Psychological Review

0033-295X/16/$12.00 http://dx.doi.org/10.1037/a0039980

The Logical Primitives of Thought: Empirical Foundations for Compositional Cognitive Models

Steven T. Piantadosi
University of Rochester

Joshua B. Tenenbaum
Massachusetts Institute of Technology

Noah D. Goodman
Stanford University

The notion of a compositional *language of thought* (LOT) has been central in computational accounts of cognition from earliest attempts (Boole, 1854; Fodor, 1975) to the present day (Feldman, 2000; Penn, Holyoak, & Povinelli, 2008; Fodor, 2008; Kemp, 2012; Goodman, Tenenbaum, & Gerstenberg, 2015). Recent modeling work shows how statistical inferences over compositionally structured hypothesis spaces might explain learning and development across a variety of domains. However, the primitive components of such representations are typically assumed a priori by modelers and theoreticians rather than determined empirically. We show how different sets of LOT primitives, embedded in a psychologically realistic approximate Bayesian inference framework, systematically predict distinct learning curves in rule-based concept learning experiments. We use this feature of LOT models to design a set of large-scale concept learning experiments that can determine the most likely primitives for psychological concepts involving Boolean connectives and quantification. Subjects' inferences are most consistent with a rich (nonminimal) set of Boolean operations, including first-order, but not second-order, quantification. Our results more generally show how specific LOT theories can be distinguished empirically.

Keywords: language of thought, concept learning, Bayesian modeling

One of the most powerful features of human cognition is our ability to create, manipulate and communicate novel structured ideas—concepts such as *prime number*, *half-sister*, *the tallest building in Cambridge*, *most*, or *most but not all*. Such concepts are interesting to cognitive psychologists because they are at the heart of how we humans think flexibly and productively, extending our thinking into novel situations, tasks, and domains. They are also central to how we produce and comprehend language. From a computational viewpoint, these concepts are interesting in part because they can be characterized using logical machinery: sketching informally, the Green Building is the tallest building in Cambridge if for all other buildings *b*, the Green Building is taller than *b*; two girls are half-sisters if there exists exactly one parent that they share; most As are Bs if (according to one standard account) the cardinality of the subset of As that are Bs is larger than the subset of As that are not Bs (Montague, 1973). Human fluency with such concepts likely reflects deep computational properties of both thinking and language, suggesting to many theorists (Hahn & Chater, 1998; Gentner, 1983; Penn, Holyoak, & Povinelli, 2008; Kemp, 2012; Goodman, Tenenbaum, & Gerstenberg, 2015) ways in which conceptual systems cannot be limited to stored examples and prototypes, but in some form must incorporate rule-like representations that are defined in terms of abstract operations and that compose flexibly to create new such representations.

Different branches of cognitive science have developed complementary approaches to study compositionally in human thought. For example, in linguistics, formal semantics has sought to characterize the logical structure necessary to capture reference for theoretically interesting fragments of natural language, often with a focus on quantifiers such as *some*, *all*, and *most* (Montague, 1973). In artificial intelligence, researchers have developed general-purpose architectures for human-like knowledge representation and reasoning based on predicate logic, often first-order logic but also using higher-order logics such as the lambda calculus, and often integrated with probability to support reasoning under uncertainty, inductive learning and abductive inference

Steven T. Piantadosi, Department of Brain and Cognitive Sciences, University of Rochester; Joshua B. Tenenbaum, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology; Noah D. Goodman, Department of Psychology, Stanford University.

This work was supported by an NSF Graduate Research Fellowship (to Steven T. Piantadosi) and a S.B.E. Dissertation Award in Linguistics (to Steven T. Piantadosi). This work was also supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award F32HD070544. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We are extremely grateful to Roger Shepard, Jesse Snedeker, Susan Carey, Roman Feiman, Evelina Fedorenko, Ted Gibson, and Leon Bergen for insightful and constructive conversations at various stages of this work. Data and code for this paper are available at http://colala.bcs.rochester.edu/data/PiantadosiTenenbaumGoodmanPsychReview.

Correspondence concerning this article should be addressed to Steven T. Piantadosi, Department of Brain and Cognitive Sciences, University of Rochester; 358 Meliora Hall, Box 270268, Rochester, NY 14627-0268. E-mail: spiantadosi@bcs.rochester.edu

(Levesque, Pirri, & Reiter, 1998; Muggleton & De Raedt, 1994; Milch, Marthi, & Russell, 2004; Russell & Norvig, 2009; Domingos & Richardson, 2007; Richardson & Domingos, 2006; Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008; Shapiro, Pagnucco, Lespérance, & Levesque, 2011). In cognitive psychology, the classic empirical method for studying compositional thought has used concept learning experiments, where participants learn various rule-based concepts from examples, and the rules are most naturally represented as simple or more complex functions of logical primitives (e.g., Bruner, Goodnow, & Austin, 1956; Shepard, Hovland, & Jenkins, 1961). By studying characteristic patterns of mistakes learners make, and which concepts are harder or easier to learn, researchers aim to discover something about the primitives and means of combination in human symbolic thought.

Our work here builds most directly on this cognitive psychology tradition of studying rule-based concept learning, but integrates elements of the formal semantics and AI traditions, along with new methods for computational cognitive modeling and web-based experimentation. This allows us to study the building blocks of compositional thought on a previously unprecedented scale, and to ask questions that have not previously been amenable to empirical inquiry. Specifically, we make three main contributions. We study empirically the dynamics of how people induce a much broader space of rules than previous work has examined, including over 100 distinct concepts varying significantly in complexity and logical structure. We also show that people's probabilistic generalization behavior in these tasks can be quantitatively well described by a memory-constrained Bayesian learning model. Finally, and most importantly, we show how the basic building blocks of the model's compositional hypothesis space can be inferred from the large-scale patterns of participants' responses. This yields insight into the primitive representations and operations in the combinatorial language of thought that people bring to this task, and presumably other settings of symbolic thinking and reasoning.

