Hypothesis generation through active inductive inference in children and adults

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

Gwyneth Heuser
Psychology Department, University of California, Berkeley, USA

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 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 symbolic concepts through active testing. Children generated substantially more complex guesses about the hidden rule and were more elaborate in their testing behavior. We take a 'computational constructivism' perspective to explaining these patterns, positing that these inferences are driven by a combination of thinking (recombining and modifying existing concepts) and exploring (actively investigating and discovering patterns in the physical world). We show how our approach and rich new dataset help explain developmental differences in their hypothesis generation, active learning and inductive generalization.

Hypothesis generation through active inductive inference in children and adults

A central question in the study of both human development and reasoning is how learners come up with new ideas and hypotheses to explain the world around them. Children excel at forming new categories, concepts, and causal theories (Carey, 2009) and by maturity, this coalesces into a unrivalled capacity for intelligent thought. Recent work in machine learning has begun to characterize computational principles driving both human development and mature human intelligence, pointing at a capacity to flexibly adapt, combine and re-purpose representations in order to generate new theories (Bramley, Dayan, Griffiths, & Lagnado, 2017; Bramley, Rothe, Tenenbaum, Xu, & Gureckis, 2018; Ellis et al., 2020); plans (Lai & Gershman, 2021; Lake, Ullman, Tenenbaum, & Gershman, 2017; Ruis, Andreas, Baroni, Bouchacourt, & Lake, 2020); solve problems (Rule, Schulz, Piantadosi, & Tenenbaum, 2018) and create tools and technologies (Allen, Smith, & Tenenbaum, 2020). A related line of work has focused on understanding developmental change in the cognitive processes and behaviors that drive learning and underpin intelligence. The "child as scientist" (Gopnik, 1996) — or more recently, "child as hacker" (Rule, Tenenbaum, & Piantadosi, 2020) — perspective casts children's cognition as driven by broadly the same processes as adults' but at an earlier stage in a journey of construction and discovery. The idea is that childlike and adultlike behavioral differences should be reflected by parametrizable differences in search and thinking, reflecting rational principles of construction. Children's hypothesis generation and search has also been framed as "higher temperature" than adults' — producing more diversity of ideas at the cost of being noisier (Lucas, Bridgers, Griffiths, & Gopnik, 2014) — more narrowly focused on a few hypotheses at a time (Ruggeri & Lombrozo, 2014), and driven more by directed exploration and less by generalization (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2017).

Constructivism is an influential perspective in developmental psychology (Carey, 2009; Xu, 2019) and philosophy of science (Fedyk & Xu, 2018; Quine, 1969) that posits learners actively construct new ideas through a mixture of thinking—recombining and modifying ideas—and play—exploring and discovering patterns in the world (Bruner, Jolly, & Sylva, 1976; Piaget & Valsiner, 1930; Xu, 2019). While influential, constructivism has so-far lacked a formal model, limiting its testable predictions and ability to contribute to AI development. Indeed, constructivist ideas are almost absent from major branches of current computational cognitive science. For instance, Bayesian models have played a dominant role in recent study of cognition, providing a principled way of modeling probabilistic inference (Howson & Urbach, 2006). However, since Bayesian accounts describe learning within a predefined hypothesis space, they are silent about how a learner might explore or extend that space. Neural networks have also received much recent

attention and achieve human-level performance in complex pattern recognition tasks (LeCun, Bengio, & Hinton, 2015), yet they require large amounts of training data compared to people and fail with unfamiliar inputs (Lake et al., 2017). From the constructivist view, this is to be expected. Neural network architectures are normally fixed ahead of training, and the weight-based representations they form are both opaque and distributed, limiting their suitability for later recombination. Information theory has also featured frequently in cognitive science as a metric of idealized information acquisition (Bramley, Dayan, et al., 2017; Gureckis & Markant, 2012). However, information-theoretic analyses also presuppose the Bayesian notion that learners have the relevant possibilities in mind and act to discriminate between them, rather than to support the task of constructing or discovering better ones. The central goal of this paper is develop a computational model of constructivist inference and use it as a lens to examine children's and adults' learning.

Traditionally, psychology research takes either a qualitative approach — studying cognition in naturalistic settings, eschewing formal models and statistics (Clarke & Braun, 2014; Piaget & Valsiner, 1930) — or a quantitative approach — distilling aspects of learning into constrained tasks to allow statistical tests of theories (Pearson, 1930). However, the lure of quantitative measurement has led to narrow focus on tasks that miss essential aspects of real-world learning. For example, learning studies are usually explicit about the possible hypotheses, and limit actions and responses, making the task one of discrimination or choice rather than generation. Bayesian models provide sensible benchmarks for these settings (Anderson, 1990; Marr, 1982), but it is unclear what behavioral alignment with them reveals (M. Jones & Love, 2011), since in real environments, good hypotheses are hard won and behavior is limited only by imagination. While constrained tasks bring advantages of convenience, the worry is they "short circuit" cognition, blocking the active, deeply generative aspects of naturalistic learning (Gureckis & Markant, 2012).

Quantitative psychology traditionally minimizes interpretation of introspective self-report data (e.g., Dennett, 1991; Johansson, Hall, & Sikström, 2008), but see also (Newell & Shanks, 2014; Szollosi, Liang, Konstantinidis, Donkin, & Newell, 2019). Indeed, a dominant perspective views concepts as similarity-based clusters of features (Medin & Schaffer, 1978; Posner & Keele, 1968; Shepard & Chang, 1963) that drive categorization of new examples (Kruschke, 1992; Love, Medin, & Gureckis, 2004) but lack symbolic structure. However, people often not only generalize successfully but can also describe their concepts to others, combine them in imagination and draw analogies (Holyoak & Thagard, 1989). It is not clear how subsymbolic approaches can account for these capabilities. We thus build on another line of work that considers concepts as symbolic and compositional

in character (Bramley et al., 2018; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, Tenenbaum, & Goodman, 2016). We stake a new methodological path, studying human hypothesis generation in an interactive context, where participants generate their own evidence as they learn and provide both forced-choice generalizations (quantitative data) and free guesses (qualitative data). We use a constructivist framework to bridge the divide between the two forms of data, by providing a formal model of hypothesis generation that can account for both quantitative and qualitative responses. Thus, our approach provides insight into the source of human flexibility in learning and thinking obscured by research focused on constrained tasks.

Our task is inspired by a tabletop game of scientific induction called "Zendo" (Heath, 2004). 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. The goal is to predict which a set of new scenes will produce the effect and identify the hidden rule that determines the general set of circumstances produce the effect (try it here). Scenes could contain a varied number of cones. Each has two immutable properties: size$\in$ {small, medium, large} and color$\in$ {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 a rich implicit space of potential concepts.

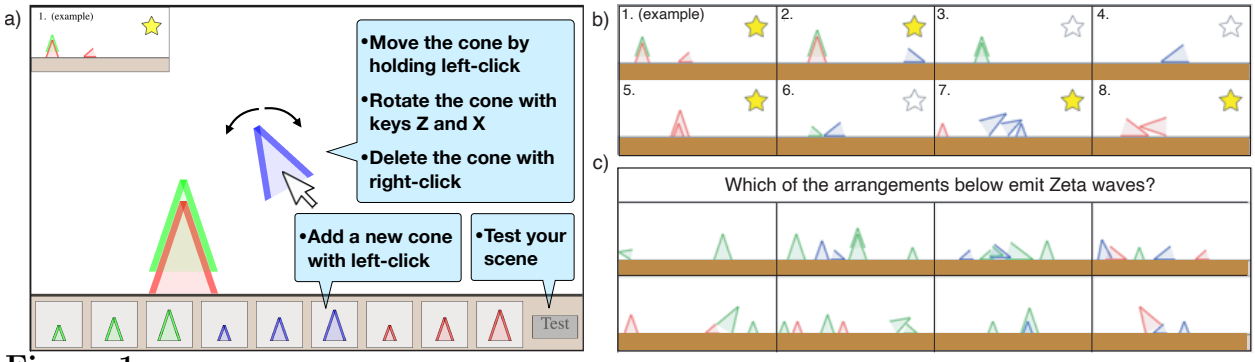


Figure 1

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.

