Active inductive inference in children and adults: A constructivist perspective

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 were more elaborate in their testing behavior and generated substantially more complex guesses about the hidden rules. We take a 'computational constructivist' perspective to explaining these patterns, arguing that these inferences are driven by a combination of thinking (generating and modifying symbolic concepts) and exploring (discovering and investigating patterns in the physical world). We show how this framework and rich new dataset speak to questions about developmental differences in hypothesis generation, active learning and inductive generalization. In particular, we find children's learning is driven by less fine-tuned construction mechanisms than adults', resulting in a greater diversity of ideas but less reliable discovery of simple explanations.

Active inductive inference in children and adults: A constructivist perspective

"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 the ideas and hypotheses they use 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; Phillips, 1995; 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). While the tenets and promise of constructivist accounts are appealing, it has historically lacked the formalization needed to distinguish it from alternative accounts of learning, limiting its testable predictions or detailed insights into cognition. We draw on recent methodological advances to formalize key aspects of constructivism and use these to analyze children and adults' behavior in an open-ended inductive learning task. We show that a virtue of the constructivist account is that it captures the wide range of ideas and testing behaviors we observe, particularly in children. We use our account to examine developmental differences in hypothesis generation and active learning. To foreshadow, we show children's hypothesis generation and active learning are driven by less fine-tuned construction mechanisms than adults', resulting in a greater diversity of ideas but less reliable discovery of simple explanations and less systematic coverage of the data space.

Concept learning

Classic work in experimental psychology suggests symbol manipulation is required for humanlike reasoning and problem solving (Bruner, Goodnow, & Austin, 1956; Johnson-Laird, 1983; Wason, 1968). However, classic symbolic accounts struggled to explain how discrete representations could be learned or effectively applied to reasoning under uncertainty (Oaksford & Chater, 2007; Posner & Keele, 1968). Meanwhile, statistical accounts of concept learning have flourished by treating concepts as driven by "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 help explain how people assign category membership fuzzily, and generalize effectively to novel stimuli (Shepard, 1987) but lack a core representation capable of capturing how people construct conceptual novelty (Komatsu, 1992).

Bayesian approaches have also played a major role in 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). Alongside Bayesian analyses, information theory has also featured frequently as a metric of idealized evidence 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, since idealized Bayesian and information theoretic accounts describe learning within a predefined hypothesis space, they do not directly explain how a learner explores or generates possibilities within an infinite latent space. That is, probabilistic accounts of induction on are generally cast at Marr's computational level (Marr, 1982), showing people behave roughly *as if* they consider and average exhaustively over what is really an unbounded space of possible concepts. Thus, while these accounts provide a jumping off point for rational analysis of cognition, we should take their limitations seriously when seeking to reverse engineer humanlike inductive inference (Simon, 2013; Van Rooij, Blokpoel, Kwisthout, & Wareham, 2019).

The goal of this paper is to examine children's and adults' inductive learning in a rich open-ended task where the space of potential hypotheses and behaviors is effectively unbounded. In doing this, we will treat constructivism as a form of rational process framework (Lieder & Griffiths, 2020), capturing how people are shaped by Bayesian and information-theoretic norms but also why they diverge from and fall short of them outside of constrained scenarios. To do 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 to do so in ways that approximate 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*, and consequently *stochasticity* and *path dependence* in a given individual's learning trajectory. 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, they must have some mechanism for generating and comparing finite numbers of discrete possibilities (Sanborn & Chater, 2016; Stewart, Chater, & Brown, 2006). Eponymously, the construction mechanism needs to be capable of recursive *construction*: composing and recomposing symbolic elements so as to achieve the systemtaticity and productivity required for a finite system to cover an infinite space of ideas (Piantadosi & Jacobs, 2016). In this way, constructivist views treat algorithmic-level cognition as necessarily symbolic and at least somewhat language-like (Fodor, 1975) in its ability to make "infinite use of finite means" (von Humboldt, 1863/1988).

For example, a constructivist learner might stochastically combine elements from an underlying concept grammar to produce new ideas that can be tested against evidence. Alternatively, they might use their grammar to describe patterns in evidence or to adapt a previous hypotheses to fit some new evidence (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 gap between ease of evaluation versus generation of hypotheses (Gettys & Fisher, 1979). We can typically generate fewer explanations on the fly—i.e., reasons why our car won't start—than we would endorse if a list was presented to us. We would likely come up with more as we looked under the hood than we would sat in the car thinking. Inference about any area of active scientific inquiry, like that reported in this journal, typically involve an enormous latent space of potential explanatory theories only a fraction of which have ever been articulated or tested and many of which were discovered only serendipitously (Shackle, 2015). It is generally accepted that the ground truth is unlikely to be among the set of theories already on the table (Box, 1976) and that challenging results are as likely to lead to theory modification as complete abandonment (Lakatos, 1976).

The constructivist perspective thus departs from a Bayesian analysis by emphasizing that induction is as much about constructing candidate possibilities, as optimizing within a set of candidates. This reframing demystifies a number of behavioral patterns that look like biases from the computational-level perspective. These include *anchoring, order effects, probability matching* and *confirmation bias*. For example, *Anchoring* is a natural consequence of generating new hypotheses by making local adjustments to an earlier hypothesis or from a salient starting point such as a number mentioned in a prompt (Griffiths, Lieder, & Goodman, 2015; Lieder, Griffiths, Huys, & Goodman, 2018). *Order effects*, where the sequence of evidence encountered affects the final belief, 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 path dependence and auto-correlation between a single learner's hypotheses over time (Bramley, Dayan, et al., 2017; Dasgupta, Schulz, & Gershman, 2016; Fränken, Theodoropoulos, & Bramley, 2022; Thaker, Tenenbaum, & Gershman, 2017; Zhao, Lucas, & Bramley, 2022). *Probability matching* is also 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, it can be shown that rather than being a choice pathology, probability matching may be better seen as 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). It has been argued that in a variety of plausible everyday settings, a single-sample–based decision can be the appropriate computation–accuracy tradeoff for 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 tests that would be rule-following under their hypothesis (such as 6–8–12), so failing to adequately challenge and disconfirm this hypothesis. On a constructivist perspective, learners can only base their exploration on testing hypotheses they have actually generated (or else behave randomly). To the extent that certain simpler hypotheses like "ascending numbers" were less likely to be generated on the basis of the provided example (cf. Oaksford & Chater, 1994; Tenenbaum, 1999), it is not surprising that participants failed to actively exclude these possibilities with their tests.

In the computational cognitive science literature, recent 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 will draw on these in examining children and adults hypothesis generation.

Constructivism in Development

The "child as scientist" (Carey, 1985; Gopnik, 1996)—or recently, "child as hacker" (Rule, Tenenbaum, & Piantadosi, 2020) — perspective casts children's cognition as driven by broadly the same inductive 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) there is also evidence that children's ability to learn effectively from active learning data is more fragile than adults'. For example, children's play can look repetitive and inefficient when held to information theoretic norms (Lapidow & Walker, 2020; McCormack et al., 2016; Meng et al., 2018; Sim & Xu, 2017). Sobel and Kushnir (2006) also 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 the mismatch between whatever idiosyncratic hypotheses are under consideration by the observer and those being tested by the active learner, making the yoked learner less able to use the data to progress their theories (Fränken et al., 2022; Markant & Gureckis, 2014). Relatedly, children have been argued to be more narrowly focused toward testing a single hypothesis at a time (Bramley, Jones, Gureckis, & Ruggeri, 2022; 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. An early emphasis on exploration has also been argued to be an effective solution to a lifelong explore–exploit tradeoff, since earlier discoveries can be exploited for longer (Gopnik, 2020). Program induction 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 model can reproduce a characteristic developmental transition from grasping a few small numbers to discovering a recursive concept of real numbers. We note that an important part of constructivism is the idea that we *cache* the useful concepts we invent (cf. Zhao, Bramley, & Lucas, 2022), meaning our conceptual library grows as we do, becoming richer and more powerful for solving the tasks we repeatedly face. We do not attempt to model this important aspect of constructivism in this paper but return to it in the General Discussion.

Differences between childlike and adultlike inductive inference 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 leaps and decrease over time, as in *simulated annealing* (Van Laarhoven & Aarts, 1987). However, a high diversity of guesses might also reflect that children have a rationally flatter latent prior than adults, inherently entertaining a wider range of hypotheses at the cost of entertaining 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, or less powerful primitive concepts to work with, children's hypotheses might more directly *describe* encountered patterns, while adults might rely more on their existing knowledge hierarchy to constrain hypothesis generation 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 these ideas.

Task

In order to study inductive learning, 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∈ {small, medium, large} and color∈ {red, green, blue} and continuous scene-specific $x \in (0,8)$, $y \in (0,6)$ positions and orientations$\in(0,2\pi)$. 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 extremely 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 in the first place. As with the 2-4-6 task, there are always infinite data-consistent possibilities and while the bulk of these may be outlandishly complex, many others may still be simpler or more salient than the ground truth. Without carefully gathered evidence with broad coverage of the space of possible scenes, a learner will frequently be unable to rule out simpler 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.[1]

We use mixed-methods (Johnson, Onwuegbuzie, & Turner, 2007), analyzing both qualitative data in the form of freely generated guesses about the symbolic rules and 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 (Church, 1932) 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

[1] 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).

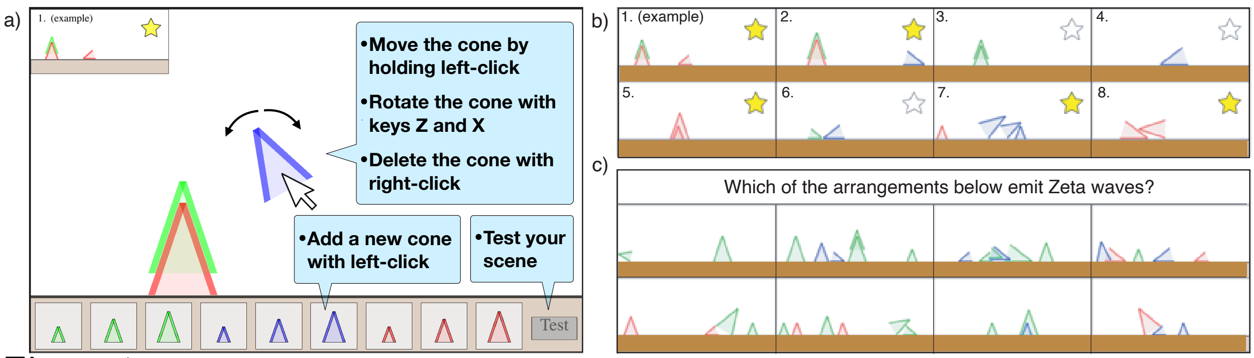


Figure 1

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.

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 of the evidence they have. Since we might expect evaluation of evidence–hypothesis consistency to be more error-prone in children, we expect adults' guesses to be more strictly consistent with their evidence. 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 differ in this dimension.

Context-free hypothesis generation

In examining children's and adults' inferences, we start by laying out a "top-down first" approach to hypothesis generation, utilizing a probabilistic context-free grammar (PCFG) to define and draw from a latent prior over concepts expressible in first order logic. A PCFG is a collection of "construction rules" that, when run repeatedly, stochastically create 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 of the scenes that the participant has tested. 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"—sampled hypotheses can be 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). Thus, one way to think of the PCFG is as a *computational level* characterization of the problem of inductive inference. However, we will argue that the generative mechanism at the heart of of the PCFG framework also elucidates important mechanistic considerations and provides the representational framework needed to ground algorithmic approximations that depart from this ideal and reflect core constructivist ideas.

At the computational level, different PCFGs, containing different primitives and expansions, can be compared against human behavior. And the probabilities for the productions in a PCFG can be fit to maximize correspondence with human judgments. In this way, recent work has attempted to infer the "logical primitives of thought" (Goodman et al., 2008; Piantadosi et al., 2016). Here we consider a single expressive PCFG architecture and examine its behavior under limited sampling. We examine its behavior with uniform production weights but also with weights engineered to produce the characteristics of "childlike' and "adultlike" symbolic guesses in our task. Crucially, under all these 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.

While naively, we might expect children to entertain simpler concepts than adults, this induction framework tends to predict the reverse. If we assume we start life at our most flexible, or "programable" (Turing, 2009), this would be like being born with concept building mechanism that is initially "untuned", growing its concepts essentially through blind mutation (Campbell, 1960) where each forking path on the road to a complete concept starts out equiprobable. However as a learner gathers a lifetime of experience, we would expect these construction weights to become tuned so as to favor certain elements or features that have proven useful in the past. A uniform-weighted PCFG hypothesis generator will thus tend to produce greater diversity than a more fine-tuned one. As such, it embodies the idea that more elaborately or implausibly structured, or "weird", concepts will come to the minds of children than adults.

