \rightarrow$	$\wedge$	$\vee$	...
Inequality	$I \rightarrow$	$\leq$	$\geq$	$>$
		$<$		
Number	$K \rightarrow$	$n \in \{1, 2, 3, 4, 5, 6\}$		

Note: Context-sensitive aspects of the grammar: [a]Bound variable(s) sampled uniformly without replacement from set; expressions requiring multiple variables censored if only one.

productions for adults or for children (Figure 3g&h) and denoted with square brackets:

1. **Observe.** either:

   (a) With probability $[A \rightarrow B]$: Sample a cone from the observation, then sample one of its features f with probability $[G \rightarrow f]$ — e.g., {#1}:[7] "medium, size" or {#3}: "red, color".

   (b) With probability $[A \rightarrow \text{Start}]$: Sample two cones uniformly without replacement from the observation, and sample any shared or pairwise feature — e.g., {#1,#2}: "size", or "contact"

2. **Functionize.** Bind a variable for each sampled cone in Step 1 and sample a true (in)equality statement relating the variable(s) and feature:

   (a) For a statement involving an unordered feature there is only one possibility — e.g, {#3}: "$= (x_1, \text{red}, \text{color})$", or for {#1,#2}: "$=(x_1, x_2, \text{color})$"

   (b) For a single cone and an ordered feature, this could also be a nonstrict inequality ($\geq$ or $\leq$). We assume a learner only samples an inequality if it expands the number of cones picked out from the scene relative to an equality

[7] Numbers prepended with # refer to the labels on the cones in the example observation in Figure 2b.

— e.g., in Figure 2b in the main text, there is also a large cone {#1} so either $\geq(x_1, \text{medium}, \text{size})$ or $=(x_1, \text{medium}, \text{size})$ might be selected with uniform probability.

(c) For two cones and an ordered feature, either strict or non-strict inequalities could be sampled if the cones differ on the sampled feature, equivalently either equality or non-strict inequality could be selected if the cones do not differ on that dimension — e.g., {#1,#2} $>(x_1, x_2, \text{size})$, or {#3,#4} $\geq(x_1, x_2, \text{size})$. In each case, the production weights from Figure 3g&h for the relevant completions are normalized and used to select the option.

3. **Extend.** With probability $\frac{[B\to D]}{[B\to D]+[B\to C(B,B)]}$ go to Step 4, otherwise sample a conjunction with probability $[C(B,B)\to \text{And}]$ or a disjunction with probability $[C(B,B)\to \text{Or}]$ and repeat. For statements with two bound variables, Step 3 is performed for x_1, then again for x_2:

(a) **Conjunction.** A cone is sampled from the subset picked out by the statement thus far and one of its features sampled with probability $[G\to f]$ — e.g., {#1} $\wedge(=(x_1, \text{green}, \text{color}), \geq(x_1, \text{medium}, \text{size}))$. Again, inequalities are sample-able only if they increase the true set size relative to equality — e.g., "$\wedge(\leq(x_1, 3, \text{xposition}), \geq(x_1, \text{medium}, \text{size}))$", which picks out more objects than "$\wedge(=(x_1, 3, \text{xposition}), \geq(x_1, \text{medium}, \text{size}))$".

(b) **Disjunction.** An additional feature-value pair is selected uniformly from *either* unselected values of the current feature, *or* from a different feature — e.g., $\vee(=(x_1, \text{color}, \text{red}), =(x_1, \text{color}, \text{blue}))$ or $\vee(=(x_1, \text{color}, \text{blue}), \geq(x_1, \text{size}, 2))$. This step is skipped if the statement is already true of all the cones in the scene.[8]

4. **Flip.** If the inspiration scene is not rule following wrap the expression in a $\neg()$.

5. **Quantify.** Given the contained statement, select true quantifier(s):

(a) For statements involving a single bound variable (i.e., those inspired by a single cone in Step 1) the possible quantifiers simply depend on the number of the cones in the scene for which the statement holds. If the statement is true of all cones in the scene Quantifier is selected using probabilities [Start→] combined with $[L\to]$ where appropriate. If it is true of only a subset of the cones then

[8] We rounded positional features to one decimal place in evaluating rules to allow for perceptual uncertainty.

$\forall(\lambda x_i : A, \mathcal{X})$ is censored and the probabilities re-normalized. K is set to match number of cones for which the statement is true.