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

Context-free hypothesis generation

In accounting for children's and adults' inferences, we entertain two related constructivist algorithms. The first takes a fully "top down" approach to inference, utilizing a probabilistic context-free grammar (PCFG) to define a latent prior over concepts expressible in first order logic. A PCFG is a collection of "productions" that stochastically build expressions in an underlying grammar (Ginsburg, 1966). A PCFG can be used to generate a prior sample of hypotheses that can then be weighted by their likelihoods of producing observations—here, their ability to reproduce the labels for the scenes that the participant has tested. The model's best guess about the hidden rule is then the *maximum a posteriori* hypothesis in the sample. The hypotheses make predictions about new scenes which can be weighted by their posterior probability and marginalized over to make generalizations. Because parts of the production process and underlying grammar involve branching—e.g., "and" and "or"—hypotheses can become arbitrarily long and complex, involving multiple Boolean functions and complex relationships between an unlimited number of bound variables. In this way, an infinite latent space is covered in the limit of infinite PCFG sampling (see Figure 2a).

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

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

Context-based hypothesis generation

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

Active learning

Children have long been seen as primarily active learners, using "play" to explore their environment and test their hypotheses (Bruner et al., 1976; Cook, Goodman, & Schulz, 2011; Piaget & Valsiner, 1930). Information theory is used to benchmark active learning (Nelson, 2005; Shannon, 1951) but assumes learners know the relevant possibilities and acts to discriminate rather than to constructing or discovering better ideas, potentially explaining why behavioral alignment with information maximization is mixed and task dependent (Coenen, Nelson, & Gureckis, 2018). The developmental literature has emphasized the utility of "control of variables" heuristic (Chen & Klahr, 1999; A. Jones, Bramley, Gureckis, & Ruggeri, in revision; Klahr, Fay, & Dunbar, 1993; Klahr, Zimmerman, & Jirout, 2011) — manipulating one design variable at a time, so that

changes in the outcome can be unambiguously attributed to the change in the input. Past research has only focused on restricted settings with a few simple variables and our task is much more complex. Thus, in exploring the active learning in our task, we will look for the empirical signature of control of variables style incremental and systematic testing.

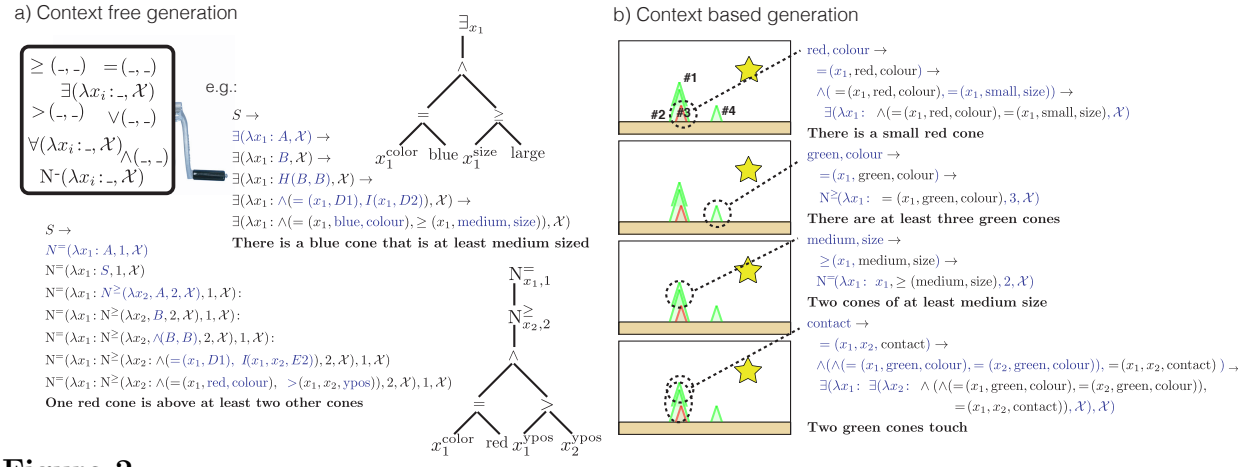


Figure 2

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.

In sum, the core contribution of this work is a close investigation of developmental differences in active open ended hypothesis generation and development of constructivist modeling approach that bridges qualitative–quantitative and symbolic–subsymbolic divides. To foreshadow, we find evidence of compositional concept formation in both adults and children; support for a bottom-up instance driven account. We find children create more complex learning data although do so less systematically than adults. They then go on to make more complex guesses while achieving a commensurate fit to evidence. Our constructivist framework suggests this behavior is a natural result of "flatter" idea and action generation mechanisms.

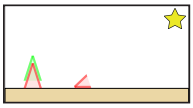
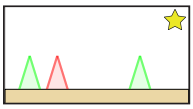
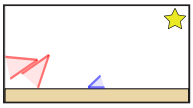
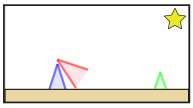
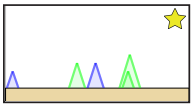
Experiment

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

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

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. 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 as well as a performance related bonus of up to $4 ($1.96±0.75). For children 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.

Materials and Procedure

Child sample.

Instructions. Participants sat in front of a laptop with a mouse attached, with the experimenter sitting next to them. The laptop displayed this webpage: `http://www.bramleylab.ppls.ed.ac.uk/experiments/zendo_kids/task.html`.

The experimenter read out the instructions displayed on the webpage 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. 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. There were five block types in total, one for each rule, and the order of these was randomized (see Table 1). Every initial example of a block type was a positive example, so 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.

Learning Phase. 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 were done moving the blocks around, they should press the "Test" button to see if stars came out of them. For positive tests, the experimenter would neutrally say: "Stars did come out of the [name]s that time" and for negative tests: "Stars did not come out of the [name]s that time."

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

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

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

Adult sample. We recruited our adult sample from Amazon Mechanical Turk and adults completed the task on their own computers. They completed the same instructions as the children with an additional section about bonuses and had to successfully answer comprehension questions, including an additional two about the bonuses, before starting the main task. Specifically, adults were bonused 5 cents for each correct generalization (up to a possible 40 cents for each of the five trials) and an additional 40 cents for an correct guess as to the hidden rule, again for each of the five trials. Aside from having no experimenter in the room, and filling out the text fields themselves, the procedure was identical to the children's task. Full materials including experiment demos, data and code are available at the Online Repository.