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 also be tested against data. Unfortunately, in our task—and perhaps even more so outside of it—the vast majority a priori generated concepts are likely to be inconsistent with whatever evidence a learner has already encountered.[2] For this reason,

[2] In our task, many more 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"), or physically impossible ("Two (different) objects have the same

the procedure is astronomically inefficient, requiring very large numbers of samples in order to reliably generate non-trivial rules. One can also use a PCFG to adapt existing hypotheses, for instance using a Markov Chain Monte Carlo scheme in which parts of a hypothesis are regrown and accepted according to their fit to evidence (cf. Fränken et al., 2022; Goodman et al., 2008). While we think this approach is promising we do not model this here, and simply return to it in the general discussion. However, we do additionally consider an alternative to the PCFG, that provides a more sample efficient and, on the face of it, more cognitively plausible mechanism for initializing new hypotheses.

Context-based hypothesis generation

Instance Driven Generation (IDG) (Bramley et al., 2018) is a recent proposal related to the PCFG framework but with a key difference. Rather than generating initial hypotheses prior to, or blind to the current evidence, the IDG generates ideas *inspired* by encountered patterns (cf. Michalski, 1969), thus incorporating bottom-up reactivity to evidence into its conceptualization process. Each IDG hypothesis starts with an observation of features of one or several 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, an IDG does not begin each learning problem with a prior over all possible concepts, but rather draws its initial ideas from a restricted space consistent with the extant patterns in a focal observation. Figure 2b illustrates this approach. While a regular PCFG effectively starts at the top level (i.e. outermost nesting) of a compound concept and works downward and inward, the IDG starts from the central content (drawn from its observation) and works upward and outward to a quantified statement, ensuring at each step that the statement is true of the scene. The result is a mechanism that uses its concept grammar to describe features and patterns in evidence. This means that the IDG does not entertain hypotheses that are possible but never exemplified by a scene. For example, “at most five reds” would only be generated if a learner actually saw a rule-following scene containing five reds. A key prediction of the IDG is an interaction between the scenes generated by the participant and the hypotheses these subsequently inspire, with simpler scenes, embodying fewer extraneous or coincidental patterns being more likely to inspire the learner to generate the true concepts.

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”).

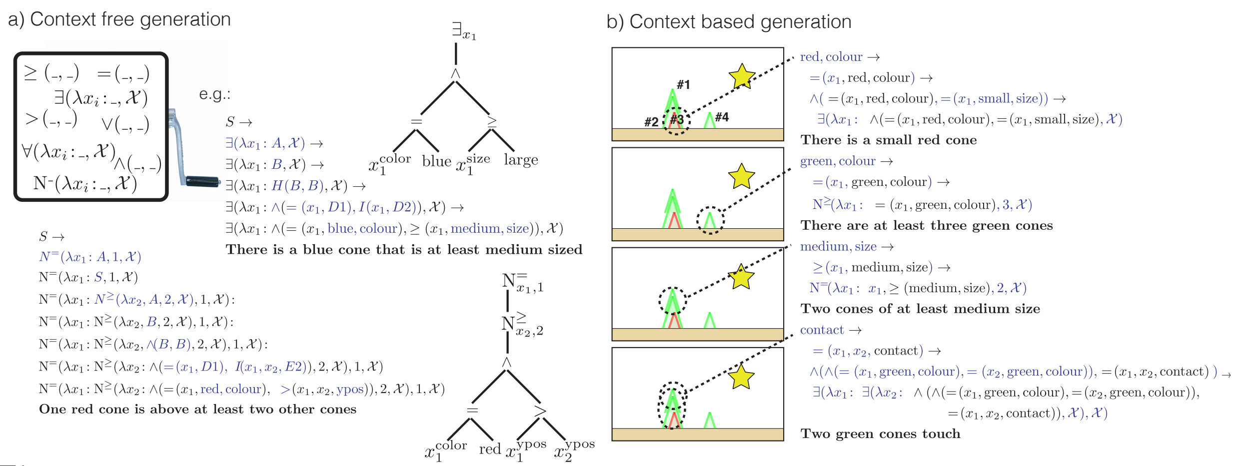


Figure 2

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.

Hypothesis-driven scene generation

Uncertainty-driven learning

Normatively, test scenes should serve to minimize expected uncertainty across the full hypothesis space. A direct way to approximate this here is to start with a prior sample of hypotheses (e.g. drawn context-free) and progressively create scenes that serve to minimize expected uncertainty over this sample by forking their predictions (Bramley et al., 2022; Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014). We visualize this in Figure 3a, imagining three labelled scenes $d_1 \ldots d_3$ that progressively divide a prior sample of hypotheses (hs) until a most-likely candidate emerges. The constructivist setting presents a challenge for this norm since the hypothesis space is latent and is initially unexplored.

Exploration-driven learning

An alternative hypothesis-free approach might be to explore the data space directly, for instance generating scenes that vary in the number and nature of objects they contain in the hope of naturally uncovering concept boundaries and inspiring hypothesis generation. We sketch this in Figure 3b. Efficient uncertainty-driven and exploration-driven learning both predict generation of scenes that differ substantially from one another, ideally being anti-correlated so as to cover the space efficiently (Osborne et al., 2012). However this does not seem well matched to constructism, wehere we rather think of the learner as entertaining a small but not completely empty set of possibilities

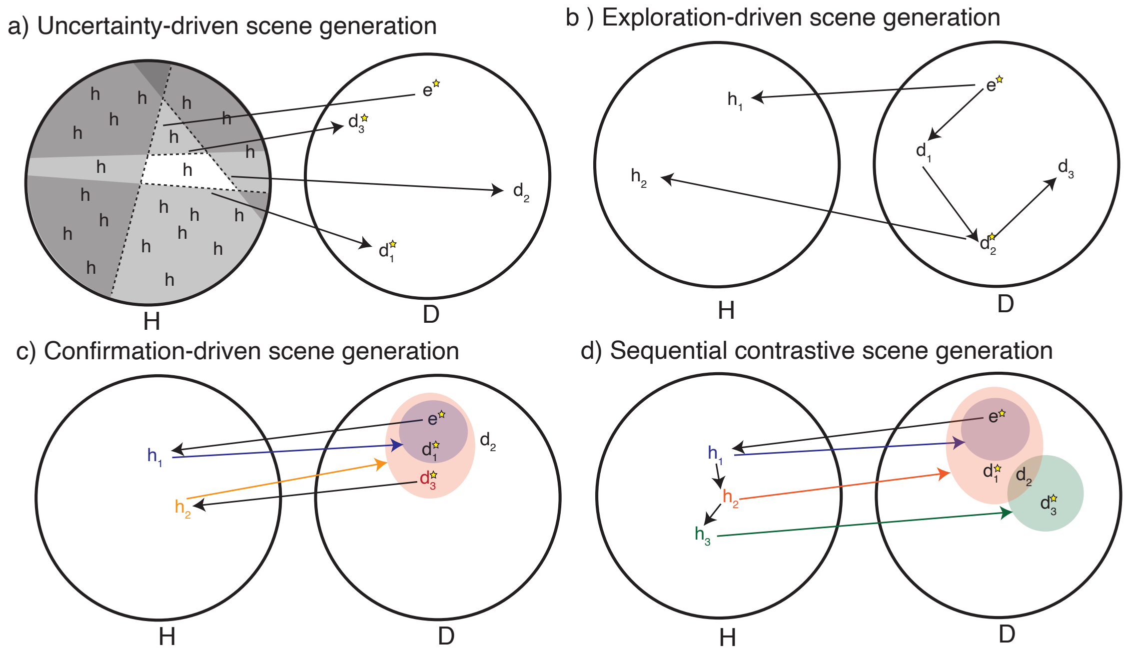


Figure 3

Aactive learning strategies: H = latent hypothesis space D = data space. Arrows indicate direction of inferences. Stars indicate scenes that followed the rule. a) Uncertainty-driven tests over prior sample $h \in H$. Dotted lines separate hypotheses by outcomes they predict for initial example e and self-generated scenes $d_1 \dots d_3$. Shading indicates which hs mis-predict each outcome. b) Exploration-driven testing. Scenes selected to explore D without regard to H. Outcomes may then inspire hypotheses. c) Confirmatory testing: Example e inspires hypothesis h_1. Scenes then test its generalization predictions. Colored circles visualize space of scenes for which each hypothesis predicts outcome will be produced. d_1 and d_2 are correctly predicted as rule following. d_3 is mispredicted by h_1 in producing the outcome, leading to a new h_2. d) Sequential contrastive testing: e inspires h_1 and h_1 inspires h_2, d_1 contrasts these leading to rejection of h_1. h_2 then inspires h_3 and d_2 contrasts these, etc.

and hence unable to capitalize on such diverse evidence.

A constructivist way to think of active learning is as acting in ways that challenge one's current hypotheses and so facilitate their refinement or the construction of better alternatives. We sketch two such approaches: Confirmatory testing and Sequential Contrastive testing.

Confirmatory testing

With a candidate hypothesis in mind, a learner can seek to challenge it through its generalizations (Nickerson, 1998; Popper, 1959). For example, after encountering the scene in row 1 of Table 1, a learner might generate the initial hypothesis that "there must be a small red" (since this describes one of the objects). To confirm this, they might try a positive generalization test, i.e. keep the small red but remove or randomize the other objects and predict the effect will still occur (e.g. d_1 in Figure 3c). Alternatively they might use it to predict a way to minimally alter d_1 so it no longer produces the effect, removing the small red and keeping the rest (e.g. d_2). So long as the learner gets the outcome they anticipate, they can stick with their hypothesis. When they don't they can either abandon or adapt it. For instance, d_3 in Figure 3c proves inconsistent with h_1, requiring a new hypothesis be generated that can explain why d_1 and d_3 produce the effect but not d_2. A limitation of a one-hypothesis-at-a-time approach is that it is unclear how distinctive the hypothesis's generalization predictions are.[3] For example, since the ground truth in this example is just "there is a red", producing new scenes containing small reds will fail to reveal that the redness but not the smallness is causative of the label. Another limitation is that it is unclear what to do when one's hypothesis is ruled out, especially if the scene if the test that differs dramatically from the ones with which it is consistent. For this reason, the education literature has long emphasized the utility of a "*control of variables*" strategy (Chen & Klahr, 1999; Klahr, Fay, & Dunbar, 1993; Klahr, Zimmerman, & Jirout, 2011). This amounts to manipulating exactly one design variable per test, such that any difference in the outcome is straightforwardly attributable to the change in the input providing a route to adapting one's hypothesis when it fails.

Sequential contrastive testing

A related scheme that might allow a constructivist learner to escape some pathologies of confirmatory testing is the *iterative counterfactual strategy* described in Oaksford and Chater (1994). That is, learners might first generate an *alternative hypothesis* h_2 by inverting some feature of their initial hypothesis and then focus their next test on separating h_1 from h_2 (e.g., Figure 3d).[4] For example, starting with h_1:"there is a small red", one local alternative would be to drop the the mention of size, leading to h_2: "There is a red". Now the learner has a pair of hypotheses and a recipe distinguishing between them: Testing a scene containing a red object that is not small (e.g. d_1). This could again be easily achieved by adapting the original scene, so the small red is a different

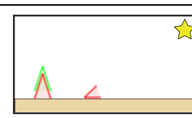
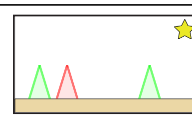
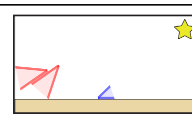
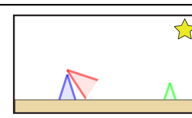
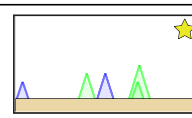
[3] A general finding is that positive confirmatory tests are valuable to the extent that the outcome of interest is rare, e.g. if most scenes are not rule following. This is not generally the case in this task.

[4] In Oaksford and Chater's (1994) formulation, the complementary hypothesis is then inconsistent with the scene that inspired the original hypothesis, such as going from "increasing by two" (inspired by seeing 2-4-6) to "decreasing by two" such that its falsification may be mistaken for confirmation of the original hypothesis. Here there are many ways to flip the content of a hypothesis both with or without rendering it inconsistent with a scene that inspired it.

size (Chen & Klahr, 1999; Klahr et al., 1993, 2011).If d_2 produces the effect, h_1 can be supplanted with h_2. Otherwise h_2 can be rejected and a new h_3 can be generated. Either way, this approach facilitates constructivism by providing a direction of travel however a test comes out, so allowing a constructivist learner to explore both the data and hypothesis spaces in parallel (Klahr & Dunbar, 1988).