We build on recently developed computational tools and empirical techniques that have allowed detailed modeling of how people learn symbolically structured concepts across a variety of domains (e.g., Nosofsky, Palmeri, & McKinley, 1994; Kemp, Goodman, & Tenenbaum, 2008a; Kemp, 2009; Piantadosi, 2011; Kemp, 2012; Ullman et al., 2012). This work formalizes learning as some kind of rational inductive inference over a compositional representation system, or *language of thought* (LOT; Fodor, 1975; Boole, 1854). For instance, learning models might posit that people initially have access to simple logical operations (e.g., *and*, *or*, and *not*), quantifiers (like *forall* and *exists*), or computational primitives (e.g., λ-abstraction or combinators). The task of learners is then to compose these primitives into a rule denoting a concept that fits the observed examples in some approximately optimal way, as measured by posterior probability in a Bayesian framework or encoding length in an information-theoretic framework. More broadly, beyond their ability to describe how people learn rule-based concepts in specific laboratory tasks, models of this sort are theoretically appealing as a complement to other frameworks for modeling human learning such as connectionism, because of their ability to explain how inductive learning integrates with compositionality, productivity and other core features of symbolic cognition (Fodor & Pylyshyn, 1988; Fodor, 2008; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Goodman, Tenenbaum, Feldman, & Griffiths, 2008).

In a prototypical example of this approach, Goodman et al. (2008) presented a rational analysis of rule-based concept learning based on Bayesian induction of the Boolean expressions (built from connectives like *and*, *or*, and *not*) most likely to have generated the observed example labels for a concept, and showed how this could explain both classic results in the literature and graded patterns of generalization in new experiments with higher-dimensional feature spaces. Their approach has been extended to learning richer logical theories (Kemp et al. 2008a; Kemp, 2012), as well as conceptual domains like magnetism (Ullman et al., 2012), semantic hierarchies (Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008), number concepts (Piantadosi, Tenenbaum, & Goodman, 2012), and function words (Piantadosi, Goodman, & Tenenbaum, 2014). Kemp (2012) provides an exhaustive characterization of logical domains such models can be formulated over (a "conceptual universe") and shows that a rule inference scheme based on minimal description length—closely related to the Bayesian posterior probability—provides a compelling quantitative fit to behavioral data across a wide range of conceptual domains.

In these and most other prior efforts, the components and rules of combination, which together characterize a representation system, have been assumed a priori with the analysis focused on demonstrating the in-principle viability of compositional, LOT-style learning. For instance, Goodman, Tenenbaum et al. (2008) assumed that learners form compositions in *disjunctive normal form* (disjunctions of conjunctions of primitives). This assumption amounts to a psychological hypothesis about the format of conceptual representations: it assumes that learners construct Boolean rules as disjunctions of conjunctions of features—as opposed to, for instance, conjunctions of disjunctions, horn clauses, or free combination of logical operations. Beyond the rules of combination, such theories also formalize a set of *primitives*, functions that are assumed to be available before the process of compositional hypothesis formation begins. It is natural to think of these primitives as the innate conceptual representations that learners use to build complex concepts. However, the effective set of primitives may also include ones defined using the innate set, and thus include operations that have been learned at an earlier age. As we study adults, the primitives can be seen as the key operations that are available to our subjects in understanding new logical concepts.

In the case of Boolean concept learning, there are many logically possible sets of primitives that have been used across different domains of cognitive science, AI, and machine learning. In one extreme, all Boolean functions can be constructed from one single primitive such as *NAND* (*not-and*, also known as the "Sheffer stroke"). In the other extreme, Boolean concepts might psychologically depend on a rich set of redundant primitives, including operations like conjunction, disjunction, negation, implication, and so forth as are used in modern computer architectures. The problem of the underdetermination of representational primitives is faced even more broadly for systems that extend beyond Boolean concepts: there are many possible primitives that would permit Turing-complete computational concepts. A coarse characterization of the right system can be made in terms of computational power: representations must be capable of supporting the knowledge people have and the computations they perform (Marr, 1982). For instance, human representational systems must extend beyond simple Boolean propositional logic since such systems that lack quantification provably cannot express concepts like *tallest*. However, descriptions based only on computational power are always

underdetermined. As the Boolean case illustrates, two representations can be equally expressive—capable of solving the same computational problems—yet distinct in how they achieve that computational power (see, e.g., Hackl, 2009; Pietroski, Lidz, Hunter, & Halberda, 2009, for examples in semantics). Any potential set of primitives and rules for combination can be regarded as a scientific hypothesis about how compositional concepts may arise in mind—a hypothesis that a mature LOT framework should seek to evaluate empirically.

Here we present a formal modeling approach that aims to discover which particular LOT features—which primitives and rules of combination—provide the best account of how people learn and reason in a given domain of compositional thought. Our approach crucially exploits the classic insight that representational *simplicity* is a major determinant of learnability, with learners preferring to infer rules that are concise in their representational system (Neisser & Weene, 1962; Haygood, 1963; Feldman, 2000, 2003c, 2003b; Chater & Vitányi, 2003; Goodman, Tenenbaum et al., 2008; Kemp et al. 2008a; Kemp, 2012). Analogously in modeling, simplicity plays a key role in model selection because simple models are more parsimonious (e.g., Conklin & Witten, 1994), explaining the data with fewer free parameters or arbitrary stipulations. This property makes a simplicity bias a sensible psychological strategy.

The existence of a human bias for representational simplicity allows us to potentially reverse engineer likely components of the LOT since different hypothesized LOTs will measure simplicity in different ways. As a result, they make different empirical predictions about what generalizations participants should make from data and what concepts should be easy to learn (e.g., what concepts are "simple"). A key example for our purposes is Feldman (2000), who showed that difficulty with learning Boolean concepts is well-modeled by the concept's description length in logic. For example, participants would find it harder to learn the exclusive disjunction (*XOR*),

$$[red\ AND\ [NOT\ square]]\ OR\ [[NOT\ red]\ AND\ square] \quad (1)$$

than the nonexclusive disjunction

$$red\ OR\ square \quad (2)$$

since the former has a longer description in standard Boolean logic, requiring more primitive logical connectives.