Results

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

Rule complexity and constituents

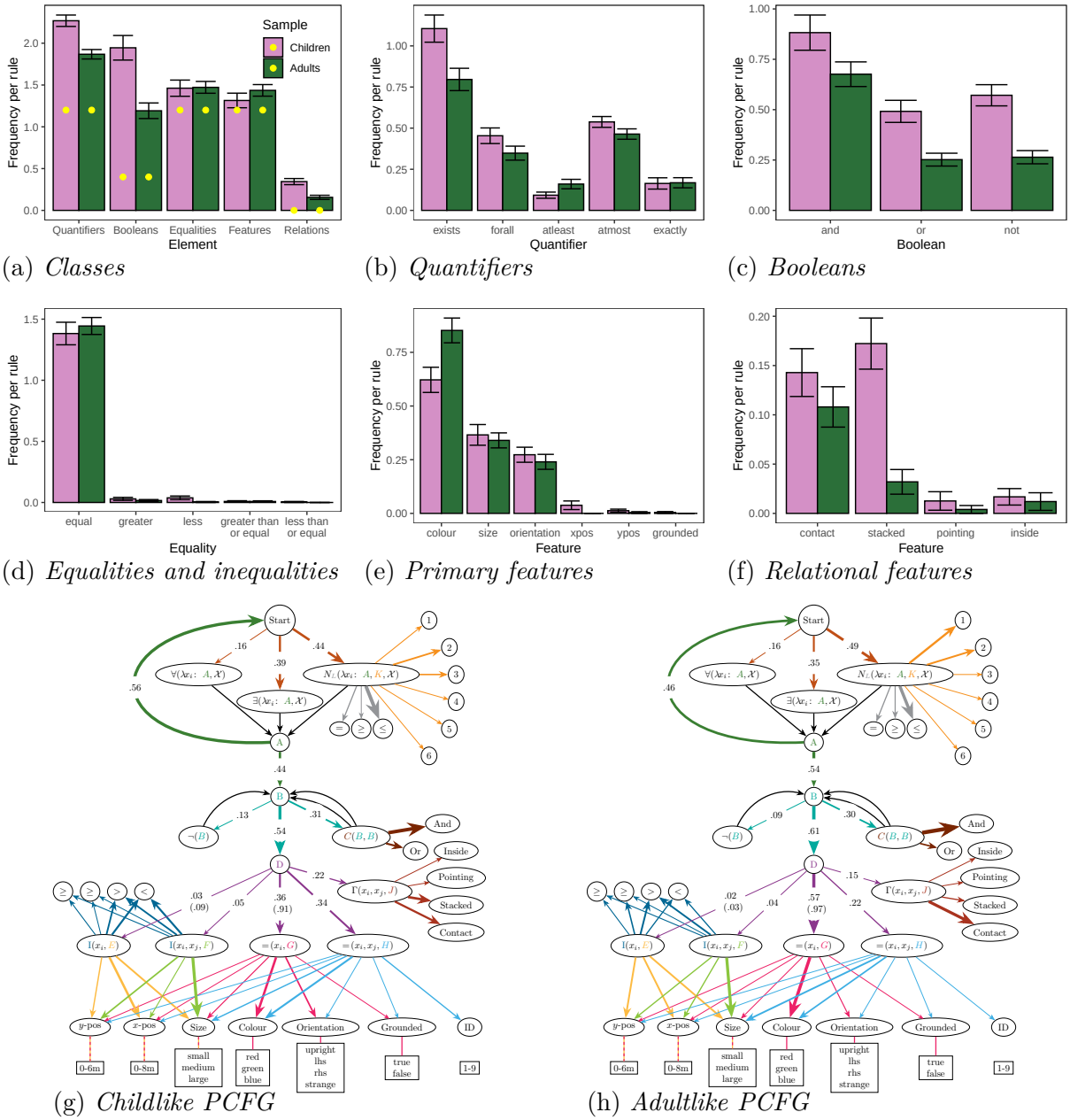
We had human coders translate participants' free text guesses about the hidden rule wherever possible into equivalent logical lambda expression using the grammatical elements available to our learning models. We were able to do this for 86% (n=205) of children's trials and 88% (n=219) of adults' trials. For example, if the participant wrote *"There must be one big red block"* this was converted into
$N^{=}(\lambda x_1 : \wedge(=(x_1, \text{large}, \text{size}), =(x_1, \text{red}, \text{color})), 1, \mathcal{X})$.[1]

To explore structural differences in children's versus adults' hypotheses, we first break down these encoded rule guesses into their grammatical parts. This reveals that children's encoded rules were substantially more complex than those generated by adults and that both were more complex than the ground truth rules. Children's and adults' rules

[1] 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 provide details about the coding in the Appendix and full coding resources and full coding data in the Online Repository.

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

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



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

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

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

Figure 3

(a–f) Relative frequency of rule elements in Children's and Adults' rule guesses. Yellow points in a show ground truth frequencies. (g&h) Visualization of the childlike and adultlike PCFGs, reverse engineered to produce rules with empirical frequencies matched to children's and adults'. 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.

Accuracy

Having observed systematic differences in the content of children's and adults' hypotheses, we now ask if these differences manifest in children's and adults' inferential

303 success; their ability to identify the ground truth and make accurate generalizations.

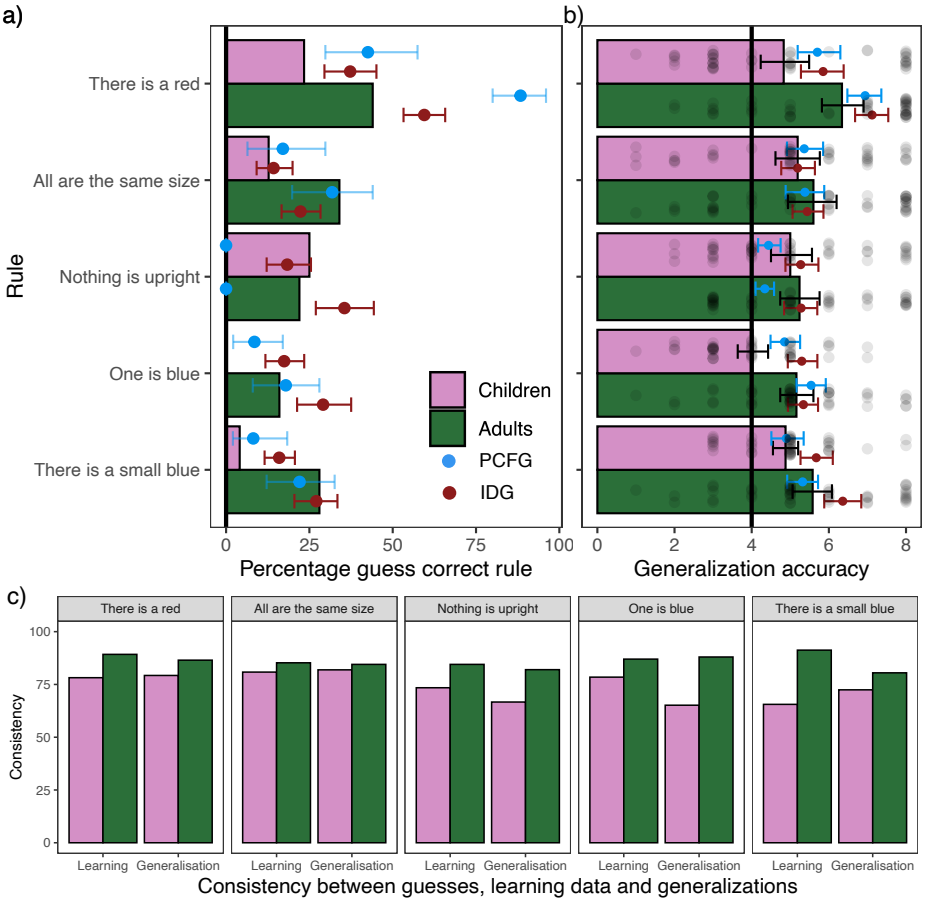


Figure 4

a) Percentage of participants guessing exactly the correct rule. Bars show mean ± bootstrapped 95% confidence intervals for children (pink) and adults (green). b) Generalization performance. Blue and red points show mean performance of PCFG and IDG simulations with fitted production weights ± bootstrapped 95% confidence intervals. Black vertical lines denote chance performance. c) Consistency between subjects' rule guess and their (self generated) learning data, and generalization judgments.

304 **Rule guesses.** Both children and adults were sometimes able to guess exactly the
305 correct rules, doing so a respective 11% and 28% of trials. Adults produced the correct rule
306 more frequently than children $t(102) = 4.0, p < .001$ and were more likely then children to
307 guess correctly (at a corrected significance level of 0.01) for the "All are the same size",
308 "One is blue" and "There is a small blue" rules (see Figure 4a). Note that chance level
309 baseline for this kind of guess is essentially 0%. There are an unlimited number of wrong
310 guesses and a small set of semantically correct guesses. It is also the nature of this
311 inductive problem that there is an infinite number of perfectly consistent rules for any
312 evidence, although as more evidence arrives the ground truth is increasingly likely to be

among "simplest" rules in this set. Thus, it is instructive to ask whether participants rules, where not exactly correct, are still consistent with the evidence they have seen.

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

We now compare this to simulated *context free* (PCFG) and *context based* (IDG) learning algorithms provided with the active learning data generated by the human participants. We produced 50,000 childlike rules $\hat{R}_{\mathrm{c}}$ and 50,000 adultlike rules $\hat{R}_{\mathrm{a}}$ that have properties matched to those in Figure 3a–f as well as an additional sample of rules based on uniform production weights $\hat{R}_{\mathrm{u}}$. These act as an approximation to the infinite latent prior over rules $P(r)$, before seeing any data. In order to approximate a posterior over rules given self-generated learning scenes $\mathbf{d}$, we then weight these samples by their likelihood of producing the scene labels observed during the learning phase

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

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

and count how often they appear in the prior sample, with indicator function $\mathbb{I}(.)$ denoting exact or semantic equivalence. To test for semantic equivalence, we computed predictions for the first 500 participant-generated scenes for each rule and clustered together those that made identical predictions. We round positional features to one decimal place in evaluating rules to allow for perceptual uncertainty.