As illustrated in Figure 3, what constructivism-compatible hypothesis-driven approaches have in common is a prediction of anchoring in data space: Each new scene shares features with the scene that inspired the earlier hypotheses that inspired it. This contrasts with the pattern we would expect if participants followed a normative uncertainty-driven approach or model-free exploration-driven approach since both tend to predict each scene should be as different as possible to earlier ones (although see Navarro & Perfors, 2011, for how this depends on the structure of the hypothesis space). While we do not collect the trial-by-trial guesses we would need to distinguish between all the accounts we mention, we will look for an empirical signature of constructivist active learning, in the form of anchored, incremental and systematic testing patterns and assess whether these differ between children and adults.

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 $\exists(\lambda x_1 : \; \wedge(=(x_1, \text{blue}, \text{color}), =(x_1, 1, \text{size})), \mathcal{X})$	

Overview

In summary, the main goal of this paper is a close investigation of developmental differences in active open-ended hypothesis generation examined through the lens of a constructivism-inspired rational-process framework that puts stochastic generation and incremental search at the center of the individuals' learning. To foreshadow, we find that children make more complex guesses about the hidden rule that are only a marginally worse fit to the evidence than adults' guesses. Children also create more complex learning data than adults but do so less systematically. We then show that both children's and adults' guesses reflect an evidence-inspired process of compositional concept formation as modeled by our Instance Driven Generation algorithm over a top-down–first PCFG norm, capturing that their guesses are inspired by discovery of patterns in their learning data. We show these behavioural patterns are a natural result of children having a less fine-tuned concept generation mechanism. Crucially, we also show that both children's and adults' symbolic guesses causally drive their generalizations, as opposed to these being driven by surface feature resemblance as emphasized in statistical views of concepts (cf. Medin & Schaffer, 1978; Posner & Keele, 1968). Finally, we show that both children's and adults' create scenes by adapting earlier scenes, which we argue is consistent with confirmatory or iterative counterfactual testing rather than uncertainty- or exploration-driven testing.

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.[5] 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`.

[5] 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.

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. The test scenes and their labels remained visible on the screen throughout the Learning and Question phases.

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 before comparing the features of the scenes produced by adults and children. We will then turn to a series of model-based analyses that attempt to reproduce participants distributions of free guesses, generalizations and scenes within the constructivist framework.

Guess complexity and constituents

We had human coders translate participants' free text guesses about the hidden rule wherever possible into an equivalent logical 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 primarily reveals that children's encoded rules were substantially *more complex* than those generated by adults and that both were substantially 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 4). As an example, one child's rule for problem 1 was *"You must have two reds and one blue"* which was translated to $N^{=}(\lambda x_1 :\ N^{=}(\lambda x_2 :\ (\wedge(=(x_1, \text{red}, \text{color}), =(x_2, \text{blue}, \text{color})), 1, \mathcal{X}), 2, \mathcal{X})$, 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 (4c), 2.06 booleans (4d), 1.55 equalities and inequalities (4e), referred to 1.39 different primary features (color, size, orientation, x- or y-position, groundedness, 4f) and 0.37 relational features (contact, stackedness, pointing, or insideness, 4g). 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. Children thus used significantly more quantification (i.e. referred to more separate entities) $t(102) = 3.98, p < .0001$, more booleans $t(102) = 3.59, p < .0001$ and relational features $t(102) = 3.12, p < .002$ than adults, but the agegroups did not differ significantly in mentions of (in)equalities $t(102) = -0.05, p = 0.96$ and references to the objects' basic features $t(102) = -.91, p = .36$. When children posited that an "at least", "at most" or "exactly" a certain number of objects must have certain features, the number they chose was substantially higher than that for adults (2.36 compared to 1.58, $t(68) = 3.72, p = 0.0004$). In terms of features, adults frequently gave rules relating to color (58% compared to 39% of children's rules, $t(102) = 2.27, p = 0.025$), while children were more likely to refer to positional properties (26% compared to 18% of adults' rules $t(102) = 2.15, p = 0.034$).

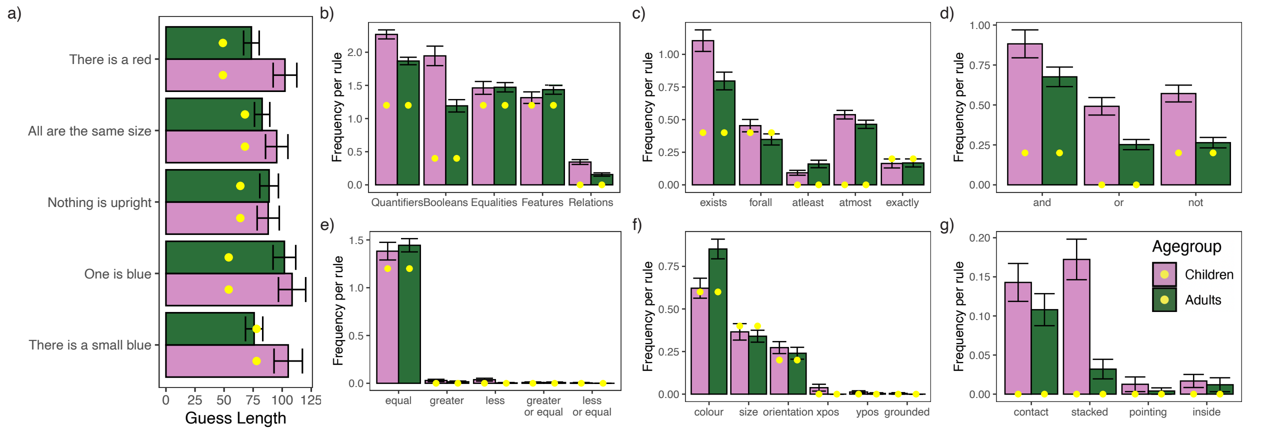


Figure 4

(a) Length of Children's and Adults' rule guesses. (b) Relative frequency of rule elements in logic coded versions of these rules, c–g with respect to quantifiers, booleans, (in)equalities, basic and relational features respectively. Error bars show normal 95% confidence intervals. Yellow points in a show ground truth frequency.

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.

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 5a). 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.[6] Thus, it

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

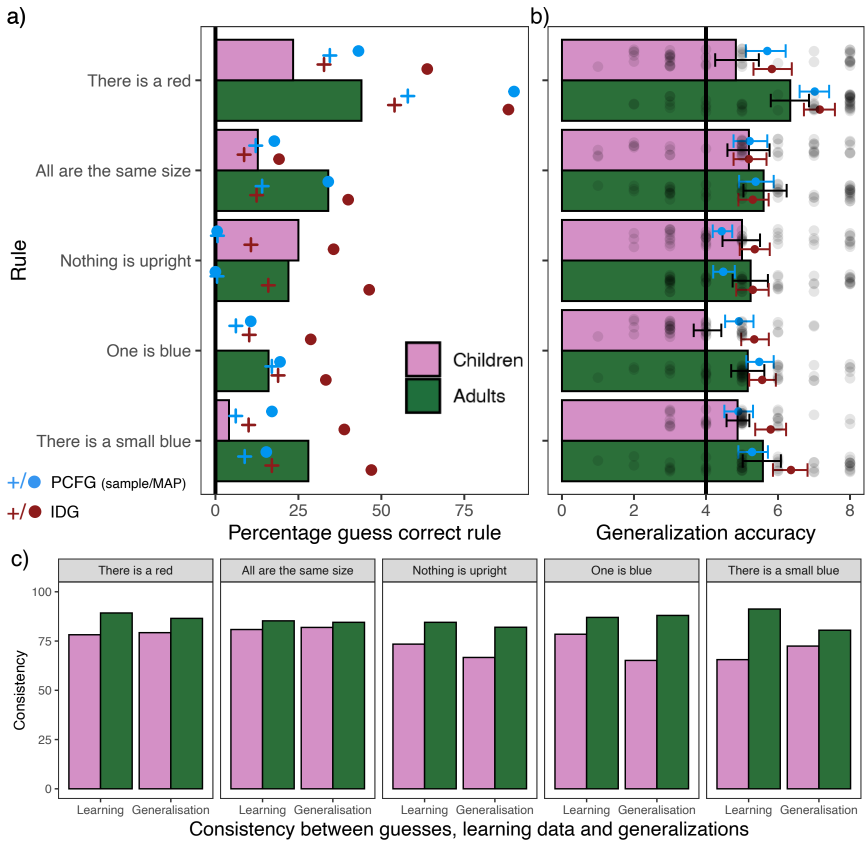


Figure 5

a) Percentage children and adults guessing correct rule. b) Generalization accuracy. Bars show mean ± bootstrapped 95% CIs. In a–b, Black vertical lines denote chance performance. Blue and red points show performance of simulated PCFG and IDG learners as described in Modeling section. Circles = guessing the MAP rule or MAP generalization (after marginalizing over posterior). "+" shows accuracy of a single posterior sample. Both models here use agegroup-consistent production weights, CIs show bootstrapped 95% confidence intervals. c) Consistency between subjects' rule guess and their (self-generated) learning data, and generalizations.

is instructive to ask whether participants' rules, where not exactly correct, are nevertheless consistent with the evidence they gathered.

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 5c for a by-rule breakdown).

Generalizations. We now report participants performance in predicting which of 8 new scenes will produce stars (i.e. follow each hidden rule). Across the five tasks, both children and adults guessed more accurately than chance (50%): *children* mean$\pm SD$ $59\% \pm 11\%, t(53) = 5.9, p < .001$; *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 5b).

Scene generation

As well as generating more complex rules, children tended to create more complex test scenes than adults. The average child-generated scene contained 3.7±0.88 objects (close to the average in the example scenes) compared to 2.8±0.57 objects for adults ($t(102) = 5.8, p < .001$). The complexity of a learner's test scenes was inversely related to their performance overall ($F(1, 102) = 39.0, \beta = -0.08, \eta^2 = .28, p < .001$) and also within both the children ($F(1, 52) =, \beta = -0.056, \eta^2 = .20, p < .001$) and adults ($F(1, 49) = 9.1, \beta = -0.096, \eta^2 = .16, p < .001$) taken individually (see Figure 6a). Within the children, age was inversely associated with scene complexity, with an average of 0.35 fewer objects per scene for each additional year $F(1, 52) = 12.6, \eta^2 = .19, p < .001$. Aside from this difference, we also assess whether children's or adults' scenes bear the hallmarks of being driven by confirming or distinguishing between a small set of possible rules.

If participants do follow a control of variables, confirmatory, or iterative counterfactual approach, we would expect the scenes generated by participants to be more similar to the initial example or one of their own preceding scenes, than to a random scene or a scene drawn from a different learning problem. If they are rather maximising information with respect to a larger set of hypotheses, or exploring the data space efficiently, we would expect the opposite pattern of indpenendence or anticorrelation. To explore this, we constructed a distance metric that we used to measure the feature–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

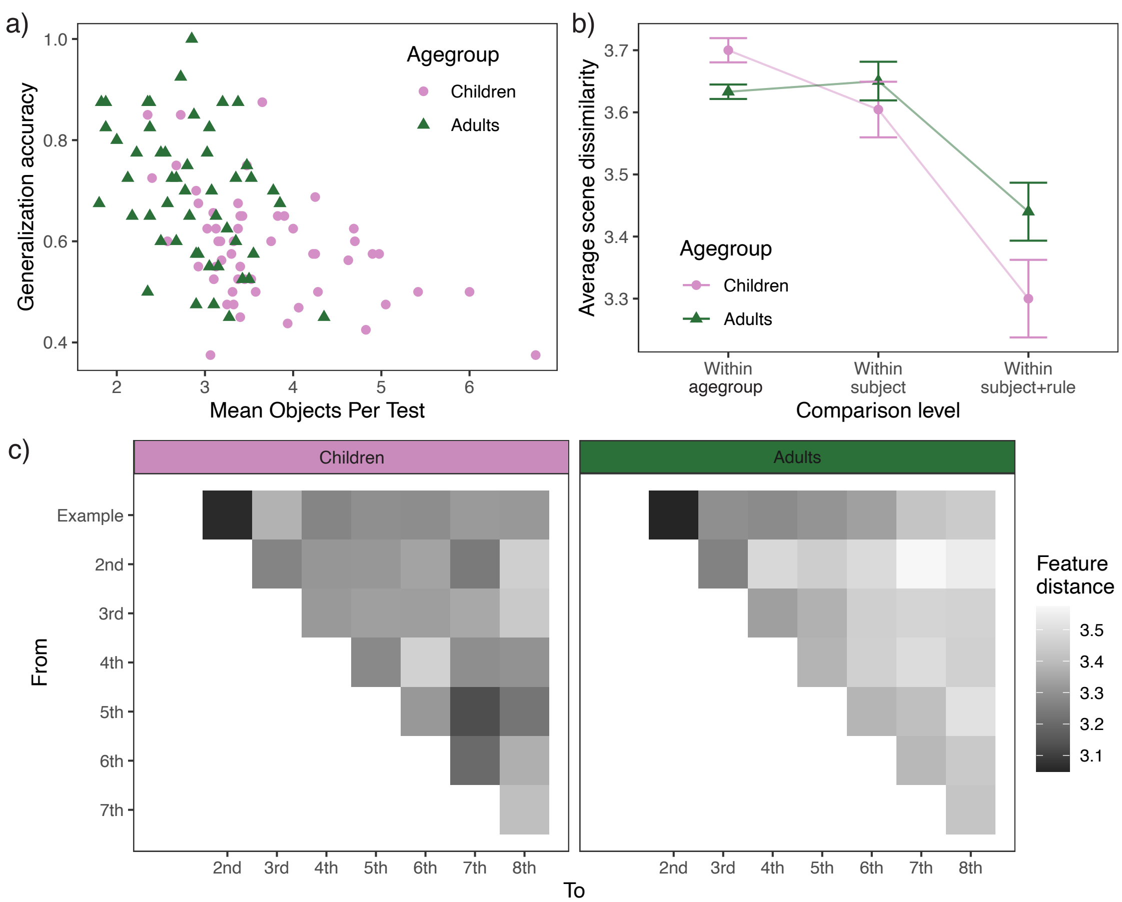


Figure 6

(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.