However, as is often pointed out in philosophical discussions of induction (Goodman, 1955), what counts as "simple" is not purely objective. For instance if people's representational system included the exclusive-or function (*XOR*) as a primitive, then the complexity—and therefore learning difficulty—for the above two concepts would be equal. Concept 2 could be expressed the same way, but Concept 1 becomes

$$red\ XOR\ square \quad (3)$$

which uses only a single logical operation. This only emphasizes a version of Goodman's *Grue problem*: Concept 1 is not more complex than Concept 2 in any independent, objective sense.[1]

This philosophical puzzle is also an experimental *tool*: if participants do find Concept 2 easier than Concept 1, that provides evidence that human cognition measures simplicity via a LOT in which *XOR* is not primitive. Here, we take this simple idea and implement it in a large-scale computational and experimental study examining a wide range of concepts with state-of-the-art learning and data analysis tools. Building on Kemp (2009, 2012), and motivated by both classic work in formal semantics (Montague, 1973), AI (Levesque et al., 1998; Muggleton & De Raedt, 1994; Milch et al., 2004; Russell & Norvig, 2009; Domingos & Richardson, 2007; Richardson & Domingos, 2006; Goodman, Mansinghka, et al. 2008; Shapiro et al., 2011), and more recent cognitive accounts of symbolic thinking (Gentner, 1983; Penn et al., 2008), our study moves beyond simple Boolean concepts and also examines those involving quantification and relational terms.

The outline of this paper is as follows. In the next section, we present a large-scale concept learning experiment that taught participants Boolean and quantificational concepts. We then describe how we formalize the LOT in terms of λ-calculus, and then present two coordinated models. First, we develop a learning model that—like participants in the study—takes observed data and infers likely LOT expressions. As we show, the learning model is capable of inferring quite complex concepts from data, and the generalizations it makes closely track those of participants. Then, we present a data analysis model that uses participants' experimental data to infer unknown parameters of the learning model including the probability of different primitives and participants' memory-decay parameters. The details of these models are provided in the Appendices, and we focus in the main text on intuitively describing their key properties. Our primary analysis is a model-based comparison that quantifies the fit of different LOTs to human learning patterns. We first apply these methods to only Boolean concepts in the experiment, and then to concepts involving quantification. Our results provide quantitative evidence against intuitively implausible logical bases, evidence for compositional logical rules with a rich set of logical connectives that include first- but not second-order quantification. Most importantly, our method shows how distinct LOTs can be firmly grounded as empirically testable scientific theories.

Before moving into the body of the paper, we should delimit our goals in several ways to avoid potential misunderstandings. Although our focus is on symbolic thinking, and the logical structure of the "language of thought" that underlies it, we do not mean to imply that all or even most human thinking is of this form. Quite the contrary: we grant that much of how people think about the world draws on other kinds of representations, such as perceptually grounded simulations (Battaglia, Hamrick, & Tenenbaum, 2013) or probabilistic expectations (Griffiths & Tenenbaum, 2006). Likewise, although we focus on a certain class of concept learning tasks with concepts defined by logical rules, we do not mean to suggest that all or even most human concepts take this form. Many concepts, especially those for basic-level natural kind categories, may be best thought of in other ways (Rosch & Mervis, 1975; Medin & Smith, 1984; Medin, 1989; Medin & Ortony, 1989; Hampton, 1998; Murphy, 2002; Hampton, 2006). Different categorization tasks may even rely on distinct systems and processes (Ashby & Maddox, 2005).

[1] Tools like *Kolmogorov complexity* (Li & Vitányi, 2008) come close to resolving this problem, providing a complexity metric that is arbitrary only up to an additive constant.

However, we do think that logically structured concepts have sometimes been unfairly maligned as "unnatural". The evidence for a rich ability to process logical concepts can be seen in many domains (e.g., Tenenbaum et al., 2011), including number and mathematics, social systems, taxonomies, and complex causal processes. The need for structured concepts becomes even more evident in natural language, where languages contain words to express a variety of logical relations, whose meanings are typically captured in formal theory only in with structured, logical systems. To illustrate in English, these words include quantifiers (e.g., *every*) and other determiners (e.g., *the*), conjunctions (e.g., *and*), kinship terms (e.g., *great uncle*), prepositions (*in*), and markers of discourse relations (e.g., *because*) expressing relations between clauses. Below the level of words, morphemes like *–est* combine with words to form superlatives whose meaning is most naturally captured with logic: someone is the "tall*est*" if their height is greater than everyone else, a sublexical concept involving first-order quantification. The full power of abstract logical structure can be seen in the compositional phrases formed in natural language—phrases like "the tallest building in Cambridge" combine simpler, constituent meanings into complex logical structures that are able to communicate a huge variety of meanings. There is logical structure in language even above the level of sentences, including in the discourse relations between sentences (Wolf & Gibson, 2005) and in recursive patterns of dialogue and pragmatics (Levinson, 2013). Our goal here is not to account for the full set of phenomena that cognitive psychologists have been interested in under the banner of "concepts" (Margolis & Laurence, 1999; Murphy, 2002), but rather to better characterize computationally those aspects of human conceptual thinking and learning that are broadly accepted across the cognitive sciences to depend on compositional language-like representations. Ultimately, we expect that a full theory of human concepts and thinking will need to integrate the kinds of approaches we develop with complementary approaches developed for studying non-rule-like concepts and nonsymbolic thought.

Experimental Paradigm

Our experiment aims to study concept learning in a domain that naturally captures both classic Boolean concept learning (e.g., Shepard et al., 1961) as well as richer types of relational and quantification concepts (e.g., Kemp, 2009, 2012). We framed the problem as one of mapping a set of objects in a feature space to a subset of those objects. For instance, one might be handed a set of objects and be asked to give back the ones that are *red or green*, a Boolean concept. Or, one could be required to hand back all objects such that *there exists another object in the set of the same shape*, a quantificational concept. This set-to-subset concept is reminiscent of the set-relational operators required for natural language semantics.

Rather than exhaustively explore the entire range of logically possible concepts (as pursued by Feldman, 2003a; Kemp, 2009, 2012), we chose to construct a space of target concepts by hand in order to focus on a particularly compelling variety of concepts. Choice of concepts by hand is both a strength and a limitation of our design. It means on the one hand that the concepts we study are ones that we believed a priori were interesting and would reveal the kinds of logical operations (e.g., quantification, logical combination) that most interest us. On the other hand, it means that our chosen concepts may not be representative of any natural category of human concepts. We believe this is a necessary property of work such as this that is very early in the effort to model operations such as quantification.