We assume the following likelihood function

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

embodying the idea that: the more training points a rule cannot explain, the less likely it is to be true. For a large b the likelihood function approaches the true deterministic behavior of the rules. However, to allow for a some noise while maintaining computational tractability in our analyses we simply assume a $b = 2$, corresponding to a likelihood function that decays rapidly from 1 for rules that predict all 8 scenes' labels to

.13 for a single misprediction, and .02 for 2 mispredictions and so on.

To generate IDG predictions, as with the PCFG, we produced a childlike, an adultlike and a uniform sample of instance driven hypotheses. This involved merging the production probabilities from the PCFG into the Instance Driven Generation procedure detailed in the Appendix. However, since each generation depended on the particular scenes for inspiration and this differed for every participant and trial, we generated smaller samples of 10,000 childlike, adultlike, and uniform rules for each trial. We spread these evenly across the 8 learning scenes. 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.

Taking the maximum *a posteriori* estimate (guessing in the event of ties) under either model leads to guessing the correct hypothesis at similar levels to participants. For a uniform-weighted PCFG sample, the MAP is correct on 9%±28% of children's trials and 12%±33% of adults' trials. Note that since these simulations use the same prior sample, the small differences we see are due to the different learning data generated by children and adults. However, accuracy improves substantially and better reproduces the empirical child–adult accuracy difference if we use samples based on reverse engineered weights that reproduce the qualitative properties of children's and adults' rules (see Appendix and Figures 3g&h). For these samples, we get correct PCFG guesses for $15\% \pm 36\%$ of children's trials and $32\% \pm 46\%$ of adults' trials. Across rules, the PCFG does not match well with children's or adults' accuracy, overperforming participants on the syntactically simpler rule "there is a red" but failing to capture participants correct guesses of "Nothing is upright".

Uniform-weight IDG simulations guess correctly on 15%±18% of trials for children's learning data and 25%±23% for adults learning data. Using the reverse-engineered weights, this increases to $21\% \pm 23\%$ and $35\% \pm 29\%$ and again provides a visually closer fit to the by-rule guess rates (Figure 4a).

Generalizations. We now analyze the quantitative response data constituted by forced choice generalizations about which of 8 new scenes will produce stars (i.e. follow the hidden rule). Across the five tasks, both children and adults guessed more accurately than chance (50%): *children* mean±SD $59\% \pm 11\%, t(53) = 5.9, p < .001$; *adults* $70\% \pm 14\%, t(49) = 10.3, p < .001$. Adults' generalizations were significantly more accurate than children's $t(102) = 4.6, p < .001$ and children's accuracy improved significantly with age $F(1, 52) = 6.2, \eta^2 = .11, p = 0.015$. Indeed, adults' generalization accuracy was above a Bonferroni-corrected chance level of $p \leq 0.01$ for all five rules and children were similarly above chance except for rules 1. "There is a red" ($t(46) = 2.5, p = .015$) and 4. "One is

blue" ($t(46) = .1, p = .915$; see Figure 4b).

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

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

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

Provided with the active learning data generated by the human participants, both performed in the human range at generalization. Using uniform production weights and taking the marginally most likely generalization labels over a posterior weighted sample of PCFG-generated rules based on the participants active learning data yielded accuracies of 60.3%±18.8% for children's and 61.7%±19.5% for adults' data. The fitted-weight PCFG models perform a little better and reproduces the empirical difference between children's and adults' accuracy: 63±20% for children's and 69±21% for adults' PCFG weights. The unfitted IDG, again, performed slightly better than the PCFG, generalizing at 66.3%±20.1% from children's active learning data and 69.5%±20.7% from adults'. Again, the fitted-weight IDG models' performance increased slightly and better reproduced the difference between children's and adults' accuracy (68±20% and 74%±21%).

The better accuracy of the IDG compared to the PCFG replicates the findings of (Bramley et al., 2018) and extends it to children as well as adults. Intuitively, this is because the bottom-up mechanism ties the hypotheses generated to features of the learning cases, effectively narrowing in on plausible hypotheses more efficiently. More broadly, these simulation results underscore the inherent difficulty of inductive inference. Even in this "small world" with known and fully observed features and allowing cognitively implausibly large numbers of hypothesis samples, it is not possible to robustly outperform human adults in this task. The PCFG and IDG were not statistically better or worse than participants at any rule inference under after Bonferroni correction with the exception that the IDG outperformed children on rule 4 $t(96) = 4.5, p < .0001$.

Interim discussion

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

mentioned in children's rules than in adults'.

A complicating factor is that children generated different 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 "sharper" set fit to adults' rules, better reproduces the accuracy patterns. We take this to suggest hypothesis generation differences are driving a large portion of the differences in children's and adult's inductive inferences.

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

Search behavior

As well as generating more complex rules, children also tended to create substantially more complex scenes than adults during the learning phase. The average child-generated scene contained 3.7±0.88 objects compared to 2.84±0.57 objects for adults ($t(102) = 5.8, p < .001$). The complexity of test scenes was inversely related to performance overall ($F(1, 102) = 39.0, \beta = -0.08, \eta^2 = .28, p < .001$) and also within both the child sample ($F(1, 52) =, \beta = -0.056, \eta^2 = .20, p < .001$) and adult sample ($F(1, 49) = 9.1, \beta = -0.096, \eta^2 = .16, p < .001$) taken individually (see Figure 5a). Within the child sample, age was inversely associated with scene complexity with an average of 0.35 fewer objects per scene for each additional year of age $F(1, 52) = 12.6, \eta^2 = .19, p < .001$. Aside from this difference, we can also assess whether children's or adults' scenes bear the hallmarks of a local "search" across possible scene dimensions.

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

If this is the case, we should expect the scenes generated by participants to be more similar to the initial example than to a random scene or scene drawn from a different

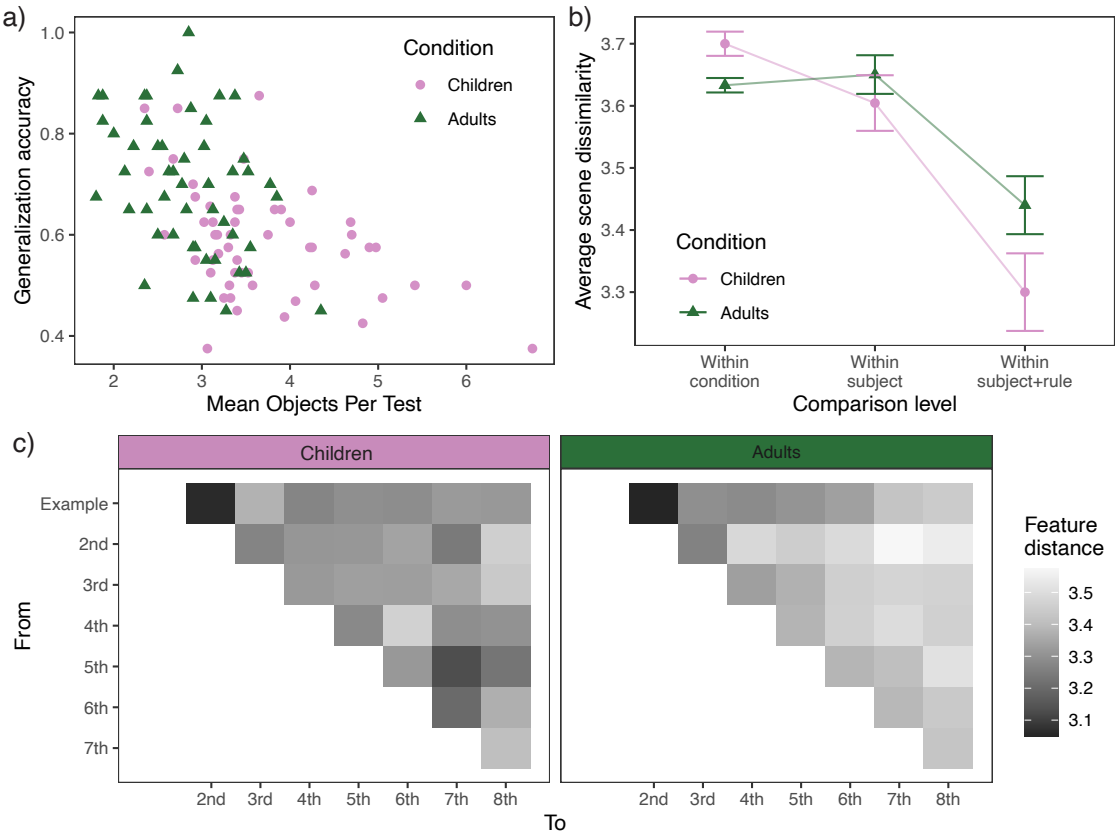


Figure 5

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

learning problem. To explore this, we constructed a distance metric that we used to measure the intuitive dissimilarity between any pair of scenes. The metric captures a form of 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. Essentially, this involved z-scoring and combining a "minimal-edit set" of feature differences and incorporating a proportional cost for additional or omitted objects. We provide a detailed procedure and example of how we computed these edit distances and break them down into their separate components in the Appendix. The mean distance between any randomly selected pair of participant-generated scenes was M±SD= 3.67 ± 0.94. Taken as a whole, the scenes generated by children were more diverse than adults' with average dissimilarity of 3.70 ± 0.14 compared to 3.63 ± 0.08, $t(102) = 2.9, p = 0.0048$.