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$= 3.67 \pm 0.94$. 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 6b&c).

Figure 6c breaks down the within-trial scene dissimilarity by test position for the two agegroups. 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 also look 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.

Interim Discussion

In sum, in our experiment we found children were only moderately less able to come up with rules that fit the evidence than adults and there were only moderate differences in the compatibility between children's and adults' rules and their subsequent generalizations. Most striking was the fact that children's guesses appeared to overfit the evidence more, producing more complex, perhaps more 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 more different objects and more complex properties of the learning scenes that were actually irrelevant to their label. As well as generating more complex concepts, children created more complex test scenes that appeared to be more repetitive overall, yet also appeared to be varied less systematically than adults'.

Model comparison

To explore the basis for the diversity of guesses and generalizations, and of the differences between children and adults' learning, we now turn to model-based characterization of the behavioral data. We focus first on the guesses, then the generalizations, and finally the scene creation. We will assess whether participants guess and generalization patterns are better captured by Bayesian inference over samples from an expressive latent prior—Probabilistic Context Free Generation (PCFG)—or rather by the partially bottom-up generation—Instance Driven Generation (IDG) limited to hypotheses inspired by patterns in scenes (Bramley et al., 2018). We then assess whether new scenes are better captured as independently generated—consistent with uncertainty-driven or exploration-driven testing—or as adaptations of earlier scenes— consistent with confirmatory or iterative contrastive testing.

To foreshadow, we find convergent evidence that both children's and adults' guesses are better accounted for by Instance Driven Generation (IDG) of hypotheses than by an approximately normative Probabilistic Context Free Grammar (PCFG) norm. We then demonstrate that neither children's nor adults' generalizations can be explained by surface similarity between rule-following and generalization probe scenes, but that they are well predicted by the learners' own symbolic guess. Finally, we show that almost all children's and adults' scenes are more likely to have been created by making simplifications and edits to either the previous or the initial scene—in line with hypothesis-driven confirmatory or contrastive testing—rather than being generated independently from scratch—consistent with uncertainty-driven or direct exploration of the data space.

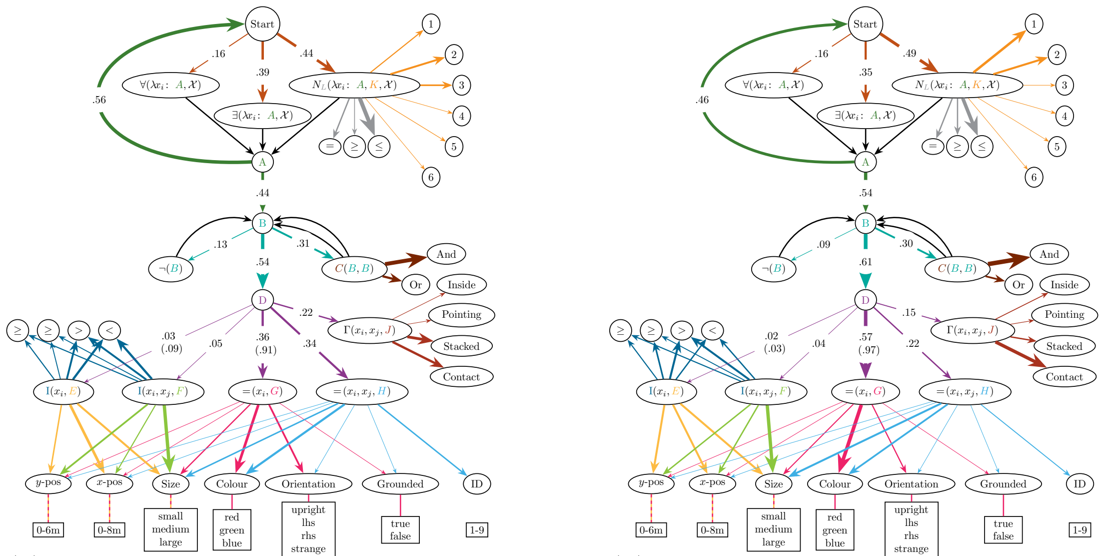
Guesses

Participants produced a huge variety of guesses but despite this, these guesses were consistent with the majority of their evidence. Children's guesses were more complex and a little less data-consistent on average than adults'. We now explore using PCFG and IDG sampling to produce similar guesses.

We first assume a PCFG as a computational level framework and reverse engineer what production weights it requires to generate the kinds of guesses we see adults and children make. Next, we contrast the prior sample-based PCFG approach to rule generation with our proposed data-inspired IDG, showing that the IDG does a better job of capturing participants' accuracy by problem type and agegroup and is also better able to produce the specific guesses made by the participants.

Reverse engineering Childlike and Adultlike production weights

Having encoded all the rule guesses from adults and children (in the section on *Rule complexity and constituents*), we created PCFG production weights that produce similar guesses as adults and children. To do this, we worked back from the observed counts for each rule element doing this separately for children's and for adults' guesses (see Appendix A). Of course, the guesses are samples from a range of different participants' posteriors, since guesses were always based on some evidence. However, since this evidence differs dramatically between trials and 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, this still provides a helpful elucidation of generation differences behind child-like and adult-like guesses. A full set of fitted prior weights for both adults and children are visualized in Figure 7. This analysis simply demonstrates that a natural way to understand children's guesses are as emanating from a less fine-tuned generation mechanism adults', with flatter, more entropic branching at 12 of the 14 forking production steps we assumed in our PCFG model. Indeed probability distibution over productions at each stage averaged 1.28 ± 0.50 bits for children compared to 1.03 ± 0.59 bits for adults, $t(13) = 3.2, p = 0.007$.



(a) *Childlike PCFG* (b) *Adultlike PCFG*

Figure 7

Visualization of (a) child-like and (b) adult-like PCFGs, 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.

Modeling accuracy by participant and rule

We now compare participants patterns of accuracy to simulated approximately normative inference over a PCFG-generated sample and IDG hypothesis generation algorithms provided with the active learning data generated by the human participants. We generated a sample of 10,000 hypotheses based on uniform production weights $\hat{H}_{\text{PCFGu}}$, and similarly for the IDG generated a sample based on uniform productions for each task $\hat{H}^{p,t}_{\text{IDGu}}$. Additionally, for each participant p—and separately for each learning task t in the case of the IDG—we generated another 10,000 possible rules using age-consistent prior production weights derived above $\hat{H}^{p}_{\text{PCFGh}}$ and $\hat{H}^{p,t}_{\text{IDGh}}$ that have statistics matched to those in Figure 4a–f.[7] The PCFG samples act as an approximation to an infinite latent prior over rules $P(h)$ before seeing any data. The uniform-weight PCFG samples capture a generic inductive bias for simpler hypotheses while fitted held-out child- and adult-like weights

[7] For these, we held out the subjects own guesses when setting the weights to avoid double dipping the data.

additionally attempt to capture "learned" inductive biases common to the requisite age-group (but not specific to the participant). The IDG samples are additionally idiosyncratically constrained in the sense of only reflecting rules referring to features or relations actually present in at least one of the learning scenes. We split the IDG sample evenly across tests 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 weighted these samples by their likelihood of producing all eight scene labels l observed during the learning phase

$$P(h|\mathbf{l};\mathbf{d}) \propto P(\mathbf{l}|h;\mathbf{d})P(h) \tag{1}$$

$$\approx P(\mathbf{l}|h;\mathbf{d}) \sum_{\hat{h}\in\hat{H}} \mathbb{I}(h = \mathrm{h}) \tag{2}$$

and combined 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 1000 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

$$P(l = 1|h;\mathbf{d}) \propto \exp(-b \times N_{\text{mispredictions}}) \tag{3}$$

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 $\propto 1$ for rules that predict all 8 scenes' labels, to $\propto .13$ for a single misprediction, and $\propto .02$ 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".

The resulting model guess accuracy is shown visualized in Figure 5a. We distinguish between two possible decision mechanisms: (1) Taking the *maximum a posteriori* (MAP) estimate from a large posterior sample (guessing in the event of ties), which we take as closer to a normative ideal and (2) taking the accuracy of a single posterior sample, which we take to be more consistent with the best-case-scenario output of a process in which a given learner searches over hypotheses driven by a combination of prior complexity and fit. Under all models, the MAP lines up with the correct hypothesis more often than participants do (15–37% based on children's active learning and 20–51% based on adults', recalling that children guessed correctly of 11% of trials and adults on 28% of trials). For instance, under a uniform-weighted prior sample, the PCFG MAP is correct on 15% of all children's trials and 20% of all 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 when we use samples based on reverse-engineered weights that reproduce the qualitative properties of other participants in the same agegroup (see Appendix A and Figure 7). For age-appropriate prior samples, the PCFG guesses correctly on 18% of children's trials and 32% of adults' trials. Using an age-inappropriate "flipped" prior sample (i.e. child-like weights for adults and adult-like weights for children) obliterates this difference, resulting in 23% for children and 22% for adults. We see a similar pattern for the IDG algorithm, but higher accuracy across the board. The IDG achieves the best accuracy on both children's and adults' trials, guessing over half of the hidden rules correctly (51%) in the case of adults' trials. However, achieving this level 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 5a).

To check what provides the better account of participants trial-by-trial accuracy patterns we fit logistic mixed-effect regression models using the response under each algorithm and prior combination to predict each participant's by-task probability of guessing correctly, including random effects for both rule type and participant. For the maximization models, we softmaxed the posterior with a low "temperature" parameter ($\tau = 1/500$, Luce, 1959), meaning predictions were close to 1 or 0 excepting where multiple hypotheses were tied, where they were close to $1/N$ for the N tied hypotheses. The "Fit" columns of Table 2 shows the log likelihood for each of these models, revealing that participants' correct judgments most in line with posterior sampling under an IDG prior, with age-appropriate production weights (log likelihood = 211.5, $\beta = 5.44 \pm 1.74, Z = 5.99, p < .001$) improving over a baseline fit of -234.3 for a model with only intercept and random effects.

Table 2

Accuracy of Rule Guesses by Simulation Models

		Accuracy MAP (%)			Accuracy Posterior Sample (%)		
Algorithm	Prior	Children's data	Adults' data	Fit	Children's data	Adults' data	Fit
PCFG	Uniform	14 ± 16	20 ± 14	-229	9±5	12±5	-226
PCFG	Agegroup	17 ± 17	32 ± 15	-230	11±7	20±7	-225
PCFG	Flipped	22 ± 20	22 ± 15	-231	15±9	15±6	-229
IDG	Uniform	26 ± 22	39 ± 21	-226	9±5	14±6	-217
IDG	**Agegroup**	36 ± 25	51 ± 18	-226	**14 ± 8**	**24 ± 8**	**-212**
IDG	Flipped	26 ± 20	52 ± 18	-230	13±8	23±8	-223

"Children" and "Adults" columns show the $M \pm SD\%$ by-subject accuracy of the requisite algorithm. "Fit" shows the log likelihood for a logistic mixed-effects regression using model accuracy to predict if the participant guesses correctly on each trial.

Modeling rule guess

As a more direct test of the constructivist PCFG and IDG models' ability to explain participants' free response guesses, we also attempted to estimate the probability of each approach generating exactly the participant's encoded guess based on their active learning data.

By definition, all 87% of trials in which participant gave an unambiguous rule, we were able to encode in our concept grammar, so all have nonzero support under a PCFG prior. Due to the stochasticity we assumed in our likelihood function, all possibilities also nonzero have posterior probability, meaning they are guaranteed to appear in a sufficiently large PCFG sample.[8] 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 rules that were generated at least once in 10,000 samples with agegroup fitted weights was highest for the IDG with fitted weights (69% for children 76% for adults), decreasing to 49% and 62% using uniform weights. This was still higher than for the PCFG which generated 42% for children's and 53% for adults' guesses with the fitted prior weights and 45% for children's and 50% for adults' rules from a uniform prior.