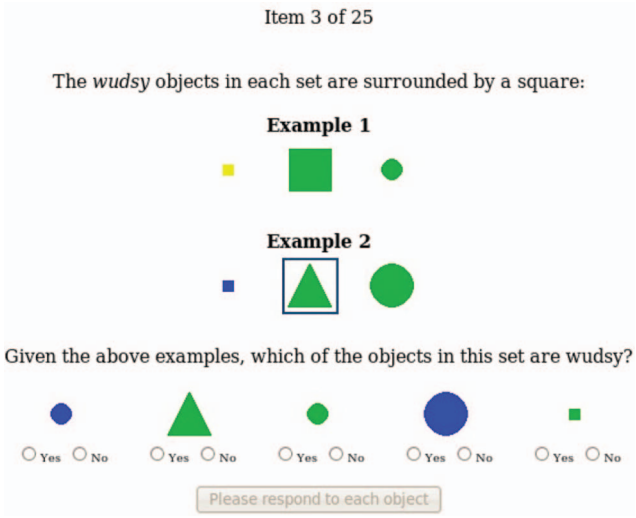

Figure 1. An example item from the concept learning experiment. Here, the participant has seen two example sets of objects, and is asked to generalize to a new set. A likely response here would be to answer in accordance with the simple concept *triangles*. See the online article for the color version of this figure.

Our set includes 108 concepts that were chosen to span a wide range of quantification and relational operations, including basic Boolean concepts (e.g., *blue objects*) and quantificational/relational terms (e.g., *the unique blue object*, *same shape as a blue object*, *every other object with the same shape is blue*, etc.). The full set of concepts is listed in Figures 2, 3, and 4.

Method

In the experiment, participants were told that they had to discover the meaning of *wudsy*, a word in an alien language. They were explicitly told that this word applied to some objects in a set, and that whether or not an object was *wudsy* might depend on what other objects were in the set. The learning paradigm was sequential: participants were shown a set and asked whether each item was *wudsy*. After responding, they were shown the right answers. The correctly labeled sets stayed visible on the screen, and participants moved on to the next set. This means that on set N, a participant could still see the correct answers to the previous $N - 1$ sets. Thus, the participant's Nth response represents their inferences conditioned on the previous $N - 1$ labeled data points. This continuous measure of generalization contrasts with previous Boolean concept learning paradigms which have typically tested only after a fixed amount of training. Our paradigm allows a substantial amount of inductive generalizations to be gathered, providing a detailed picture of learning curves and specific patterns of mistakes. An example experimental item is shown in Figure 1, showing participants being asked to generalize to a set containing five

elements after seeing the two preceding sets, only one of which contained a positive instance of a *wudsy* object.

To aid in motivation, participants were required to wait 5 seconds when they made a mistake in any element of a set. The space of objects included squares, circles, and triangles, that were green, blue, or yellow. Object sizes ranged through three logarithmically spaced sizes, denoted here Size 1 (smallest), 2, and 3 (largest). Sets were generated from this space of objects at random by first uniformly choosing a set cardinality between 1 and 5, and then randomly sampling objects without replacement. Random generation was used to ensure that participants do not assume sets were chosen to be informative about target concepts (as in, e.g., Shafto & Goodman, 2008). Subjects were shown 25 sets of objects in total, requiring on average a total of 75 responses to individual objects.

Subjects were randomly assigned to a concept and one of two lists in that concept, where each list was a different sequence labeled according to the target concept. Within the same list, the presented sequence of data was identical. Subjects were allowed to do multiple concepts, but could not repeat the same concept twice. The small number of lists allowed us to run more participants within each list to get higher confidence in the exact learning curves for any particular sequence of labeled data. The specific shapes and colors in each target concept were randomized across participants. For example, *red and circle* was randomized to *blue and triangle*, *blue and square*, *green and circle*, and so forth across participants.

At sets 5, 10, and 25, participants were asked to describe what they thought *wudsy* meant. In general, verbal descriptions proved extremely difficult to analyze because participants often wrote ambiguous descriptions. For instance, we ran concepts such as *the unique tallest* (cannot be tied for tallest shape in the set) and *one of the tallest* (can be tied for tallest shape in the set). Subjects with both of these concepts wrote "tallest," which, in English, might mean either concept. While we do not analyze this linguistic data here, the ambiguity in participant descriptions provides some evidence that participants did not represent target concepts in natural language—doing so often leaves the target concept underspecified.

Subjects were run online using Amazon's Mechanical Turk. Subjects who fell more than 2 standard deviations below the mean accuracy in their concept were removed. This removed on average only 3.9% of data from each concept (standard deviation = 2.6% across concepts). Data from participants who completed fewer than five sets of a given concept was also removed, but otherwise partial data from participants was included in our analysis. A total of 1,596 participants were run across the 108 concepts. While we did not gather detailed demographics on subjects, a study of the population of people who tend to complete Turk experiments can be found in Berinsky, Huber, and Lenz (2012), who show that the study population is often more representative of the U.S. population than in-person convenience samples typically used in research (see also Paolacci, Chandler, & Ipeirotis, 2010; Behrend, Sharek, Meade, & Wiebe, 2011). In our experiment, individual participants completed an average of 4.24 concepts (median 2), with the maximum number of concepts run by a participant at 80. Overall accuracy in the experiment was 78% with a chance rate of 56%, though the accuracies varied substantially by concept. Mean accuracies on concepts were highly consistent across the two lists, with a correlation of $R^2 = 0.81$. Subjects were run until each concept and list was completed by approximately 20 participants.

Model-Free Results

Though our primary analysis is model-based, here we present the general patterns in the learning experiment. Figures 2–4 list the 108 concepts tested as well as raw participant performance on these concepts. Performance here is computed as the mean accuracy across the fixed number of trials in each concept. Each horizontal line in these figures shows a compressed learning curve with two points, representing mean participant performance on the first and last quarter of the experiment. Each figure shows one third of the total concepts tested, sorted by overall mean accuracy: Figure 2 shows the most easily learned concepts, Figure 3 shows concepts that are likely learnable with some difficulty, and Figure 4 shows concepts that are extremely difficult to learn, some of which may not have been learned by anyone over the course of the experiment. These plots also include blue bars corresponding to chance performance. Chance was computed by assuming that participants guess with the correct base rate: thus, if the concept is true of set elements 30% of the time, then participants matching the base rate would be correct with probability $0.3 \cdot 0.3 + (1 - 0.3) \cdot (1 - 0.3)$. The horizontal lines in these plots are green for simple Boolean concepts and black for concepts that have no equivalent in Boolean logic.