However, this diversity seems to be *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 5bc).

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

These patterns manifest in small differences in the quality of the total evidence generated according to an information gain analysis. Adults' scenes are more informative under a uniform prior or either the childlike or adultlike prior (see Appendix).

Experiment Discussion

Our constructivist analysis suggests we cannot attribute the differences in hypothesis generation to the differences in active learning, nor can we attribute the differences in active learning directly to differences in hypothesis generation. That is, assuming the same generation process for the children's and then the adults' data does not reproduce the differences in rule guesses and accuracy and assuming scene creation is driven by distinguishing among the hypotheses manifesting the childlike or adultlike latent prior does not explain the developmental differences in the complexity and systematicity of the scenes creation. Rather, these data support the idea that developmental shift in hypothesis generation and active search are manifestations a gradual tuning of the constructivist generative mechanisms that produce novelty in cognition.

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

Model fitting

Guesses

To evaluate our constructivist PCFG and IDG models' ability to explain participants' free response guesses, we computed probability of each generating exactly the participant's encoded guess conditioned on their active learning data.

By definition, all 87% of participants' rules that we were able to encode in our concept grammar have nonzero support under a PCFG prior, and due to the stochasticity we assumed in our likelihood function, all should also nonzero have posterior probability. However, in practice it is impossible to cover an infinite space of discrete possibilities with a finite set of samples, meaning there are a substantial number of cases in which the complex rule produced by the participant did not appear in our PCFG or IDG samples at all. The proportion of Children's and Adults' rules that were generated at least once in 50,000 samples by the fitted PCFG was almost identical: 68% for Children's and 67% for Adults' guesses. The IDG fits were based on 10,000 samples generated separately for each trial and these samples included children's rules 50% of the time and adults' rules 60% of time using fitted weights and 49% and 62% of the time using uniform weights. Intuitively, the slightly lower coverage is due to the practical constraint that the IDG samples had to be computed separately for each trial limiting the number we were able to store and evaluate and due to the constraint that the current form of our IDG algorithm produces hypotheses with a maximum of two bound variables. The larger difference in coverage between children's and adults guesses for the IDG is consistent with the idea that the more complex learning scenes generated by the children resulted in a wider spread of scene-inspired hypotheses and a concomitantly lower chance of this including the children's explicit guess. For this reason, we visualize the posterior guess probabilities based on the samples at the trial level (Figure 6a).

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

Generalizations

A standard benchmark for models of concept learning is a fit with participants' generalizations to new exemplars. Thus, we complete our analyses by comparing 18 models ability to account for participant's generalizations. The set of models we consider allows us

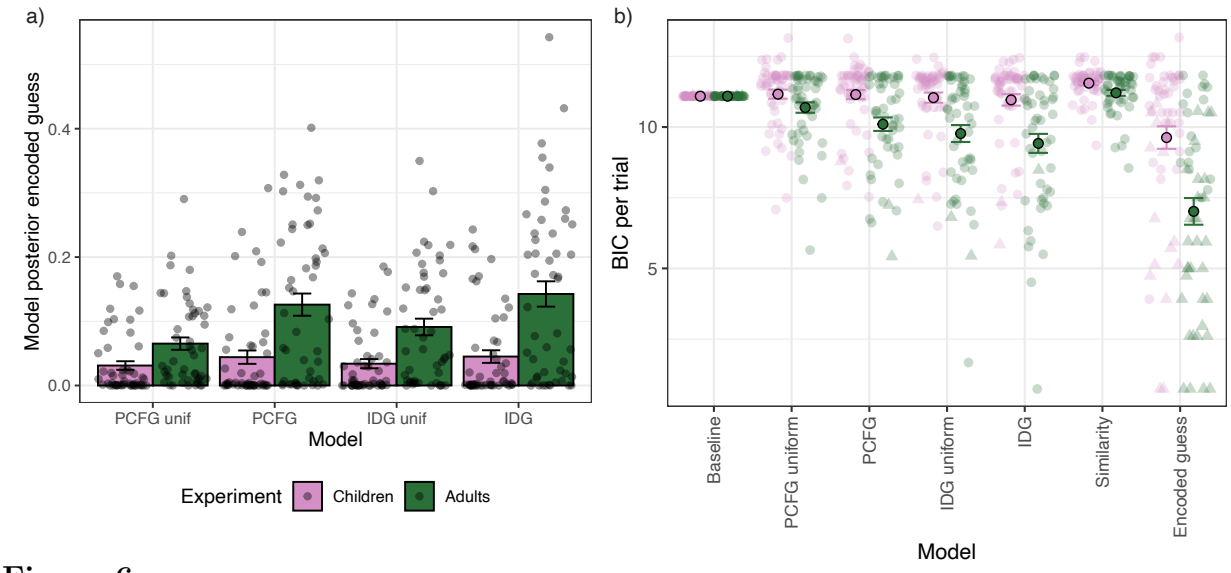


Figure 6

a) Encoded guess model fits: Bars and Points show each model's per participant average (±SE) posterior probability for all trials in which participants provided an encodeable guess (85%, N=424). b) Generalization model fits. BIC-per-trial completed for individual level model fits to participants generalization choices (lower is better). Opaque points show model mean±SE, faint points denote individual fits, with triangles used to mark where the model is the best fit (of all 18 tested) for that participant.

to test our core claims that children's and adults' induced representations are symbolic and compositional, as opposed to statistical and similarity-driven.

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

The models we fit were:

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

2. **Encoded Guess**. This model takes participants' free guess of the hidden rule, coded in lambda abstraction, and uses it to generate a prediction vector $r \in R$:{*rule-following*=1, *not rule-following*=0} for each scene. These predictions are then softmaxed using $P(\text{choice}) = \frac{e^{r\tau}}{\sum_{r \in R} e^{r\tau}}$,with inverse temperature parameter $\tau \in (-\infty, \infty)$ (Luce, 1959) optimised to maximize model likelihood. Large positive τ

indicates a hard maximization. $\tau \approx 0$ indicates random selection and negative $\tau < 0$ indicates anticorrelation between model predictions and choices. For trials in which the participant does not provide an unambiguous rule, the model assigns a .5 likelihood to each generalization choice.

3. Similarity. 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 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 the Appendix. We computed the mean difference between rule-following and not-rule following similarities as a $\Delta Similarity$ score for each participant×trial×item combination. Positive scores mean generalization item has a greater feature similarity to the rule following learning scenes than the not rule-following learning scenes. Negative scores mean the reverse. To convert these into choice probabilities, we take the inverse logit of these scores $r = \frac{\Delta \text{Similarity}}{\Delta \text{Similarity}+1}$ and again fit these r values to maximize the likelihood of participants' choices using a softmax function with inverse temperature parameter $\tau \in (-\infty, \infty)$. Intuitively, this model provides a non-symbolic alternative account of generalization behavior.

4-6. PCFG {uniform, off}. These models use the marginal likelihood of each generalization scene as rule following under the Probabilistic Context Free Generation (PCFG) posterior as the predicted r values. "Uniform" uses the prior with uniform production weights. "Off" uses the mismatched weights —- that is, adultlike weights for children's generalizations and childlike weights for adults' generalizations. As above, in each case, these prediction are fit to participants' choices using temperature parameter $\tau \in (-\infty, \infty)$.