Table 3 details model fits to participants' guesses. The IDG is again the stronger hypothesis generation candidate, assigning higher probabilities on average to the rules that

[8] They would not necessarily appear in an infinitely large IDG sample because many of the more complex concepts are merely possible without being positively present. For example "there is a red and fewer than five small blues" is consistent with the Figure 1b but would never be generated by the IDG procedure inspired by these scenes.

Table 3

Model Probability of Producing Participants' Exact 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**

Note: N best columns show the number of participants in each agegroup best fit by each model.

participants provided. As expected, the variants of the PCFG and IDG with agegroup-consistent production weights were 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 8a 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.

To better understand why we were not able to generate all of participants guesses, we also examined those frequently generated by the models and contrasted these with those never generated under any of our model variants. Table 4 shows two examples of each for children and adults and the full set is available in the Online Repository. Unsurprisingly, the participant guesses our models failed to generate tended to have more complex forms and a concomitantly low generation probability. Assuming uniform weights, the syntax of the children's guesses that we did generate had marginally higher log prior generation probabilities Median (Inter-Quartile Range) -10.2 (5.0) than those we didn't were unable to generate -13.9 (16.31) (Mood's median test, $Z = 1.9$, $p = 0.053$). For adults this difference was more pronounced -9.9 (5.0) compared to -14.9 (14.0) (Mood's median test, $Z = 4.5, p =< .001$).[9] This examination revealed that one class of rules our participants

[9] Note that these prior generation probabilities are a lower bound on the chance of of generating a particular semantic rule since many syntactic forms can express the same semantic content (Fränken et al., 2022). This captures why some relatively frequently generated semantic classes of guess nevertheless had a low probability for each specific syntactic expression .

Table 4
Example Guesses

Agegroup	Rule	Example syntax	log Prior Uniform	log Prior Age-group	log(Likelihood)	N/10k
Children	*"One is on top of the other"*	$\exists(\lambda x_1 : \exists(\lambda x_2 : \Gamma(x_1, x_2, \text{stacked}), \mathcal{X}), \mathcal{X})$	-9.5	-8.4	0	117
Children	*"Only different colors*	$\forall(\lambda x_1 : \forall(\lambda x_2 : \vee(=(x_1, x_2, \text{ID}), \neg(=(x_1, x_2, \text{color}))), \mathcal{X}), \mathcal{X})$	-9.8	-8.0	0	260
Adults	*"If there are multiple small blocks."*	$N_{\geq}(\lambda x_1 := (x_1, 1, \text{size}), 2, \mathcal{X})$	-9.9	-19.6	0	609
Adults	*"There is at least one small green triangle."*	$\exists(\lambda x_1 : \wedge(=(x_1, \text{green}, \text{color}), =(x_1, 1, \text{size})), \mathcal{X})$	-13.8	-21.3	0	532
Children	*"They have to be with all three different colors"*	$\exists(\lambda x_1 : \exists(\lambda x_2 : \exists(\lambda x_3 : \wedge(\wedge(=(x_1, \text{red}, \text{color}), =(x_2, \text{green}, \text{color})), =(x_3, \text{blue}, \text{color})), \mathcal{X}), \mathcal{X}), \mathcal{X})$	-22.3	-16.6	-2.0	0
Children	*"There has to be one small blue piece and there has to be more than one piece"*	$\exists(\lambda x_1 : N_{\geq}(\lambda x_2 : \wedge(=(x_1, 1, \text{size}), =(x_1, \text{blue}, \text{color})), 2, \mathcal{X}), \mathcal{X})$	-12.5	-11.3	0	0
Adults	*"When there is a cone from each color of the same size"*	$\exists(\lambda x_1 : \exists(\lambda x_2 : \exists(\lambda x_3 : \wedge(\wedge(\wedge(\wedge(=(x_1, \text{red}, \text{color}), =(x_2, \text{green}, \text{color})), =(x_3, \text{blue}, \text{color})), =(x_1, x_2, \text{size})), =(x_1, x_3, \text{size})), \mathcal{X}), \mathcal{X}), \mathcal{X})$	-20.5	-11.11	-2.0	0
Adults	*"one piece has to be leaning on another"*	$\exists(\lambda x_1 : \exists(\lambda x_2 : \wedge(\Gamma(x_1, x_2, \text{contact}), \neg(=(x_2, \text{upright}, \text{orientation}))), \mathcal{X}), \mathcal{X})$	-18.5	-21.3	-3.9	0

Note N/10k shows how many times we generated this rule in 10,000 samples assuming agegroup-specific weights and counting any semantically equivalent expressions.

guessed but our models did not generate were those that could be expressed much concisely with more powerful logical grammar. For example, we saw a number of cases of universal quantification over feature values, such as "one of each color", mentioned in both a child and an adult guess in Table 4. This kind of rule can be expressed parsimoniously in second order logic with a single universal quantifier over color properties while in our grammar it required a separate quantification for each color. The fact that children produced about as many apparently higher-order-logic rules as adults seems to suggest that the PCFG we assumed, despite its ostensively complex structure, is still a simplification of the basis from which children constructed their ideas (cf. Piantadosi et al., 2016).

Generalizations

We next examine our models' ability to account for participant's generalization performance. As with the guesses, we first examine patterns of accuracy by comparing participants to simulated constructivist PCFG and IDG learner benchmarks before fitting a range of models to the specific generalizations participants made.

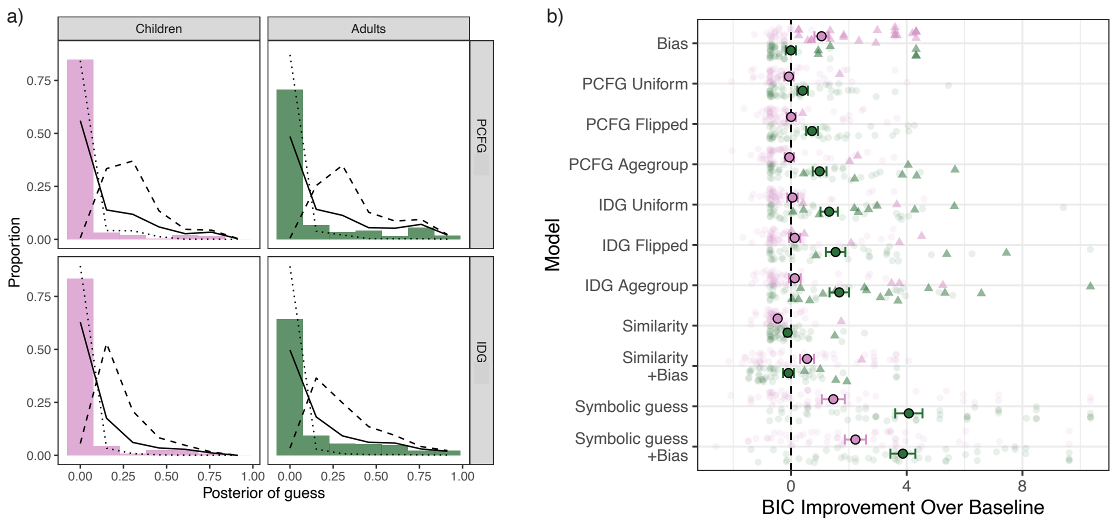


Figure 8

a) Posterior probability of participants' guesses under PCFG and IDG samples with agegroup weights. 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 (of of the 17 blind to the symbolic guess) is the best fit for that participant.

Modeling generalization accuracy

To do this, we use their requisite predictive distributions to model labelling generalizations $\mathbf{l}^*$ to the set of test scenes $\mathbf{d}^*$

$$P(\mathbf{l}^*|\mathbf{l};\mathbf{d},\mathbf{d}^*) = \int_H P(\mathbf{l}^*|H;\mathbf{d}^*)P(H|\mathbf{l};\mathbf{d})\; dH \tag{4}$$

$$\approx \sum_{h\in\hat{H}} P(\mathbf{l}^*|h;\mathbf{d}^*)P(h|\mathbf{l};\mathbf{d}) \tag{5}$$

Provided with the active learning data generated by the human participants, both performed in the human range at generalization. As with predicting the guesses, taking the marginally most likely generalization labels over a posterior weighted sample of agegroup-appropriate IDG prior productions performed best overall and reproduced the difference between children's and adults' generalization accuracies (68.8±20.1% and 74.2%±21.7%). The uniform-production IDG still 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'. Using agegroup-appropriate priors, the PCFG also reproduces the empirical difference between children's and adults' accuracy: 62.8±19.8% for children's trials and 68.8±20.9% for adults'trials. Using the PCFG with uniform production weights yielded accuracies of 61.4%±19.6% for children's and 63.5%±20% for adults' data.

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 inspiration 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 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.[10] This also reveals that building in human inductive biases boosts generalization performance (cf Lake et al., 2017) and the idea that adults' have formed stronger inductive biases than children goes some way to explain differences in how they generalize.

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.

Modeling specific generalizations

A standard benchmark for models of concept learning is a fit with participants' generalizations to new exemplars. Thus, we compared 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-based.

We fit a total of 18 models to the generalization data. All models had between 0

[10] It is likely that other approximate inference methods, such as an MCMC or greedy posterior search approach, could improve on this sample efficiency. However they also introduce other challenges for the learner (i.e. escaping local minima) and the modeler (getting good coverage of the response space and aggregating auto-correlated samples).

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 compared 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 $\in$ {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 $r = P_{\text{PCFG}}(\mathbf{l}* | \mathbf{l}; \mathbf{d}, \mathbf{d}*)$. "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 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 a logistic function of these scores $r = \frac{e^{\Delta\text{Similarity}}}{e^{\Delta\text{Similarity}}+1}$ 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 capture generalization patterns even though they are blind to the individual 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 8b. At the individual level, the PCFG+Bias and IDG+Bias models performed no better than the unbiased PCFG or IDG models, thus we omit these from Figure 8b for simplicity.

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.

If we restrict our comparison to models blind to the participant's symbolic guess then the IDG with the Agegroup-derived prior is the best fitting model for both children and adults. In this set, at the individual level, IDG Agegroup best fits the most adults (15/50), with 28/50 best fit by one of the IDG variants, compared to 6/50 by a PCFG variant and 5/50 by a Similarity model. The majority of children were better fit by Bias (25/54) or Baseline (13/54), but of the 16 individually better fit by one of the inference models, 11 were best captured by an IDG variant, 3 by a PCFG variant and 2 by a similarity variant (see triangles in Figure 8b and Appendix Table A-3).

Overall, children's generalizations were 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: with a clear hypothesis in mind, there is little rationale to select more or fewer than the generalization scenes consistent with that rule.

As with the free rule guesses, the IDG was robustly more aligned with participants' generalizations than the PCFG, particularly for adults, and particularly when using agegroup-appropriate weights rather than Uniform or age-inappropriate Flipped production weights. Thus, this model comparison 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 the IDG than our PCFG algorithm across all analyses.

These results support a key aspect of the constructivist framework, participant's idiosyncratic symbolic guesses seem to do the work in driving generalizations, rather than these being driven by family resemblance in the features of the scenes. The constructivist account anticipates that generalization patterns are dependent on what concept the learner has arrived at by the end of learning, and our end-to-end models of this process demonstrate the sheer breadth of concepts that learners can reasonably end up with in this task.

Scene generation

We finally turn to participants' scene generation. We compare participants generated scenes to several benchmarks before comparing a set of models of scene generation to test the idea that participants adapted earlier scenes to isolate and test the role of features mentioned in their hypotheses.

Comparison with information norms

According to an information gain analysis, children's and adults' scene generation result in some differences in the quality of the total evidence generated. 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 9). Relative to the agegroup-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, as we noted information-theoretic analyses as limited in what can reveal. It is predicated on an implausibly complete

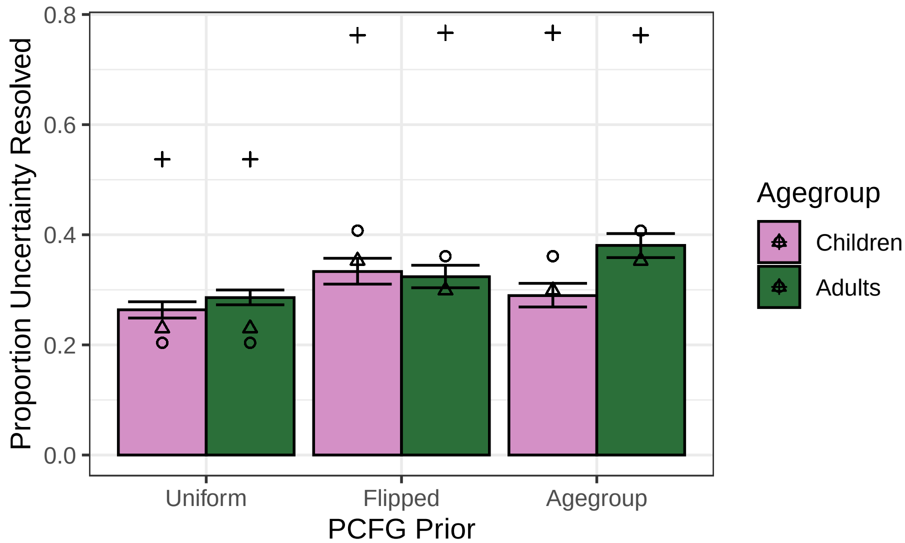


Figure 9

Uncertainty reduction under different priors. Triangles = random scene selection. Circles = greedy expected information maximizing scene selection. "+" symbols = Ideal teaching scenes.

representation of uncertainty that we approximated by using a large sample of prior hypotheses, while we have characterized constructivist learning as driven by more focal testing of a handful of similar possibilities.