This graph demonstrates several basic patterns previously found in Boolean concept learning. For instance, complex concepts (*circle and blue*) are learned less quickly than simple ones (*circle*) (Feldman, 2000). The graph also shows that conjunctions (*circle and blue*) are easier than disjunctions (*circle or blue*). The *and/or* asymmetry is one of the oldest findings in rule-based concept learning (Bourne, 1966; Shepard et al., 1961). These results also suggest selective attention effects where multiple references to the same feature dimension (*blue or green*) are easier than references across dimensions (*circle or blue*). Figure 2 also demonstrates rapid high performance for many non-Boolean concepts—for instance, *the unique element and is* (blue and circle), *one of the smallest*, *exists another object with the same shape and color*.

The richness of our data is more clearly illustrated by Figure 5, which shows response patterns to six examples concepts: Figure 5a and 5b from the upper most accurate third of concepts, Figure 5c and 5d from the middle third, and Figure 5c and 5f from the lower third. Each subfigure shows a participant on a row, and their response to each object in each set over the course of the experiment. Black squares in this plot represent incorrect responses, and white represent correct responses. Columns on the left of these plots correspond to early responses in the experiment; columns on the right correspond to later ones when correct answers for all previous (leftward) examples have been observed. So for example, the top two rows of Figure 5c show two participants who did not learn the target concept and made mistakes throughout the entire experiment. The rows (participants) in these plots are sorted by clustering to reveal participants whose patterns group together. The blue and gray bar at the bottom is a key showing which individual responses were responses to objects in the same set: adjacent columns with the same color in the key were correspond to objects presented in the same set. Thus, for instance in Figure 5a, the first set contained four objects (four blue in a row in the key), the second contained one (one gray), the third contained five objects (five blue), and so forth.

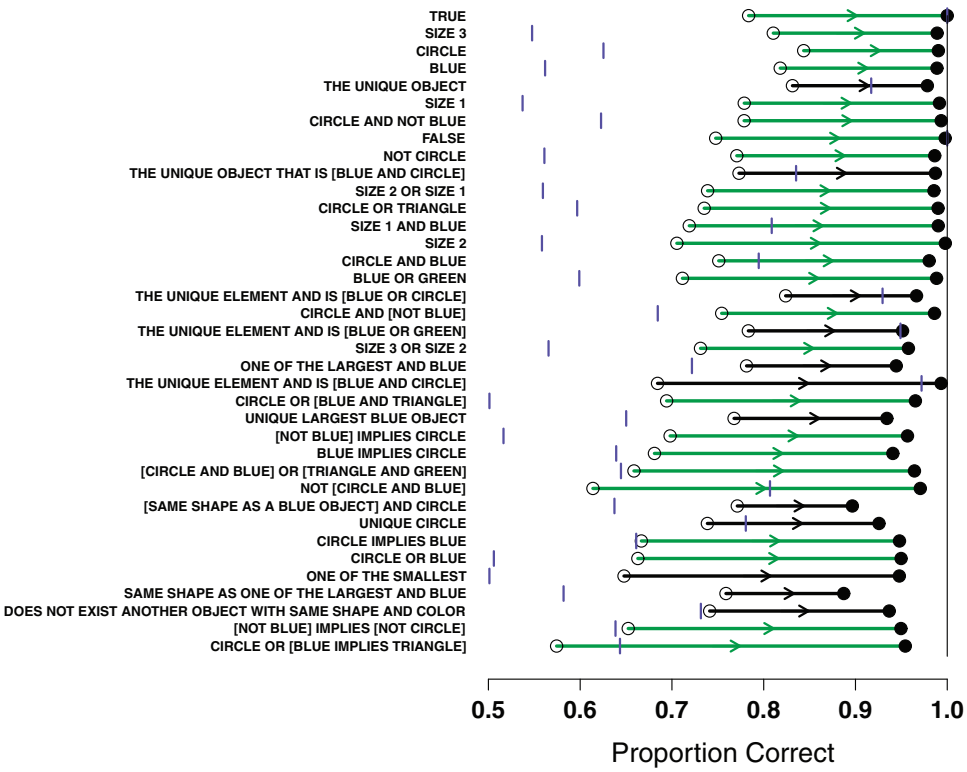

Figure 2. Proportion correct on the first 25% of the experiment (open circle) and last 25% (closed circles) for the top third of concepts most easily learned. Green (gray) lines denote concepts that can be written in simple Boolean (propositional) logic. Blue (black) bars denote chance guessing at the correct base rate. See the online article for the color version of this figure.

This plot illustrates several interesting patterns. Even in concepts that participants readily learn (e.g., Figure 5a), they still make occasional errors. These errors, however, appear not to be systematic across participants or sets. In other situations, such as Figure 5d, participants make highly systematic patterns of mistakes, often incorrectly labeling the same elements of the same sets (appearing here as black vertical lines). There are three participants in the middle of this plot, however, who appear to correctly get the target early concept and answer perfectly for most of the experiment. This pattern is also found in harder concepts, Figure 5e and 5f, where only a few participants achieve high accuracy and the rest appear to make regular patterns of responses. Note that even though mean accuracy is low on these concepts participants show clearly systematic patterns of inference. This motivates data analysis based on predicting specific generalizations—responses to individual items in sets—rather than overall concept accuracy.

These plots also demonstrate that while the average participant may show graded performance, individual participants likely have very rule-like hypotheses in mind (Nosofsky et al., 1994): at some point, participants appear to "get it" and respond perfectly or nearly perfectly for the remainder of the experiment. Subject averages, though, show more gradual patterns of learning since participants often "get it" in slightly different places. Thus, it is important to recognize that a model of average learning does not necessarily represent how individuals act. The individual differences in this task may represent distinct strategies and likely further insight could be gained by models of individuals (Bruner et al., 1956; Levine, 1966; Ashby & Maddox, 2005; Gluck, Shohamy, & Myers, 2002; Visser, Raijmakers, & van der Maas, 2009; Visser, Jansen, & Speekenbrink, 2010). We describe and justify the linkage between our model and subject data below.