7-9. IDG {uniform, off}. 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 $\tau \in (-\infty, \infty)$.

10. Intercept. This model 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 are 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 $P(\text{choice}) = \lambda b + (1 - \lambda).5$.

11-18. {Other model} + Intercept. These model variants combine the above model predictions (r = e.g. the Similarity Score or Encoded Guess prediction), with an overall generalization bias b as in 10. This involves jointly optimizing both the mixture λ and inverse temperature τ parameter such that:

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

We fit all models to to the children's and adults' data, and then separately to each individual participant. Individual level results are highlighted in Figure 6b. In all cases, the Intercept+PCFG and Intercept+IDG models performed no better than the pure PCFG or IDG models. Therefore we visualise individual participant fits for just models 1.–12 in Figure 6b. The full table of model fits is presented in the Appendix (Table A-3).

In line with our core hypotheses that participants inferences were symbolic, Encoded guess + Intercept is the best fitting model for both children's and adults' generalizations outperforming all the models we considered based just on only the learning data. For children, Encoded guess + Intercept has BIC 2149, improving 490 over Baseline with bias term mixture parameter $\lambda = .26$ and choice temperature parameter $\tau = 1.25$. For adults, this is BIC 1776 with a larger BIC improvement of 996 over Baseline, with a $\lambda = 0.08$ indicating lower bias and temperature $\tau = 2.0$ indicating higher fidelity alignment with the guessed-rule's predictions. Probing this bias, we see children undergeneralized 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, 16/50 children were best fit by the Baseline+Intercept model, followed by 15 by the Encoded Guess model, 9 by the Encoded Guess + Intercept model and a further 7 by Baseline. No other model best fit more than 2 children. For adults, 32/52 were best fit by Encoded guess, 6 by Baseline + Intercept, 4 by Encoded Guess + Intercept and no other model best fit more than 2 participants.

Thus, confirming our key constructivist hypothesis, there is a clear alignment between participant's symbolic rule guesses and their generalizations. Our end-to-end constructivist sampling models (blind to the participant's explicit guess) also received empirical support, predicting adults generalizations well above baselines and above the similarity-driven account which had no correlation with adults' or children's generalizations (see Appendix Table A-3). Thus, these results suggest participants' inductive inferences

can be explained by a computational constructivism framework combined with approximate Bayesian inference. As with the free responses, the IDG is better aligned with participants than the PCFG, particularly when using reverse engineered rather than uniform production weights.

Discussion

We explored children and adults' hypothesis generation and inductive inferences in an interactive task where the space of possibilities and actions is open and practically unbounded. We showed our computational constructivism framework can explain participants successes and helps us unpack the differences between children's and adults' behavior as consequences of differences in their generation and search mechanisms. In particular, we found support for a partially bottom-up Instance Driven Generation account over a fully top down Context Free (PCFG) approach, replicating (Bramley et al., 2018). We take this to support the idea that human inductive inferences spontaneously utilize compositional construction (Piantadosi, 2021) and further, that our introspective descriptions pick out genuine structural features of the consequent representations. Our formal model comparison supports these conclusions, with both end-to-end PCFG and IDG models predicting generalizations even while blind to the free responses while a feature-similarity model received no empirical support.

Developmental differences

Our analyses revealed a variety of developmental differences. Children's guesses were more complex than adults', and consequently we could capture them with a significantly "flatter" generation process that inherently produced a wider diversity of hypotheses. This is broadly normative: Having been exposed to less evidence, with less idea what conceptual compositions and fragments will be useful in understanding their environment, children are justified in entertaining a wider set of ideas. Children were more likely to refer to relational and positional properties in their guesses, while adults were by most likely to make guesses that pertained to the primary object features (color and size). This is an independently interesting finding, since relational features are structurally more complex than primitive features, we might have predicted they would be more readily evoked by adults. It could be that children bought in more to the scientific reasoning cover story, treating mechanistic explanations, such as that objects must touch or be positioned in particular ways to produce stars, as credible (Gelman, 2004). Conversely, adults may have been more likely to expect Gricean considerations to apply, e.g. that experimenters

would likely set simple rules using salient but abstract features like color over perceptually ambiguous properties like position (Szollosi & Newell, 2020).

Children also produced more elaborate scenes during active learning than adults. One possible characterization is that children's active scene construction also used a "flatter" generative prior, resulting in more diversity of exploration approaches. Indeed, differences in active exploration are the other side of the coin of the high temperature search idea (Friston et al., 2016; Gopnik, 2020; Klahr & Dunbar, 1988; Schulz, Klenske, Bramley, & Speekenbrink, 2017). On the other hand, adults' testing behavior was more systematic, potentially reflecting a more top-down, or strategic, *control of variables* approach to gathering evidence and updating beliefs (Kuhn & Brannock, 1977). Within each trial, children's testing was more repetitive, suggesting that they made slower progress in exploring the problem space.

Children's guesses were also moderately less consistent with their evidence than adults'. This might be because they were less able to extract common features across learning scenes (Ruggeri & Feufel, 2015; Ruggeri & Lombrozo, 2015). However, it could also be a consequence of limited hypothesis generation. With a flatter prior and limited sampling, one has a lower chance generating a hypothesis that can explain all the evidence. Children also under-generalized, selecting only 1 or 2 of the 8 test scenes (there was actually always 4) and even when their explicit guesses predicted more should be selected. On the face of it, this reflects Wu et al.'s (Wu et al., 2017) finding that children are weaker generalizers than adults.

Limitations and future directions

While this dataset provides an exceptionally rich window on developmental differences in inductive inference, some of what this task gains in open-endedness it loses in experimental control. There is residual ambiguity about the extent that differences in active learning cause differences in hypothesis generation and visa versa. Partialing this out would require controlled studies that fix the evidence and probe the hypotheses generated, or that fix the hypotheses and probe what evidence is sought. However, we have argued that constrained tasks run the risk of short-circuiting natural cognition. Learners may struggle to test hypotheses they did not concieve themselves, or to use data they have not generated to evaluate their hypotheses (Markant & Gureckis, 2014), meaning sole quantitative focus on studies that fix one or other aspect of the problem may provide a misleading perspective on active inference in the wild.

We also note that there are many ways we could have set up the primitives and productions of our PCFG and IDG models. Combined with non-uniform weights, this

makes for a dangerously expressive set of set of theories of cognition. We do not claim to have explored this space systematically here but that our modeling lends support to the idea that some symbolic and compositional process drives inductive inference. Identifying the computational primitives of thought may not be a realistic 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). Thus we expect different learners to take different paths in an inherently stochastic learning trajectory, limiting universal claims about representational content.

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.

Finally, as it stands, neither our PCFG or IDG are plausible process models. The PCFG is a framework for normative top-down inference in the limit of infinite sampling, and IDG is a hybrid that is be less sample inefficient as a brute force approach to inference in situations where a learner already has some evidence. A process account needs to explain how a learner searches the latent posterior in either framework with limited memory and computation. We see incremental adaptation of one or a few hypotheses in the light of evidence as a promising approach (Bramley, Mayrhofer, Gerstenberg, & Lagnado, 2017; Dasgupta, Schulz, & Gershman, 2017; Ullman, Goodman, & Tenenbaum, 2012). A learner might use an observation to generate an initial idea akin to our IDG, but then explore permutations to this to account for the rest of the evidence. Such a process account could also help differentiate recent perspectives on the source of developmental differences. For example, it would allow measurement of whether children's search patterns are more "high temperature" than adults' (Gopnik, 2020), over and above differences in the latent search space.

Conclusions

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

bring us closer to a computational understanding of human learning across the lifespan.

References

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

Anderson, J. (1990). *The adaptive character of thought*. Erlbaum.

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., 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.

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

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.

Clarke, V., & Braun, V. (2014). Thematic analysis. In *Encyclopedia of critical psychology* (pp. 1947–1952). Springer.

Coenen, A., Nelson, J., & Gureckis, M., Todd. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*, 1548–1587.

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

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

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.

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.

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

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.

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.

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

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

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive science*, *13*(3), 295–355.