(b) Statements involving two bound variables in lambda calculus have two nested quantifier statements each selected as in (a). The inner statement quantifying x_2 is selected first based on truth value of the expression while taking x_1 to refer to the cone observed in '1.'. The truth of the selected inner quantified statement is then assessed for all cones to select the outer quantifier — e.g., {#3,#4} "$\wedge(=(x_2, \text{green}, \text{color}), \leq (x_1, x_2, \text{size}))$" might become "$\forall(\lambda x_1 \colon \exists(\lambda x_2 \colon \wedge(=(x_2, \text{green}, \text{color}), \leq (x_1, x_2, \text{size})), \mathcal{X}), \mathcal{X})$". The inner quantifier $\exists$ is selected (three of the four cones are green {#1, #2, #4}), and the outer quantifier $\forall$ is selected (all cones are less than or equal in size to a green cone).

Note that a procedure like the one laid out above is, in principle, capable of generating any rule generated by the PCFG in Figure 3g&3h, but will only do so when exposed to an observation that is actually consistent with that rule, and will do so more often when the observation is inconsistent with as many other rules as possible (i.e., a minimal positive example). Step 4. allows that non-rule following scenes can be used to inspire rules involving a negation, for instance that "something is not upright" – which is semantically equivalent to saying that "nothing is upright". Basing hypotheses on instances may improve the quality of the effective sample of hypotheses that the learner generates.

One way to think of the IDG procedure is as a partial inversion of a PCFG. As illustrated by the blue text in the examples in Figure 2b in the main text. While the PCFG starts at the outside and works inward, the IDG starts from the central content and works outward out to a quantified statement, ensuring at each step that this final statement is true of the scene.

We note that it is possible, in principle, to calculate a lower bound on the prior probability for the PCFG or IDG generating a hypothesis that a participant reported, even if it does not occur in our sample. This can be achieved by reverse engineering the production steps that would be needed to produce the precise encoded syntax. This is a lower bound because it does not count semantically equivalent "phrasings" of the hypothesis that e.g. mention features in different orders or use logically equivalent combinations of booleans. We found that complex expressions tend to have a large number of "phrasings". In our sample-based approximation we implicitly treat semantically equivalent expressions as constituting the same hypothesis but note that determining semantic equivalence is an nontrivial aspect of constructivist inference that we do not fully

address here.

Full generalization model fits

As described in main text, we fit 18 model variants to participant's data. All models have between 0 and 2 parameters. For each model, we fit the parameter(s) by maximizing the model's likelihood of producing the participant data, using R's `optim` function. We compare models using the Bayesian Information Criterion (Schwarz, 1978) to accommodate their different numbers of fitted parameters.[9] Full results are in Table A-3.

Appendix B: Free response coding

To analyze the free responses, we first had two coders go through all responses and categorize them as either:

1. Correct: The subject gives exactly the correct rule or something logically equivalent
2. Overcomplicated: The subject gives a rule that over-specifies the criteria needed to produce stars relative to the ground truth. This means the rule they give is logically sufficient but not necessary. For example, stipulating that "there must be a small red" is overcomplicated if the true rule is "there must be a red" because a scene could contain a medium or large red and emit stars.
3. Overliberal: The opposite of overcomplicated. The subject gives a rule that under-specifies what must happen for the scene to produce stars. For example, stipulating that "there must be a blue" if the true rule is that "exactly one is blue". This is logically necessary but not sufficient because a scene could contain blue objects but not produce stars because there is not exactly one of them.
4. Different: The subject gives a rule that is intelligible but different from the ground truth in that it is neither necessary or sufficient for determining whether a scene will produce stars.
5. Vague or multiple. Nuisance category.

[9] On one perspective, our derivation of the child-like and adult-like productions constitutes fitting an additional 39 parameters ($m - 1$ for each production step), so evoking an additional BIC parameter penalty of $39 \times \log(3940) = 323$ for PCFG over PCFG Uniform and similarly for the IDG. If we were to apply this penalty, the uniform weighted variants would be clearly preferred under the BIC criterion at the aggregate level. It is less clear how to apply this penalty at the individual level since the held out priors are fit to different data than that being modeled. We chose to include the fitted versions alongside the uniform versions here without penalty as demonstrations of the differences that arise from different generation probabilities.