We also compared participants against three scene selection benchmarks. In Figure 9, black triangles show the reduction in uncertainty resulting from supplementing the initial example with 7 scenes selected at random from from among participant generated scenes. Circles show the result of repeatedly selecting from a sample of 1000 of the participant-generated scenes, greedily selecting whichever one maximizes the expected information gain with respect to the prior at that test. Plus symbols show the reduction in uncertainty resulting from observing scenes selected by an ideal teacher—i.e. the seven scenes that, in combination with the initial example, best reveal the true concept.[11] One striking feature of these benchmarks is the low performance of the uncertainty-driven norm under all PCFG priors. Expected information gain slightly outperforms participants and random selection assuming the agegroup priors, but is actually worse than random scene selection under a flat uniform prior sample. This poor performance stems from the fact that the prior space of hypotheses is just so large and symmetric, making most scenes similarly informative at first. Furthermore, a large class of PCFG hypotheses predict that

[11] We selected these by generating 10,000 sets of seven scenes for each rule, and selecting the set that best reduced entropy.

all possible scenes will be rule following, or that all possible scenes will be non-rule following. These hypotheses are incorrect and rarely entertained by participants, yet have an outsized effect on the greedy selection of scenes that maximize expected information gain. Scenes selected to maximally convey each concept are far more informative, highlighting gulf between self-teaching and optimal teaching in inductive settings.

Figure 10 compares an example scene sequence selected by a child and an adult against a random selection from all participant scenes, uncertainty-driven selection and those selected to maximally convey the concept. This visual comparison highlights how human scene selection involves recognizable repetition and patterning that look quite unlike random and uncertainty-driven selection. In particular, several of the scenes selected to minimize expected uncertainty are very complex compared to participants' selections. Theoretically uncertainty driven scenes do an excellent job of dividing the hypothesis space, shown by their ceiling-level EIG (Figure 10f). However, since the target rule in this case turns out to be a simple, this sophistication does not benefit the uncertainty-driven learner overall (Figure 10g).

Models of scene selection

We hypothesized participants might adopt incremental hypothesis-driven testing strategies to deal with the challenges of the inductive setting. We suggested this might involve testing nearby confirmatory generalizations of a focal hypothesis (Klayman & Ha, 1989), or contrasting nearby variants to this hypothesis (Oaksford & Chater, 1994). In either case, we argued this would result in patterns of similarity (retention of rule-critical elements and creation of minimal contrast pairs) and simplification (removal of non-rule critical elements) quite distinct from the predictions of information-driven or uncertainty-driven testing. We indeed observed anchoring within learning problems. In particular, participants scenes appeared to be anchored both persistently to the initial positive example and sequentially (Figure 6c). We here operationalize this by creating a family of scene adaptation models that assume learners create new scenes by mutating either the initial positive example, or their own previous scene. We compare these against baselines that rather assume learners generate each new scene from scratch. Concretely, the models we fit were:

1. **Generate {Uniform}:** Adds a random number of objects to each scene. Uniform assumes each object uniformly selected features (color, size, orientation and groundedness)[12]. This model has zero fitted parameters so acts as an overall baseline.

[12] We do not attempt to predict the relational features or absolute positions in this analysis.

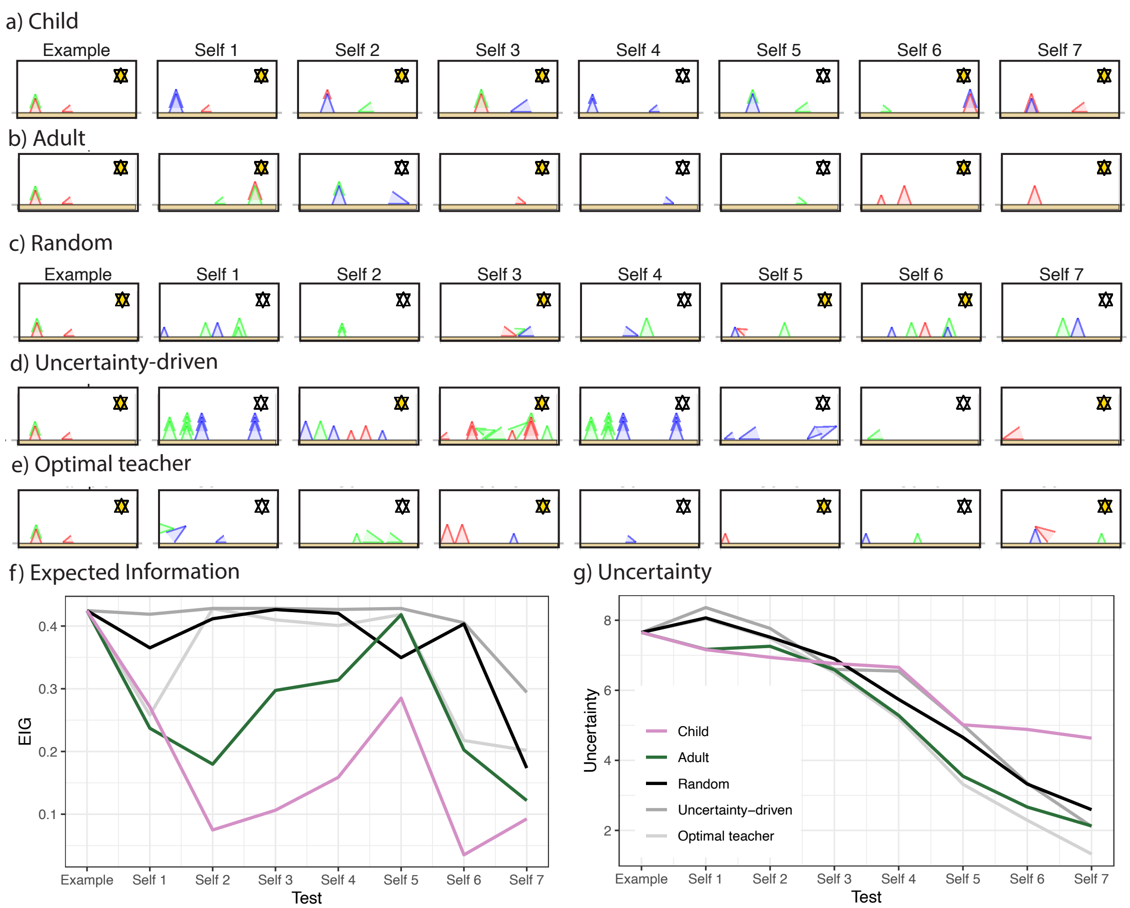


Figure 10

Example sequences for the "There is a red" problem. a) A child's scenes b) An adult's scenes c) Random selection from all participant generated scenes d) Uncertainty driven selection from all participant scenes e) Optimal scene selection for communicating the concept. f) Expected Information Gain and g) achieved uncertainty reduction for sequences in a–e.

Otherwise with this and all subsequent models we assumed each feature was sampled from its mean prevalence to act as a stronger baseline.

2. **Generate Simple:** Adds a number objects to each scene drawn from an exponential distribution (truncated to the maximum allowable number of objects) with fitted rate parameter λ, selecting the features of these objects at random. This models a tendency to create simple scenes containing fewer objects, with the mean number of objects per generated scene given by $\frac{1}{\lambda}$.

3. **Adapt Initial {Simple}:** Assumes the learner creates each new scene by adapting the initial scene. Concretely, we assume the learner samples either the same number of objects as in the initial scene with probability η, or a random number with

probability $1 - \eta$. The objects in new scene are assumed to be a mixture of the features of the matching object in the initial scene (replicating the original feature with probability η) or selected randomly from their support (with probability $1 - \eta$). We marginalize over all possible object mappings between scene i and $j.\eta = 1$ corresponds to perfectly reliable copying of the number and nature while $\eta = 0$ denotes always resampling the feature. The simple variant assumes the number of objects in the scene, if not drawn from the inspiration scene, is drawn from an exponential distribution with parameter λ as above.

4. **Adapt Previous {Simple}:** This model works as above but uses the preceding scene rather than the initial scene as its starting point.

5. **Adapt Mixed {Simple}:** This model simply mixes the predictions of Adapt Initial and Adapt Previous to capture the behavior of a learner who sometimes adapts the initial scene (with probability θ) or by their own preceding scene with probability $(1 - \theta)$.

We fit the models to each agegroup, and separately every individual participant (see Appendix B for details). Table 5 shows the resulting agregroup-level BICs the number of individuals best fit by each model and the spread of parameter values for each. Adapt Mixed Simple was the best model for both agegroups overall and the best model for 48% of children and 38% of adults. No participant was better fit by Generate or Generate Simple, capturing that every single participant exhibited some degree of positive anchoring on the number or nature of the earlier scenes. 80% of children and 96% of adults additionally showed an additional preference for simple scenes. Almost half of adults (48%) were best characterized as adapting the previous scene than repeatedly adapting the initial scene or a mixture of both while this was only true for 19% of children. Fitted simplicity rate λ was larger for adults (≈ 0.5) than children (≈ 0.3) capturing their stronger tendency to create scenes with fewer objects. Fidelity of copying features of inspiration scenes η was similar for children and adults ($\approx .3$). Note that this is an underestimate due to the need to marginalize over many possible object-object mappings and two potential inspiration scenes. Mixture parameter θ was below .5 on average for both children and adults suggesting dominance of the initial scene over the previous scene.

In sum, this model comparison supports the idea that learners adapted their earlier tests often retaining the same number of objects and tending to keep many of the same features. Adults were more likely than children to reduce the number of objects and had more tendency to adapt sequentially, gradually traveling further away from the initial example.

Table 5
Models of Scene Generation

Model	BIC/scene	N Best	λ	η	θ
Children					
Generate Uniform	40.2	0			
Generate	34.9	0			
Generate Simple	30.7	0	0.34 ± 0.1		
Adapt Initial	30.4	2		$.29 \pm .19$	
Adapt Previous	30.1	8		$.25 \pm .18$	
Adapt Mixed	30.0	1		$.27 \pm .19$	$.40 \pm .29$
Adapt Initial Simple	29.3	7	0.33 ± 0.11	$.34 \pm .16$	
Adapt Previous Simple	29.0	10	0.34 ± 0.13	$.31 \pm .17$	
Adapt Mixed Simple	28.7	26	0.34 ± 0.12	$.33 \pm .17$	$.40 \pm .24$
Adults					
Generate Uniform	32.8	0			
Generate	27.8	0			
Generate Simple	23.1	0	0.50 ± 0.18		
Adapt Initial	23.6	0		$.23 \pm .14$	
Adapt Previous	23.4	1		$.21 \pm .13$	
Adapt Mixed	23.3	1		$.21 \pm .13$	$.35 \pm .26$
Adapt Initial Simple	22.4	5	0.50 ± 0.20	$.29 \pm .12$	
Adapt Previous Simple	21.9	24	0.54 ± 0.30	$.23 \pm .13$	
Adapt Mixed Simple	21.8	19	0.54 ± 0.27	$.24 \pm .13$	$.32 \pm .25$

Note: BIC/scene shows the fit of the model at the agegroup level divided by the number of scenes for easier comparison. λ (simplicity), η (fidelity) and θ (mixture) show $M \pm SD$ of best fitting model parameters variant across subjects. Boldface indicates the best fitting model.

General Discussion

In this paper, 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. Children's guesses and tests were more complex than those of adults.

2. We could synthesize the diversity and distribution of children and adults' guesses with a constructivist—symbolic, generative—inference framework, reproducing both their sporadic correct guesses but also capturing the spread of their incorrect ideas

and offering a framework for modeling differences between children's and adults' inductive inference.

3. Children's guesses reflected less fine-tuned construction mechanisms than adults', producing more diversity but were consequently less predictable.

4. Both children's and adults' hypothesis generation appeared data-inspired, shown by better fit throughout our model-based analyses by our Instance Driven Generation account—inspired by patterns in the learning scenes—over our approximately normative (PCFG) account—that generated hypotheses a priori and weighted them with the evidence.