Motivation for a Model-Based Analysis

Although we can compare learning rates and accuracies for these concepts, such comparisons are not obviously straightforward and useful. For one, the concepts vary in their base rate accuracy (blue points) and so it is difficult to know if differences in accuracy result from difference in chance performance. Worse, though, is that participants can achieve high accuracy on concepts not by learning the correct concept but by learning a closely related one. Subjects may, for instance, learn *one of the largest* (the object can be tied for largest) for *the unique largest* (the object cannot be tied for largest). A third problem is that it is not clear how informative learning rates are for comparisons since the observed data may be differentially informative as to the target concept. For instance, an ideal learner who is equibiased between *circle and blue* and *circle or blue* may nonetheless find the former easier to learn because it is true less often, meaning that the positive examples may be more diagnostic for the target concept, or perhaps maybe more psychologically available.

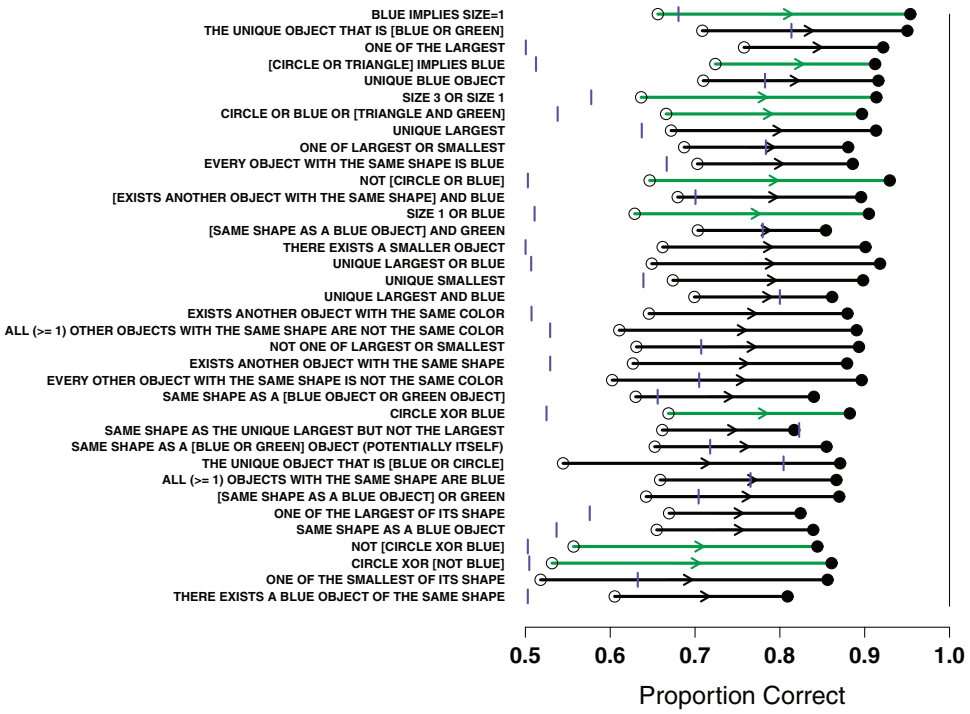

Figure 3. Proportion correct on the first 25% of the experiment (open circle) and last 25% (closed circles) for the second third of concepts most easily learned. Green (gray) lines denote concepts that can be written in simple Boolean (propositional) logic. Blue (black) bars denote chance guessing at the correct base rate. See the online article for the color version of this figure.

Note that from these accuracies, we are not able to identify which particular concepts subjects learned due to similar issues. Indeed, we do not even need to try: instead we can model their individual responses with a formal rule-learning model. Different rule-learning models will predict different patterns of responses, and we can determine which best describe people's responses. In the next section, we describe the learning model that attempts to capture the details of the learning curves rather than overall patterns of accuracy.

Languages of Thought

Modeling participant inferences in the experiment requires a formal system for expressing both concepts and hypothesized LOTs. For this, we use λ-*calculus* (Church, 1936; Hindley & Seldin, 1986; Smullyan, 1985), a computational system for elegantly expressing compositions of functions. This framework allows for a flexibility in theorizing that is not possible in other systems such as propositional or first-order logic because it permits any degree of computational power, ranging from the expressiveness of Boolean logic, up to Turing-completeness. In previous work, we (Piantadosi et al., 2012; Piantadosi, Goodman, Ellis, & Tenenbaum, 2008, 2014) and others (Zettlemoyer & Collins, 2005; Liang, Jordan, & Klein, 2010) have demonstrated the computational plausibility of learning in frameworks like λ-calculus.

For the present purposes, the consequences of choosing λ-calculus are relatively minor: it primarily allows us to explicitly express compositions of functions, including syntax to denote what parts of expressions are variables and which are functions. For instance, the λ-expression

$$\lambda x \ . \ (f \ (g \ x)) \tag{4}$$

represents a *function* of a variable x, computed by first applying g to x (denoted $[g\ x]$) and then applying f to $(g\ x)$, yielding $(f\ (g\ x))$. Here, the "λx" simply denotes that x is a variable.

We formalize a number of *grammars* for λ-calculus expressions, each formalizing a distinct representational theory about which compositions are cognitive permissible and likely. An example λ-calculus grammar for propositional logic is shown in Figure 6a. Each row in this table represents an expansion rule: the left hand side is a *type* and the right hand side is an expression that the type expand to. Thus, for instance, we could create the expression λ*x.* (*or (green? x) (blue? x)*) by first expanding the *START* symbol with *START to* λ*x.BOOL.* We then expand the *BOOL* in the right hand side of λ*x.BOOL* to (*or BOOL BOOL*), yielding the intermediate expression λ*x. (or BOOL BOOL*). Then, each of these *BOOL*s are expanded to (*F x*) yielding λ*x.* [or *(F OBJECT) (F OBJECT)*]. Finally, the first *F* is expanded to *COLOR*, then *green?* and the second *F* is expanded to *COLOR* and then *blue?*. Both *OBJECT*s are expanded to *x*es, yielding the full expression.