Table A-3
Models of Participants' Generalizations

	Model	Group	log(Likelihood)	BIC	λ	τ	N	Accuracy
1.	Baseline	children	-1319.75	2639.50			7	50%
2.	Bias	children	-1218.96	2445.47	0.32		**16**	50%
3.	PCFG Uniform	children	-1319.72	2647.00		58.17	0	61%
4.	PCFG Uniform + Bias	children	-1208.93	2432.97	0.35	2.18	0	
5.	PCFG Flipped	children	-1318.46	2644.47		8.97	1	66%
6.	PCFG Flipped + Bias	children	-1207.28	2429.67	0.34	2.07	0	
7.	PCFG Agegroup	children	-1319.58	2646.71		24.17	1	63%
8.	PCFG Agegroup + Bias	children	-1208.63	2432.36	0.35	2.15	0	
9.	IDG Uniform	children	-1298.73	2605.02		1.78	1	65%
10.	IDG Uniform + Bias	children	-1193.90	2402.90	0.32	1.19	0	
11.	IDG Flipped	children	-1315.49	2638.54		4.35	1	66%
12.	IDG Flipped + Bias	children	-1199.22	2413.54	0.35	1.38	0	
13.	IDG Agegroup	children	-1308.05	2623.65		2.51	2	69%
14.	IDG Agegroup + Bias	children	-1193.41	2401.93	0.34	1.19	0	
15.	Similarity	children	-1316.44	2640.42		-1.99	0	41%
16.	Similarity + Bias	children	-1214.71	2444.52	0.32	-1.30	1	
17.	Symbolic Guess	children	-1143.69	2294.92		1.02	15	62%
18.	**Symbolic Guess + Bias**	**children**	**-1067.18**	**2149.47**	**0.26**	**0.80**	9	
1.	Baseline	adults	-1386.29	2772.59			2	50%
2.	Bias	adults	-1364.90	2737.40	0.15		6	50%
3.	PCFG Uniform	adults	-1320.64	2648.89		1.27	0	63%
4.	PCFG Uniform + Bias	adults	-1253.52	2522.25	0.26	0.68	0	
5.	PCFG Flipped	adults	-1294.91	2597.42		1.06	1	66%
6.	PCFG Flipped + Bias	adults	-1229.18	2473.55	0.24	0.63	0	
7.	PCFG Agegroup	adults	-1266.96	2541.51		0.94	1	69%
8.	PCFG Agegroup + Bias	adults	-1203.64	2422.47	0.23	0.59	0	
9.	IDG Uniform	adults	-1228.21	2464.02		0.67	2	69%
10.	IDG Uniform + Bias	adults	-1179.12	2373.44	0.20	0.48	0	
11.	IDG Flipped	adults	-1245.56	2498.72		0.76	0	73%
12.	IDG Flipped + Bias	adults	-1179.23	2373.65	0.24	0.48	0	
13.	IDG Agegroup	adults	-1188.28	2384.17		0.62	2	74%
14.	IDG Agegroup + Bias	adults	-1134.58	2284.37	0.20	0.44	0	
15.	Similarity	adults	-1359.05	2725.70		-0.73	0	37%
16.	Similarity + Bias	adults	-1337.55	2690.30	0.14	-0.61	0	
17.	Symbolic Guess	adults	-893.49	1794.58		0.56	**32**	70%
18.	**Symbolic Guess + Bias**	**adults**	**-880.59**	**1776.38**	**0.08**	**0.50**	4	

NB: Accuracy column shows performance of the requisite model across 100 simulated runs through the task using participants' active learning data with τ set to 100 (essentially hard maximizing over the model's predictions). The Biased models perform strictly worse due to their bias so are not included in this column.

6. No rule. The subject says they cannot think of a rule.

We were able to encode 205/238 (86%) of the children's responses and (219/250) 87% for adults as correct, overcomplicated, overliberal or different. Table A-4 shows the complete confusion matrix. The two coders agreed 85% of the time, resulting in a Cohen's Kappa of .77 indicating a good level of agreement (Krippendorff, 2012).