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

6. Both children and adults scenes generation seemed to involve modifying previous scenes, with adults doing so more systematically and with more tendency to simplify them.

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 inductive inference, some of what is gained in open-endedness is lost in experimental control. There is considerable residual ambiguity about the extent that differences in active learning shaped differences in hypothesis generation and visa versa. One way to try and partial this out could be to run more 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). Sole focus on scenarios fix one or other aspect of the the inductive inference loop 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. In future work hope, we plan to elicit more fine-grained online measures of learners' thought process—e.g. asking them to list their hypotheses after each guess or describe how they construct test scenes. This would support comparison of process-level accounts of both hypothesis adaptation and active search and allow identification of individual differences.

Theoretical Expressivity

There are many ways we could have set up the primitives, parameters and productions of our PCFG and IDG models. 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.

Elicitation 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

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; 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).

While neither the PCFG or IDG are oven-ready process models of human concept formation, they provide a useful starting point for thinking about process accounts. The PCFG framework describes normative inference in the limit of infinite sampling, but also provides a mechanism for both generating and adapting samples. The IDG is a hybrid that seeds hypotheses by trying to describe patterns that are present in observations rather than merely possible, making it more sample-efficient as a brute force approach to inference in situations where a learner already has some positive or demonstrative evidence of a concept. However its success is dependent on the learner generating or encountering scenes that exemplify and isolate causally relevant features. With enough evidence both approaches should favor the ground truth but with little evidence the PCFG will tend to entertain many concepts that the IDG does not.

While the IDG captured the data better here, it is not a complete account because, even with instance-inspired stating point, we still need to explain how a learner adapts in light of new evidence. Following a number of recent research lines (Bramley, Mayrhofer, Gerstenberg, & Lagnado, 2017; Dasgupta, Schulz, & Gershman, 2017; Ullman, Goodman, & Tenenbaum, 2012), we see incremental mutation of one or a few focal hypotheses in the light of evidence as a promising approach. For instance, a learner might use an observation to generate an initial idea akin to our IDG, but then explore permutations to this to generate new scenes to test (Oaksford & Chater, 1994), and to account for these tests (Fränken et al., 2022). While older models like RULEX (Nosofsky & Palmeri, 1998; Nosofsky et al., 1994) provide candidate heuristics for achieving such a search over theories, their long run behavior lacks a clear relationship with computational-level rationality (Navarro, 2005). However, if a learners' adaptations approximate a valid approximation scheme, for instance accepting proposed permutations with the Metropolis-Hastings probability $\max(1, \frac{P(h')}{P(h^t)})$ (Bramley, Dayan, et al., 2017; Dasgupta et al., 2016; Hastings, 1970; Thaker et al., 2017), they can start to explain why more probable hypotheses are discovered more often as well as explaining probability matching and order effects are inevitable consequences of approximation (see Fränken et al., 2022). Since the endpoint of an MCMC search approaches an independent posterior sample, we would expect a population of such 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). Taking the idea that earlier hypotheses carry information about older evidence and inference, we might also think of a population of such hypotheses as a kind of particle filter (Bramley, Dayan, et al., 2017; Daw & Courville, 2008). While acting primarily as a computational level norm, the PCFG prior provides useful infrastracture for hypothesis search. For example, prior production weights can be used to adapt an existing hypothesis by partially "regrowing" it (Goodman et al., 2008). Furthermore, prior production weights implied by a generative prior mechanism combined data likelihoods allows for the principled acceptance or rejection of new proposals in an MCMC-like search scheme. This could result in much greater sample efficiency than either the PCFG or IDG presented here, and it would be interesting to consider combinations of prior- or instance-driven initializations with permutation-based search. For this to become a fully satisfying account of constructivist inference this would need to be paired with a mechanism for scene generation in line with those we sketch in Figure 3c&d, so explaining anchoring, order effects, probability matching and confirmation bias in a unified account (Klahr & Dunbar, 1988).

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 the Similarity model performing at chance and also being completely uncorrelated with participants while all our symbolic model variants received support. While this is far from an exhaustive comparison with sub-symbolic concept models, even a successful similarity-driven 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). Our generalization analysis also showcases the difficulty of predicting human behavior in a setting where there is such a large and long-tailed space of similarly plausible rules an individual might be using to drive their generalizations. Modeling symbolic inference directly from the learning input had some predictive power for adults' generalizations, but simply by asking participants for their best guess, we could immediately get a far better handle on how they would generalize.

While we did not provide a fully satisfying model of scene generation, we did show that participant-generated scenes were better understood as adapting earlier scenes than as being created from scratch. We argued that this is consistent with testing driven by one or a couple of conceptually neighboring hypotheses, either generalizing their predictions or contrasting them. This is in some ways a return to pre-Bayesian ideas in philosophy of science in testing permits falsification but not confirmation. Even when a hypothesis h survives repeated confirmatory tests, or repeated head-to-head challenges from local alternatives, we might think of it as gaining a degree of confirmation, but there always remains the specter of potential future falsification (cf. Popper, 1959). We think this better reflects the state of a constructivist learner who cannot know, until discovering it, whether some better hypothesis is waiting in the wings.

For a learner limited to a few hypotheses at a time, the approach has clear virtues: It links the process of adapting a hypotheses with that of coming up with new scenes to test and links the outcome of tests to the subsequent inferential step of supplanting or reinforcing the current favored hypothesis. Since learners are always reusing at least some feature or other, it allows the learner's two tasks to support the other, with reuse of modified previous tests and minimal positive examples minimizing the cognitive and physical costs of generating both new tests and new hypotheses (Gershman & Niv, 2010).

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 potentially normative: Having been exposed to less evidence, with less idea what conceptual compositions and fragments will be useful in understanding their environment, we should 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 there are several 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 theory space 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, Schulz, Speekenbrink, Nelson, & Meder, 2018) 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 measure children and adults' prior distributions directly. If children's guesses shift within a problem in a way that is less sensitive to their own 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. Importantly, while the endpoints of children's theorizing were more diverse than adults', the cognition required to produce their hypotheses still highly systematic. Children were able to implement a stable-enough symbolic generation or adaptation mechanism to produce meaningful symbolic hypotheses on the large majority of trials, referring to the features and relations they encountered. Even when their hypotheses did a poor job of explaining all the learning data, the hypothesis construction process did not break down entirely as it would if childlike brain activity were simply random and disorganized. However, the issue remains whether there is just more noise in children's behavior—e.g., they are just a bit more easily distracted compared to adults—as opposed something like a greater inclination to explore.

Another aspect of constructivism that we did not focus on here but that is critical to understanding development, is the idea that over time, learners can chunk, cache and recursively reuse concepts to build ever richer ones (cf. Zhao, Bramley, & Lucas, 2022). As such the conceptual library of an adult ought to be more advanced, containing more powerful and complex concepts that can be readily reused to build new concepts. This might lead to a prediction of a different pattern of guesses than we found here. That is, we might have expected adults' concepts to look more complex than children's, not because they are built from more parts, but because the parts they are built from are, themselves, more complex. We suspect that the reason we did not find this sort of pattern here is that our task used very basic abstract features. Presumably our shape and geometric relation concepts are fairly established by around the age of 10. We predict that this would not hold in more applied domains where adults are able to draw on advanced concepts. For instance, when theorizing about economic conditions an adult might refer advanced primitives like “power laws”, “compound growth” or “arbitrage” that we would not expect to exist yet in the conceptual repertoire of many 9-11 year olds.

As well as producing more complex guesses, children also produced more elaborate scenes during learning. One possible characterization is that children's active scene construction was more exploration-driven and less hypothesis-driven than adults' (Wu et al., 2018), perhaps mixing more hypotheses-free exploration-driven actions in with hypothesis-driven systematic ones (Meder, Wu, Schulz, & Ruggeri, 2021). 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). However 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. 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, 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 uncertainty and potentially reflecting a more metacognitive control over learning (Kuhn & Brannock, 1977; Oaksford & Chater, 1994).

Curiously, children were more likely to refer to relational and positional properties in their guesses, while adults were 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.

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 all eight 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 of ever 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. In sum, it seems children were less able to come up with a concise description of all the evidence generated, reflecting both a less developed metacognitive awareness and the skills needed (both verbal and conceptual) to extract patterns.

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 a constructivist 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., Jones, A., Gureckis, T. M., & Ruggeri, A. (2022). Changing many things at once sometimes makes for a good experiment, and children know that.

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.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review*, *67*, 380–400.

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.

Church, A. (1932). A set of postulates for the foundation of logic. *Annals of mathematics*, 346–366.

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

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

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.

Daw, N., & Courville, A. (2008). The pigeon as particle filter. *Advances in neural information processing systems*, *20*, 369–376.

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 Ppress.

Fränken, J.-P., Theodoropoulos, N. C., & Bramley, N. R. (2022). Algorithms for adaptation in inductive inference. *Cognitive Psychology*.

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.
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.
Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?* (pp. 205–259). Springer.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.
Lapidow, E., & Walker, C. M. (2020). The search for invariance: repeated positive testing serves the goals of causal learning. *Language and concept acquisition from infancy through childhood*, 197–219.
Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).
Lewis, O., Perez, S., & Tenenbaum, J. (2014). Error-driven stochastic search for theories

and concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36).

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, *25*(1), 322–349.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological Review*, *111*(2), 309.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147.

Luce, D. R. (1959). *Individual choice behavior*. New York: Wiley.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Marr, D. (1982). *Vision*. New York: Freeman & Co.

McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. A. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22.

Meder, B., Wu, C. M., Schulz, E., & Ruggeri, A. (2021). Development of directed and random exploration in children. *Developmental Science*, *24*(4), e13095.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207.

Meng, Y., Bramley, N., & Xu, F. (2018). Children's causal interventions combine discrimination and confirmation. In *Proceedings of the* 40^{th} *Annual Meeting of the Cognitive Science Society.*

Michalski, R. S. (1969). On the quasi-minimal solution of the general covering problem. In *Proceedings of the* 5^{th} *Annual Symposium on Information Processing* (Vol. A3, pp. 125–128).

Morris, A. (2021). Invisible gorillas in the mind: Internal inattentional blindness and the prospect of introspection training.

Navarro, D. J. (2005). Analyzing the rulex model of category learning. *Journal of*

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

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175.

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. (1994). Another look at eliminative and enumerative behaviour in a conceptual task. *European Journal of Cognitive Psychology*, *6*(2), 149–169.

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

Osborne, M., Garnett, R., Ghahramani, Z., Duvenaud, D. K., Roberts, S. J., & Rasmussen, C. (2012). Active learning of model evidence using bayesian quadrature. *Advances in neural information processing systems*, *25*.

Phillips, D. C. (1995). The good, the bad, and the ugly: The many faces of constructivism. *Educational researcher*, *24*(7), 5–12.

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., & Jacobs, R. A. (2016). Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science*, *25*(1), 54–59.

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.

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

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. In *Neural Information Processing Systems.*

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 *Proceedings of the 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. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*.

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

Schulz, E., Klenske, E. D., Bramley, N. R., & Speekenbrink, M. (2017). Strategic exploration in human adaptive control. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.* The Cognitive Sicence Society.

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.

Shackle, S. (2015). Science and serendipity: famous accidental discoveries: Most scientific breakthroughs take years of research–but often, serendipity provides the final push, as these historic discoveries show. *New Humanist*, *2*.

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.

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.

Simon, H. A. (2013). *Administrative behavior*. Simon and Schuster.

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.

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the turing test* (pp. 23–65). Springer.

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. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924.

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

Zhao, B., Bramley, N. R., & Lucas, C. (2022). Powering up causal generalization: A model of human conceptual bootstrapping with adaptor grammars. In *Proceedings of the 44th annual meeting of the cognitive science society* (Vol. 44).

Zhao, B., Lucas, C. G., & Bramley, N. R. (2022). How do people generalize causal relations over objects? a non-parametric bayesian account. *Computational Brain & Behavior*, *5*(1), 22–44.

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 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.[13] 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. We chose 10,000 simply because this provided reasonable coverage of the task without exhausting our storage and computational capacity.

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.

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

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.

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 productions for adults or for children (Figure 7) and denoted with square brackets:

1. **Observe.** either:

   (a) With probability $[A \to B]$: Sample a cone from the observation, then sample one of its features f with probability $[G \to f]$—e.g., {#1}:[14] "medium, size" or {#3}: "red, color".

   (b) With probability $[A \to \text{Start}]$: 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

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

expands the number of cones picked out from the scene relative to an equality—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 7 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.[15]

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

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

with $[L \rightarrow]$ where appropriate. If it is true of only a subset of the cones then $\forall(\lambda x_i : A, \mathcal{X})$ 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 "$\forall(\lambda x_1 :\ \exists(\lambda x_2 :\ \wedge(=(x_2, \text{green}, \text{color}), \leq(x_1, x_2, \text{size})), \mathcal{X}), \mathcal{X})$". 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 7a&7b, but will only do so when exposed to an observation that exemplifies 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.