Using the grammar, one expands expressions until contain no more nonterminals, the uppercase symbols on the left side of this grammar. These grammars are meant to capture a core generative capacity of learners—that learners can in principle construct a huge number of potential concepts corresponding to every expression derivable in the grammar. The appeal of the grammar is that

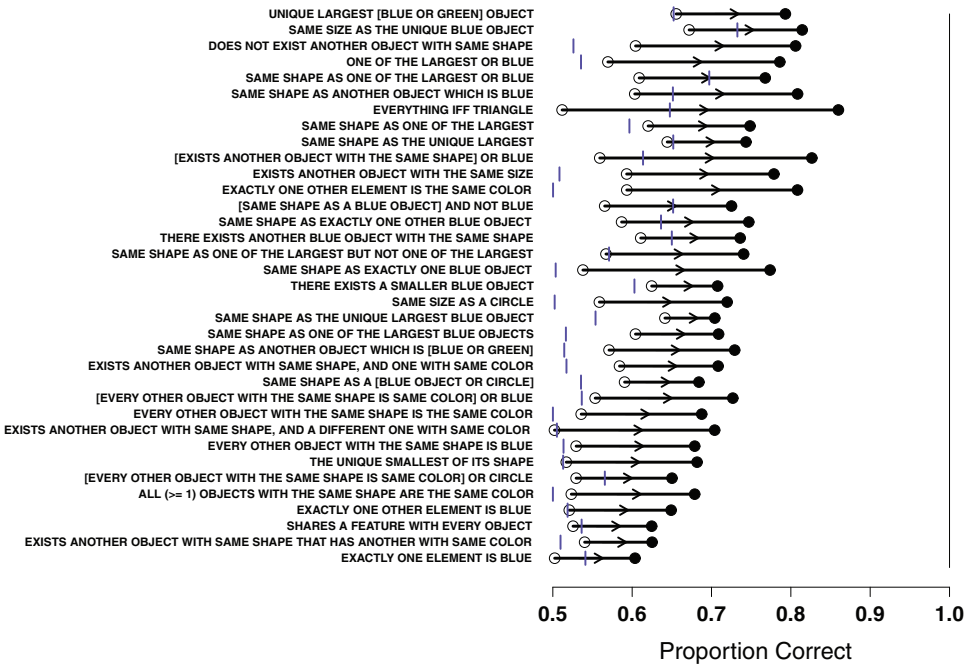

Figure 4. Proportion correct on the first 25% of the experiment (open circle) and last 25% (closed circles) for the third of concepts hardest to learn, none of which are simple Boolean expressions. Blue (black) bars denote chance guessing at the correct base rate. See the online article for the color version of this figure.

a rich potential for concepts comes from a simple generating system. The majority of rules in the grammar are actually methods of accessing perceptual primitives. The core logical or computation parts of the grammars are few—in this case, just three.

However, there are many other ways to write down expressions in Boolean logic, corresponding to different LOTs. Figure 6b shows one other: the NAND grammar uses only a single logical connective, *NAND*, yet can provably express all concepts the SIMPLEBOOLEAN grammar Figure 6a can. For instance (*OR A B*) can be computed with NAND as (*NAND (NAND A A) (NAND B B)*). As discussed above, these two grammars provide distinct representational hypotheses of equivalent computational power, but distinct computational processes and give rise to distinct inductive biases because they differ in what counts as simple (short derivations in the grammar).

We can also define grammars that go beyond simple Boolean logic. Figure 7a, defines a grammar that includes simple first-order quantification. Here, we introduce two more functions that return truth values, *exists* and *forall* ($\exists$ and $\forall$ respectively). These functions themselves take a *function*, *F*, as an argument, as well as a set. *exists* returns true if its argument *F* evaluates to *true* on any element of the set; *forall* returns true if *F* evaluates to *true* on all elements of the set. We must therefore include sets to quantify over. Here we choose two that are natural: the set of all elements in the current set, *S* (e.g., the range of *x*), and the set of all elements in the context other than the current argument *x*, (*non-Xes S*). Note here that *F* can only be expanded according to the features defined in Figure 6a, namely size, shape, and color. Using Simple-FOL we may write concepts such as *There exists a red object* [λx. *(exists red? S)*] or *There exists a triangle other than x* [λx. *(exists triangle? (non-Xes S)*].

A much more interesting kind of quantification can be created if the grammar can potentially define new functions from sets to objects. FOL is one such grammar, where now *F* can expand to a *new* λ-expression using the rule $F \rightarrow (\lambda x_i.BOOL)$. Such a grammar is shown in Figure7b. This reveals the power of λ-expressions: we can use the same syntactic form to specify functions, and functions of functions, and so forth.

In this case, we can create concepts like *There exists a red object that is the same shape as x in (non-Xes S*):

$$\lambda x\, S\ .\ (exists\ (\lambda x_2\ .\ (and\ (red?\ x_2)\ (equal\text{-}shape?\ x\ x_2)))\ (non\text{-}Xes\ S)). \quad (5)$$

Here, the *F* on the right hand side of the rule for *exists* was expanded to a new λ-expression, (λx_2 . (*and (red? x_2) (equal-shape? x x_2)*), itself representing a new function that learners might hypothesize. This ability to define new functions introduces a small technical complication: every time a new λ-expression is created, it requires a name for a new bound variable, here x_i (for $i = 1, 2, 3, \ldots$). To deal with this, any time a λ-expression is generated, we also add a rule to the grammar that expands *OBJECT* to the new x_i for all lower nodes (e.g., nodes contained by the novel λ-expression). For simplicity, we make all expansions to any x or x_i equally likely. With this setup, our actual grammar is not a context-free grammar, but is closely related: the expressions without distinct labels on the bound variables are context-free, and the bound variables are uniformly generated from those that are possible at each depth. Note that with the ability to create new λ-expressions (involving another variable x_i), it makes sense to introduce relations between objects, such as *size>* and *equal-shape?*, which respectively check if an