We then had one coder familiar with the grammar go through each free response that was not assigned vague or no rule, and encode it as a function in our grammar. The second coder then blind spot checked 15% of these rules (64) and agreed in 95% of cases 61/64. The 6 cases of disagreement were discussed and resolved. In 5/6 cases, this was in favor of the primary coder. The full set of free text responses along with the requisite

Table A-4
Agreement Matrix for Independent Coders' Free Response Classifications

	correct	overliberal	overspecific	different	vague	no rule	multiple
correct	**93**	1	5	0	0	0	0
overliberal	5	**13**	1	8	0	1	0
overspecific	1	2	**42**	12	0	0	0
different	0	5	3	**224**	15	3	0
vague	0	1	2	3	**11**	6	0
no rule	0	0	0	0	0	**31**	0
multiple	0	1	0	2	0	0	**0**

classification, encoded rules are available in the Online Repository.

Appendix C: Scene similarity measurement

To establish the overall similarity between two scenes, we need to map the objects in a given scene to the objects in another scene (for example between the scenes in FigureA-1 a and b) and establish a reasonable cost for the differences between objects across dimensions. We also need a procedure for cases where there are objects in one scene that have no analogue in the other. We approach the calculation of similarity via the principle of minimum edit distance (Levenshtein, 1966). This means summing up the elementary operations required to convert scene (a) into scene (b) or visa versa. We assume objects can be adjusted in one dimension at a time (i.e. moving them on the x axis, rotating them, or changing their color, and so on.

Before focusing on how to map the objects between the scenes we must decide how to measure the adjustment distance for a particular object in scene a to its supposed analogue in scene b. As a simple way to combine the edit costs across dimensions we first Z-score each dimension, such that the average distance between any two values across all objects and all scenes and dimensions is 1. We then take the L1-norm (or city block distance) as the cost for converting an object in scene (a) to an object in scene (b), or visa versa. Note this is sensitive the size of the adjustment, penalizing larger changes in position, orientation or size more severely than smaller changes, while changes in color are all considered equally large since color is taken as categorical. Note also that for orientation differences we also always assume the shortest distance around the circle.

If scene (a) has an object that does not exist in scene (b) we assume a default adjustment penalty equal to the average divergence between two objects across all comparisons (3.57 in the current dataset). We do the same for any object that exists in (a) but not (b).

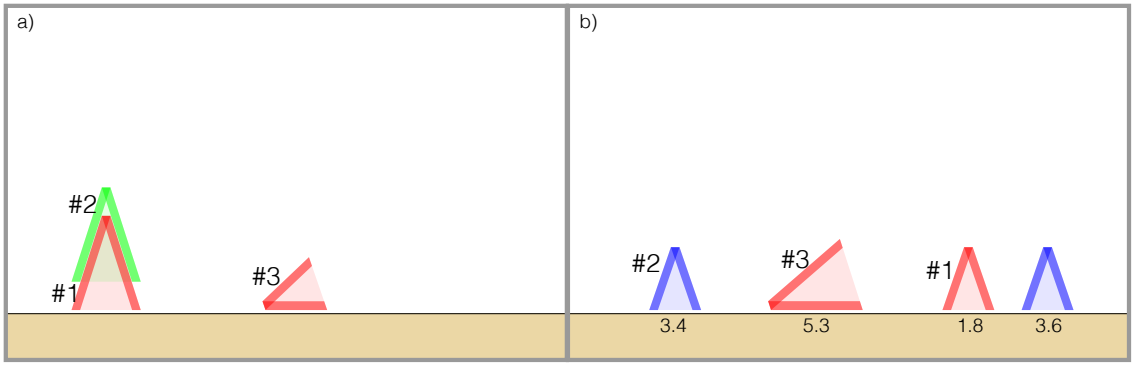


Figure A-1

Three example scenes. Objects indices link the most similar set of objects in b to those in a. Numbers below indicate the edit distance for each object (i.e. the sum of scaled dimension adjustments). Intuitively scene a) is more similar to scene b) than to scene c) and this is reflected in the similarity scores.

Calculating the overall similarity between two scenes involves solving a mapping problem of identifying which objects in scene (a) are "the same" as those in scene (b). We resolve this "charitably", by searching exhaustively for the mapping of objects in scene (a) to scene (b) that minimizes the total edit distance. Having selected this mapping, and computed the final edit distance including any costs for additional or removed objects, we divide by the number shared cones, so as to avoid the dissimilarities increasing with the number of objects involved.