Reverse engineering production child-like and adult-like production weights

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 $\{263+5,\ 108+5,\ 297+5\}/(263+108+297+15) = \{.39, .17, .44\}$. To avoid double counting the data in modeling subjects' specific guesses, we created a separate agegroup-appropriate prior production weighting for each participant based on the guesses of the other participants' from the same agegroup, but omitting their own guesses.

Appendix B - Model fitting details

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.[16] Full results are in Table A-3.

Scene generation model fits

We used a grid search in increments of 0.05 to optimize η and θ and directly optimized λ for each setting of η and θ.

Appendix B: Free response coding

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

[16] 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 Agegroup 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	N blind	Accuracy
1.	Baseline	children	-1319.75	2639.50			7	13	50%
2.	Bias	children	-1218.96	2445.47	0.32		**16**	**25**	50%
3.	PCFG Uniform	children	-1319.72	2647.00		58.17	0	1	61%
4.	PCFG Uniform + Bias	children	-1208.93	2432.97	0.35	2.18	0	0	
5.	PCFG Flipped	children	-1318.46	2644.47		8.97	1	1	66%
6.	PCFG Flipped + Bias	children	-1207.28	2429.67	0.34	2.07	0	0	
7.	PCFG Agegroup	children	-1319.58	2646.71		24.17	1	1	63%
8.	PCFG Agegroup + Bias	children	-1208.63	2432.36	0.35	2.15	0	0	
9.	IDG Uniform	children	-1298.73	2605.02		1.78	1	2	65%
10.	IDG Uniform + Bias	children	-1193.90	2402.90	0.32	1.19	0	0	
11.	IDG Flipped	children	-1315.49	2638.54		4.35	1	4	66%
12.	IDG Flipped + Bias	children	-1199.22	2413.54	0.35	1.38	0	0	
13.	IDG Agegroup	children	-1308.05	2623.65		2.51	2	5	69%
14.	IDG Agegroup + Bias	children	-1193.41	2401.93	0.34	1.19	0	0	
15.	Similarity	children	-1316.44	2640.42		-1.99	0	1	41%
16.	Similarity + Bias	children	-1214.71	2444.52	0.32	-1.30	1	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	5	50%
2.	Bias	adults	-1364.90	2737.40	0.15		6	6	50%
3.	PCFG Uniform	adults	-1320.64	2648.89		1.27	0	0	63%
4.	PCFG Uniform + Bias	adults	-1253.52	2522.25	0.26	0.68	0	0	
5.	PCFG Flipped	adults	-1294.91	2597.42		1.06	1	1	66%
6.	PCFG Flipped + Bias	adults	-1229.18	2473.55	0.24	0.63	0	0	
7.	PCFG Agegroup	adults	-1266.96	2541.51		0.94	1	5	69%
8.	PCFG Agegroup + Bias	adults	-1203.64	2422.47	0.23	0.59	0	0	
9.	IDG Uniform	adults	-1228.21	2464.02		0.67	2	8	69%
10.	IDG Uniform + Bias	adults	-1179.12	2373.44	0.20	0.48	0	0	
11.	IDG Flipped	adults	-1245.56	2498.72		0.76	0	5	73%
12.	IDG Flipped + Bias	adults	-1179.23	2373.65	0.24	0.48	0	0	
13.	IDG Agegroup	adults	-1188.28	2384.17		0.62	2	**15**	74%
14.	IDG Agegroup + Bias	adults	-1134.58	2284.37	0.20	0.44	0	0	
15.	Similarity	adults	-1359.05	2725.70		-0.73	0	1	37%
16.	Similarity + Bias	adults	-1337.55	2690.30	0.14	-0.61	0	4	
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		

Note: Boldface indicates best fitting model overall. N blind restricts comparisons to models blind to the symbolic guess. Underlines indicate best fitting blind model. Accuracy column shows performance of the requisite model on 100 simulated runs through the task using participants' active learning data with τ set to 1/100 (i.e. hard maximizing over the model predictions). Biased models perform strictly worse so are not included in this column.

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.

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).

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**

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 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).

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).

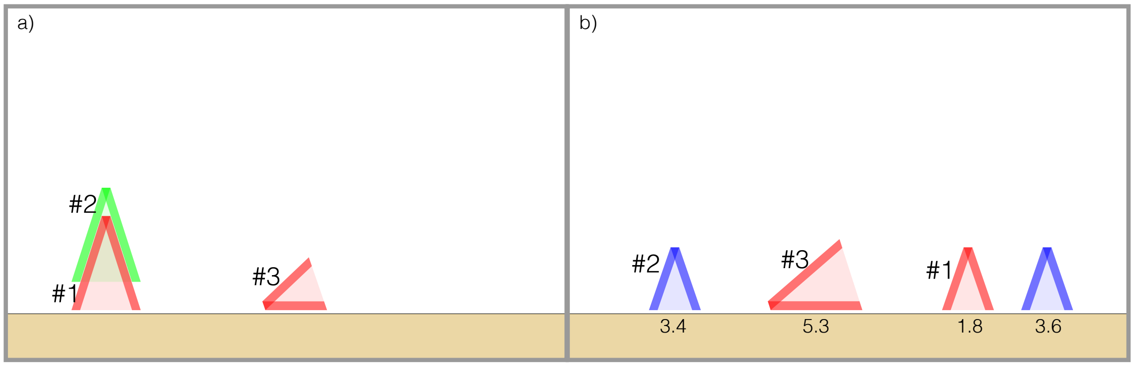


Figure A-1
Three example scenes. Objects indices link the most similar set of objects in b to those in a. Numbers below indicate the edit distance for each object (i.e. the sum of scaled dimension adjustments).

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).

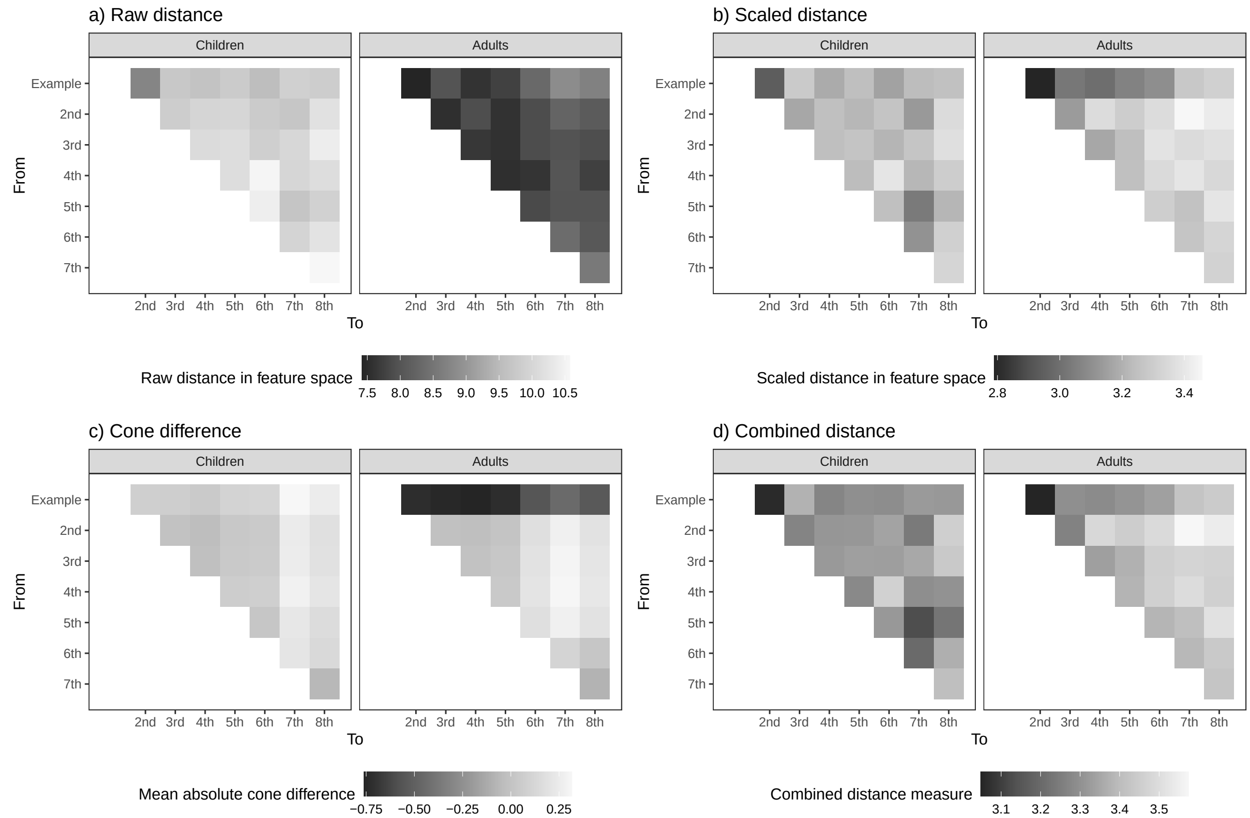


Figure A-2

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 modeled 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.

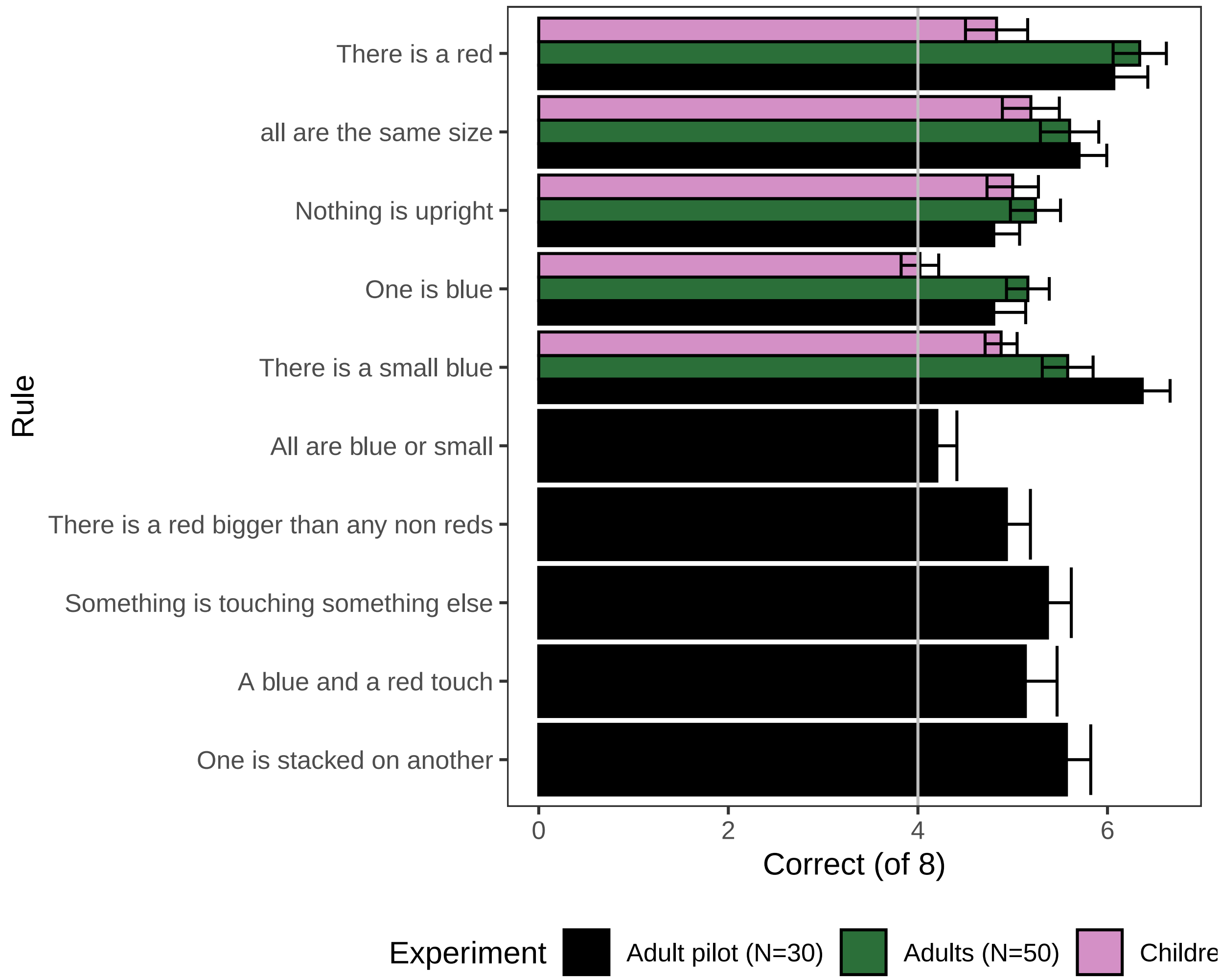


Figure A-3

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