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

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

Jones, A., Bramley, N. R., Gureckis, T. M., & Ruggeri, A. (in revision). Changing many things at once sometimes makes for a good experiment, and children know that. Retrieved from `psyarxiv.com/9qv5y` doi: 10.31234/osf.io/9qv5y

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

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.

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.

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

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*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

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.

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.

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

Michalski, R. S. (1969). On the quasi-minimal solution of the general covering problem. In (Vol. A3, pp. 125–128).

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

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and brain sciences*, *37*(1), 1–19.

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*.

Pearson, K. (1930). *The life, letters and labours of francis galton*. Cambridge University Press.

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

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

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

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

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

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

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

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216.

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*.

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

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

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

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

Szollosi, A., Liang, G., Konstantinidis, E., Donkin, C., & Newell, B. R. (2019). Simultaneous underweighting and overestimation of rare events: Unpacking a

paradox. *Journal of Experimental Psychology: General*, *148*(12), 2207.

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

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).

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

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

Appendix A: Models

Generating context free (PCFG) model predictions

Here, 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. The grammatical primitives 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.[2] 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 the empirical samples, as described in the main text.

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.

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.

---

[2] 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. **?**).

Table A-2
Prior Production Process

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

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

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 3gh) and denoted with square brackets:

1. **Observe.** either:

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

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

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

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

   (b) For a single cone and an ordered feature, this could also be a nonstrict inequality ($\geq$ or $\leq$). We assume a learner only samples an inequality if it

[3] 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 3g&h for the relevant completions are normalized and used to select the option.

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

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

(b) **Disjunction.** An additional feature-value pair is selected uniformly from *either* unselected values of the current feature, *or* from a different feature — e.g., $\vee(=(x_1, \text{color}, \text{red}), =(x_1, \text{color}, \text{blue}))$ or $\vee(=(x_1, \text{color}, \text{blue}), \geq(x_1, \text{size}, 2))$. This step is skipped if the statement is already true of all the cones in the scene.[4]

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$\to$] combined with $[L\to]$ where appropriate. If it is true of only a subset of the cones then

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

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

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

Note that a procedure like the one laid out above is, in principle, capable of generating any rule generated by the PCFG in Figure

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.

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

Appendix B: Free response coding

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

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

Table A-3

Models of Participants' Generalizations

	Model	Group	log(Likelihood)	BIC	λ	τ	N	Accuracy
1.	Baseline	children	-1319.75	2639.50			7	50%
2.	Encoded Guess	children	-1143.69	2294.92		0.98	15	62%
3.	Similarity	children	-1316.44	2640.42		-0.50	0	41%
4.	PCFG Uniform	children	-1319.75	2647.05		-0.01	0	60%
5.	PCFG Off	children	-1318.85	2645.26		0.09	0	65%
6.	PCFG	children	-1319.57	2646.69		0.04	1	63%
7.	IDG Uniform	children	-1299.72	2607.00		0.55	2	66%
8.	IDG Off	children	-1304.92	2617.39		0.45	1	**70%**
9.	IDG	children	-1308.52	2624.59		0.39	2	68%
10.	Intercept	children	-1218.96	2445.47	0.32		**16**	50%
11.	**Encoded Guess + Intercept**	**children**	**-1067.18**	**2149.47**	0.26	1.24	9	
12.	Similarity + Intercept	children	-1214.71	2444.52	0.32	-0.77	1	
13.	PCFG Uniform + Intercept	children	-1210.30	2435.70	0.35	0.43	0	
14.	PCFG Off + Intercept	children	-1207.64	2430.39	0.34	0.48	0	
15.	PCFG + Intercept	children	-1208.74	2432.59	0.35	0.46	0	
16.	IDG Uniform + Intercept	children	-1195.19	2405.48	0.32	0.83	0	
17.	IDG Off + Intercept	children	-1193.01	2401.12	0.34	0.83	0	
18.	IDG + Intercept	children	-1194.19	2403.49	0.34	0.82	0	
1.	Baseline	adults	-1386.29	2772.59			2	50%
2.	Encoded Guess	adults	-893.49	1794.58		1.78	**32**	70%
3.	Similarity	adults	-1359.05	2725.70		-1.38	0	36%
4.	PCFG Uniform	adults	-1333.95	2675.50		0.69	0	62%
5.	PCFG Off	adults	-1293.60	2594.79		0.94	1	66%
6.	PCFG	adults	-1267.89	2543.38		1.06	2	69%
7.	IDG Uniform	adults	-1229.69	2466.97		1.50	2	69%
8.	IDG Off	adults	-1208.11	2423.83		1.52	0	73%
9.	IDG	adults	-1185.64	2378.89		1.62	1	**74%**
10.	Intercept	adults	-1364.90	2737.40	0.15		6	50%
11.	**Encoded Guess + Intercept**	**adults**	**-880.59**	**1776.38**	0.08	2.01	4	
12.	Similarity + Intercept	adults	-1337.55	2690.30	0.14	-1.63	0	
13.	PCFG Uniform + Intercept	adults	-1268.87	2552.93	0.26	1.35	0	
14.	PCFG Off + Intercept	adults	-1226.61	2468.42	0.25	1.60	0	
15.	PCFG + Intercept	adults	-1203.66	2422.53	0.24	1.69	0	
16.	IDG Uniform + Intercept	adults	-1179.02	2373.24	0.20	2.13	0	
17.	IDG Off + Intercept	adults	-1147.46	2310.13	0.22	2.26	0	
18.	IDG + Intercept	adults	-1131.92	2279.04	0.20	2.28	0	

NB: Accuracy column shows performance of the requisite model across 100 simulated runs through the task using participants active learning data and τ set to 100 (essentially hard maximizing over the model's predictions). The +Intercept models perform strictly worse due to their bias 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: Comparison with Bramley et al (2018)

Finally, for interest and to demonstrate replication of our core results. We provide a direct comparison between the generalization accuracies in the current sample of children and adults and those in the sample of 30 adults modelled in (Bramley et al., 2018). Bramley et al (2018) included 10 ground truth concepts, and the current paper uses just

the first five of these. Figure A-1 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.