Figure A-2 computes the inter-scene similarity components that go into Figure 6c in the main text. Summing up the edit distances across all objects, children's scenes seem much more diverse than adults (Figure A-2a). However this is primarily due to their containing a greater average number of objects. Scaling the edit distance by the number of objects in the target scene gives a more balanced perspective (Figure A-2b) but does not account for the fact that the compared scene may contain more or fewer objects in total. Figure A-2c visualizes just the object difference showing that children's scenes contain roughly as many objects on average as the initial example while adults' scenes contain around 0.75 fewer objects than are present in the initial example (dark shading in top row). Thus, we opted to combine b and c by weighting the unsigned cone difference by the mean inter-object distance across all comparisons to give our combined distance measure (Figure A-2d and Figure 6c in the main text).

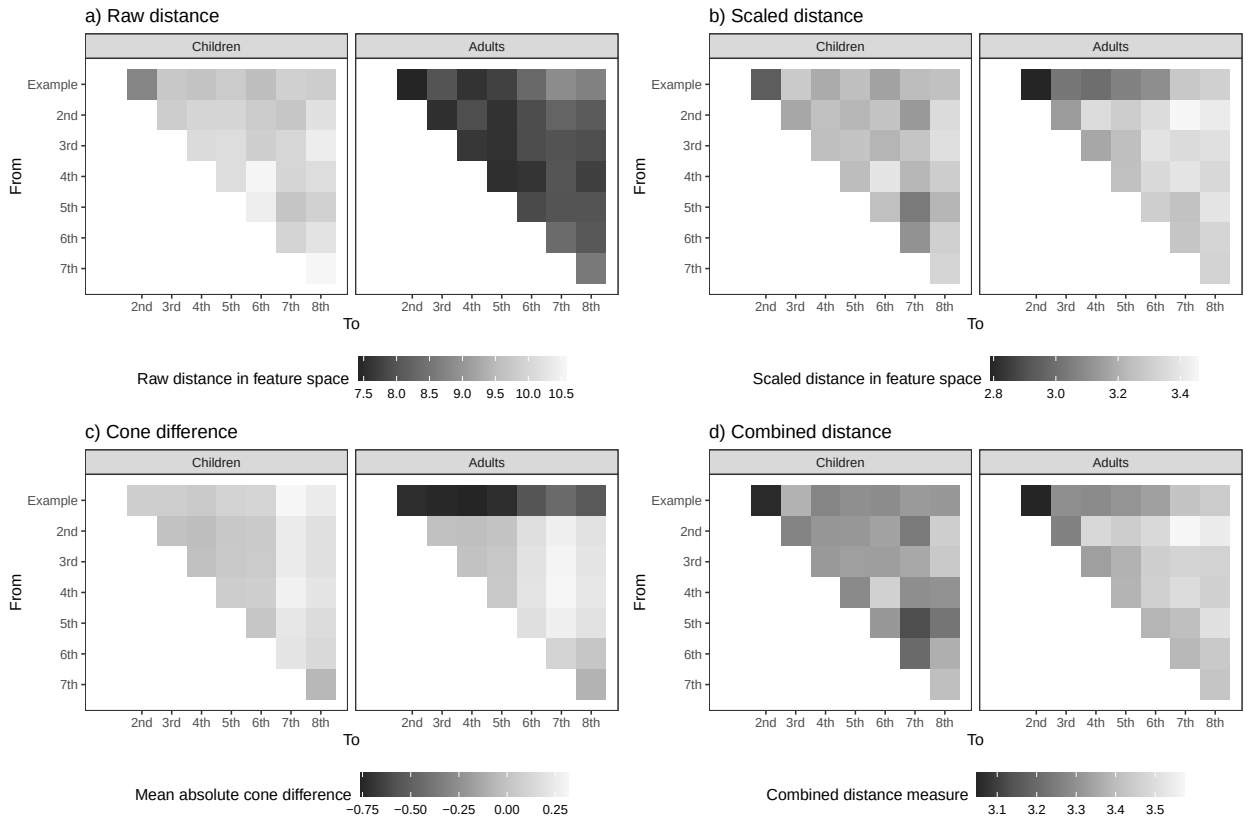


Figure A-2

a) The average minimum edit distance summed up across shared objects. b) Rescaling a by dividing by the number of objects. c) The penalty for additional or omitted objects. d)Combined distance as in main text.

Appendix D: Comparison with Bramley et al (2018)

Finally, for interest and to demonstrate replication of our core results. We provide a direct comparison between the generalization accuracies in the current sample of children and adults and those in the sample of 30 adults modelled in (Bramley et al., 2018). Bramley et al (2018) included 10 ground truth concepts, and the current paper uses just the first five of these. Figure A-3 shows these accuracy patterns side by side, revealing the adults in the current experiment performed approximately as well as those in the original conference paper.

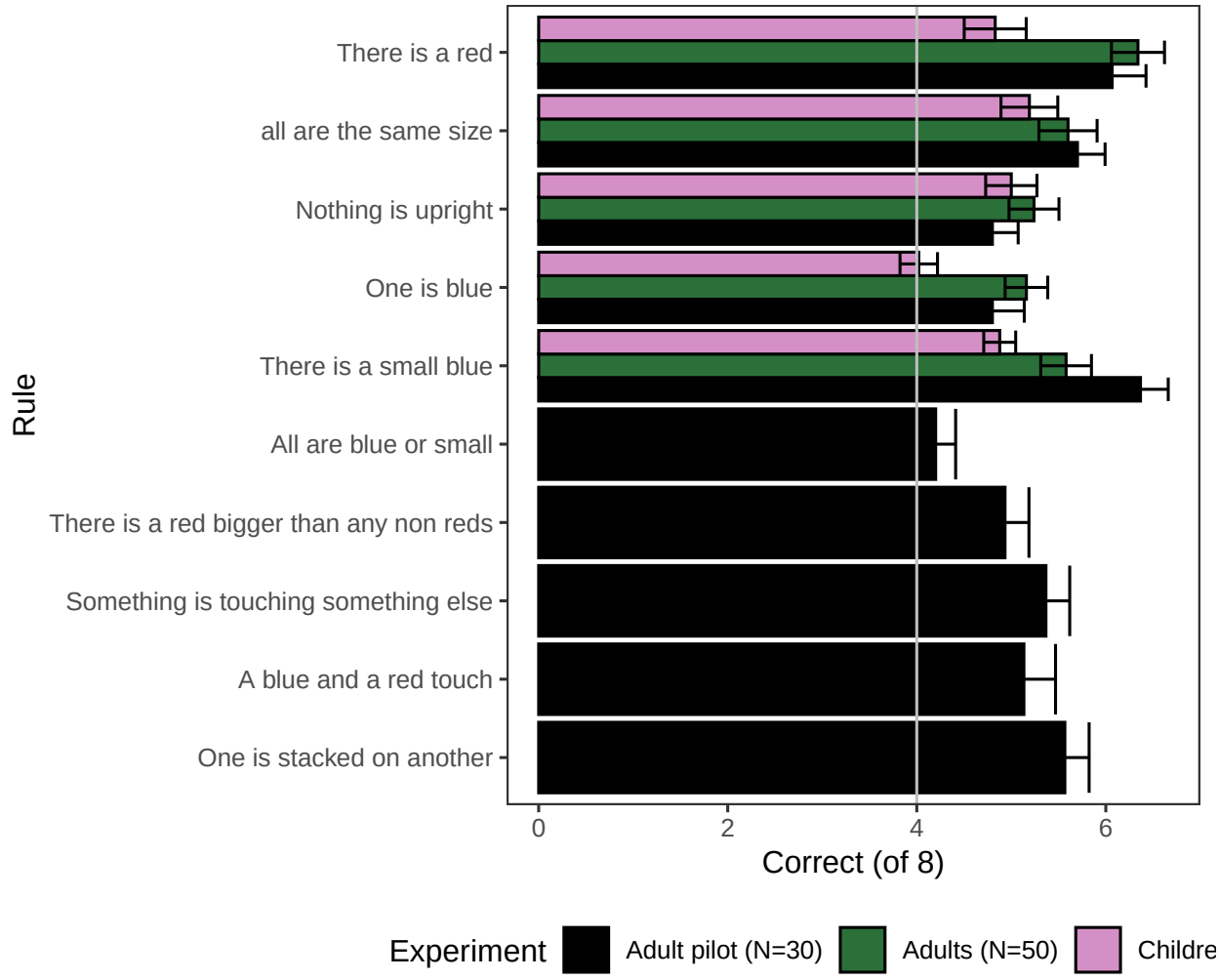


Figure A-3

Generalization accuracy by number of objects per test scene comparing with 10 rule adult pilot from Bramley et al. (2018).