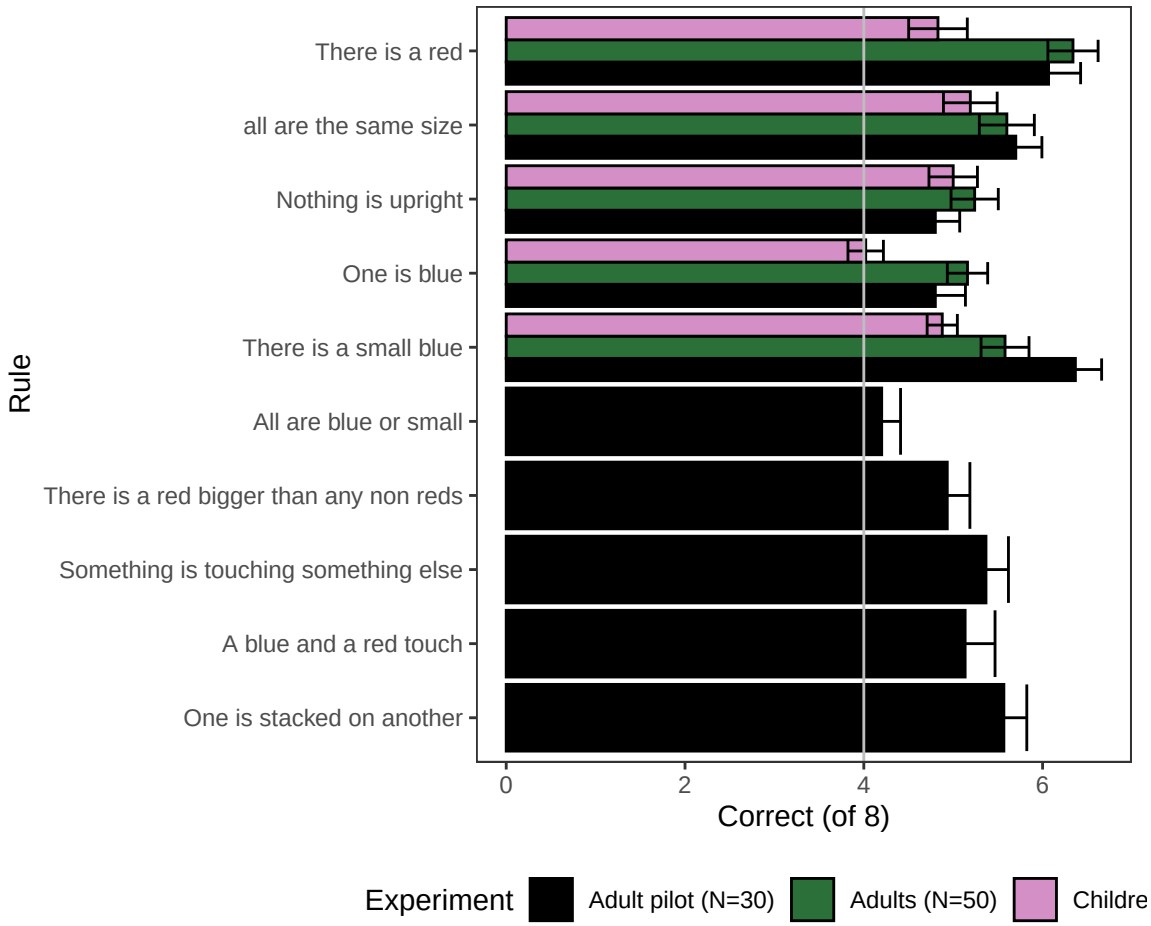


Figure A-1
Generalization accuracy by number of objects per test scene comparing with 10 rule adult pilot from Bramley et al. (2018).

Appendix D: 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-2 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-3 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-3a). 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-3b) but does not account for the fact that the compared scene may contain more or fewer objects in total. Figure A-3c visualizes just the object difference showing that children's scenes contain roughly as many objects on average as the initial example while adults' scenes contain around 0.75 fewer objects than are present in the initial example (dark shading in top row). Thus, we opted to combine b and c by weighting the unsigned cone difference by the mean inter-object distance across all comparisons to give our combined distance measure (Figure A-3d and Figure 6c in the main text).

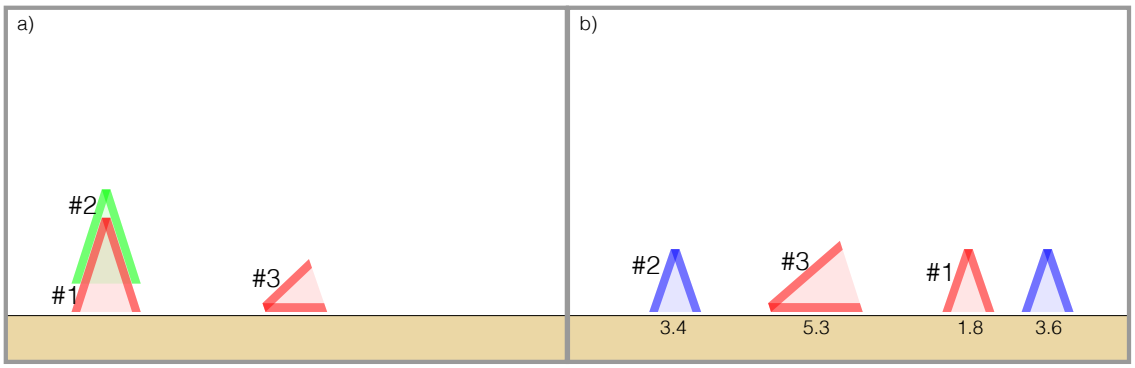


Figure A-2

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

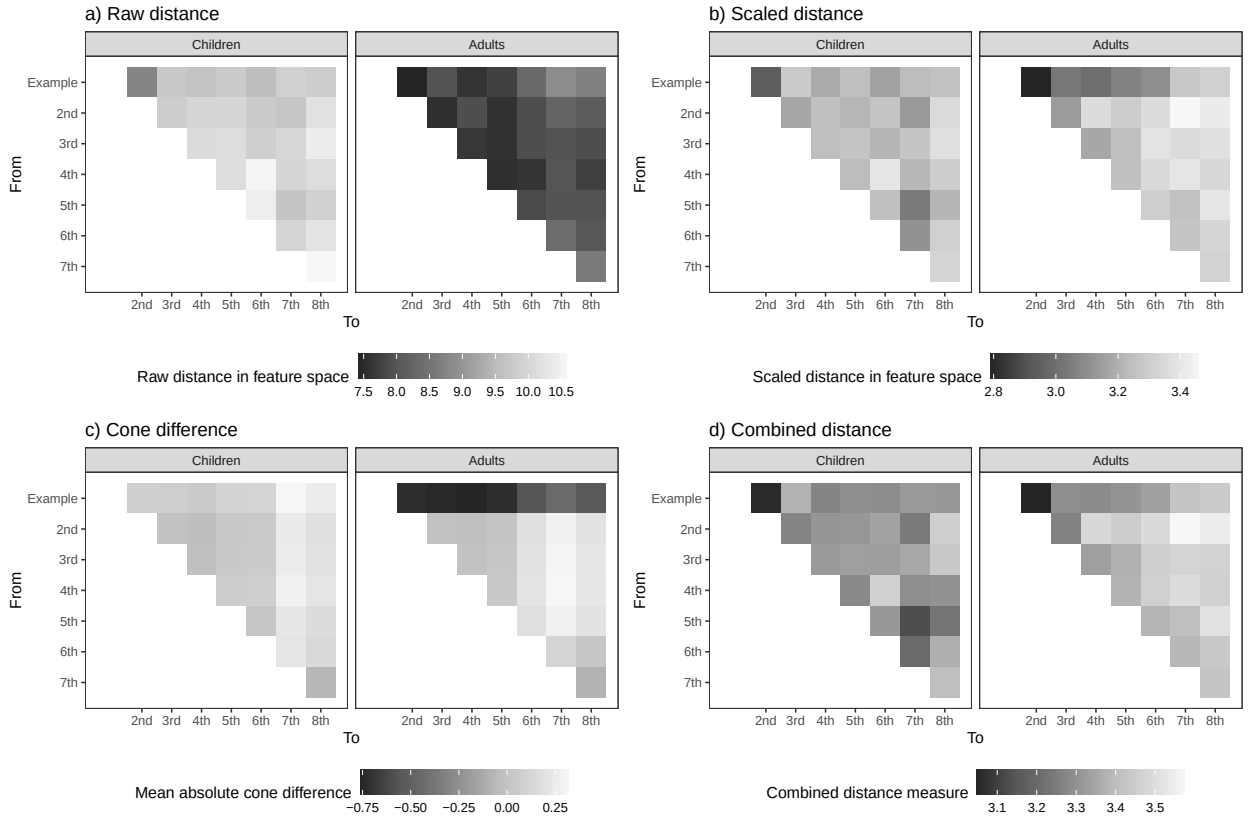


Figure A-3

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 E: Information gain analysis of active learning data

Children's and adults' scene generation patterns manifest in small differences in the quality of the total evidence generated according to an information gain analysis. For example, using the unweighted PCFG sample, prior entropy is 13.31 bits and children's evidence produces an information gain (reduction in uncertainty) of 6.86±0.55 bits while adults data allows for marginally higher information 7.04±0.44 bits $t(102) = -1.8, p = 0.068$. Relative to the fitted PCFG priors, the difference in information gains is rather larger, with children's scenes leading to information gain at 6.92±0.70 bits (prior entropy 12.94), and adults' at 7.50±0.66 (prior entropy 12.65) $t(102) = 4.4, p < .0001$. Under the mismatched priors — that is, taking the adultlike PCFG prior for children and childlike PCFG prior for adults — children's tests look slightly more informative than under their own prior, generating 7.14±0.72bits, and adults' tests slightly less informative than under their own prior 7.21±0.61 bits, eliminating the statistical difference $t(102) = 0.5, p = 0.62$. On the face of it, this is evidence against the idea that children's more elaborate hypothesis generation and concomitantly flatter latent prior is driving them rationally toward more elaborate testing patterns. However, we see this information-theoretic analyses as limited in what reveals. This is because is predicated on an implausibly complete representation of uncertainty, e.g. using a large sample of prior hypotheses, while we might expect constructivist search behavior to be driven by more focal testing of a smaller number of possibilities. Nevertheless, we present these information scores as norms for completeness and comparison with other active learning tasks.