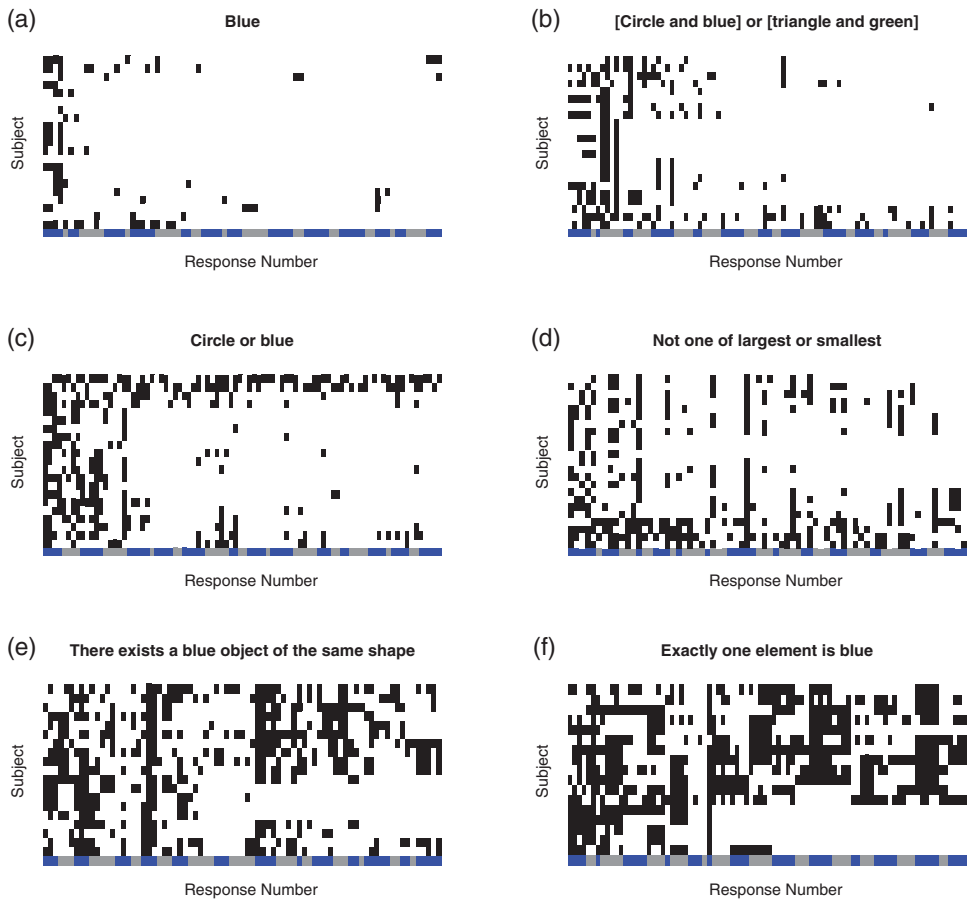

Figure 5. Each row of (a)–(f) shows a single participant's responses throughout the course of the experiment (left to right). Black responses are errors and white responses are correct. The rows have been sorted by similarity in order to illustrate common patterns of generalization. The key at the bottom shows which elements are grouped together in each set. This shows systematic patterns of mistakes during learning, and often all-or-none acquisition by individual participants. See the online article for the color version of this figure.

object is larger than another object or if two objects are the same shape.[2]

In sum, the most important fact about these grammars is that they each allow a large number of concepts to potentially be defined, yet they "build in" very little, including only a few logical and feature-based operations. We conceptualize these languages as essentially like programming languages. Programming languages are useful precisely because they allow a huge range of computational processes to be defined using a highly formalized, small set of primitives, some of which are instantiated directly in hardware. With this view, the task of learning becomes one of program induction: given some observed data, which expression in the LOT best captures the process generating the observed data?

Inference and the Language of Thought

Our modeling is based around two intertwined computational models. One model is a model of *what is going on in subjects' heads*—namely using Bayesian methods as a model of cognitive processes. This is a model of subjects' inferences in learning the concept for *wudsy* and follows largely the concept learning model of (Goodman, Tenenbaum et al., 2008). This model observes the uses of *wudsy* on prior concepts and infers what composition of primitives most likely captures the concept.

The second Bayesian model is a Bayesian *data analysis*, which takes subjects' behavioral responses and infers the parameters of the cognitive model. As experimentalists, this permits us to make efficient inferences about unknown values in each cognitive model. In particular, for each assumed grammar, it takes the behavioral responses and infers the parameters of the grammar, under the assumption that concept learning proceeds according to the assumptions of the cognitive model. For data analysis, we are also interested in comparing different models entirely. Due to the computational difficulty of model comparison in Bayesian inference, we rely on other methods to quantitatively evaluate models (Shiffrin, Lee, Kim, & Wagenmakers, 2008) once parameters have been estimated using Bayesian tools (e.g., sampling and priors).

[2] In the case of SIMPLEBOOLEAN, there is no way to call functions on anything except *x*, meaning that it would be useful to have primitives for these kinds of comparison.

SIMPLEBOOLEAN

START	→	*lambda x . BOOL*
BOOL	→	*(and BOOL BOOL)*
		(or BOOL BOOL)
		(not BOOL)
		true
		false
BOOL	→	*(F OBJECT)*
OBJECT	→	*x*
F	→	COLOR
		SHAPE
		SIZE
COLOR	→	*blue?*
		green?
		yellow?
SHAPE	→	*circle?*
		rectangle?
		triangle?
SIZE	→	*size1?*
		size2?
		size3?

(a)

NAND

START	→	*lambda x . BOOL*
BOOL	→	*(nand BOOL BOOL)*
		true
		false
BOOL	→	*(F OBJECT)*
OBJECT	→	*x*
F	→	COLOR
		SHAPE
		SIZE
COLOR	→	*blue?*
		green?
		yellow?
SHAPE	→	*circle?*
		rectangle?
		triangle?
SIZE	→	*size1?*
		size2?
		size3?

(b)

Figure 6. Two bases for Boolean logic: (a) writes expressions using the standard logical connectives (and, or, not), while (b) uses only one connective (not-and). Both are universal, in that all propositional formulas can be written using either set of primitives.

These two models are nested and inherently connected, yet reflect very different aspects of our scientific analysis. It could be true or false, for instance, that people use the Bayesian concept learning model, independent of whether or not the Bayesian data analysis model is appropriate and effective here. This double use of Bayesian models has been used and discussed previously in cognitive literature (Kruschke, 2010b; Huszár, Noppeney, & Lengyel, 2010; Lee & Sarnecka, 2010, 2011; Lee, 2011; Hemmer, Tauber, & Steyvers, 2014). We leave the mathematical details of these models to the appendixes, and here focus on an intuitive description of both of these inferences.

A Cognitive Model of Concept Inference

The learning model takes previously seen sets labeled with true/false values and infers a hypothesis generating the labels from the objects in each set. This hypothesis represents a learner's guess about the meaning of the word *wudsy*. We create a model of this inference on Marr's computational level (Marr, 1982)—that is, by formalizing the problem that learners solve rather than the specific algorithm they use to do so. For reasons of computational and experimental tractability, we model only subject averages, with the assumption that the distribution of subject responses will track our idealized learning model. Note that this approach leaves unspecified what goes on in the mind of an individual learner. In particular, we do distinguish whether individuals represent one or many rules at a single time—and for good reason: as discussed above, the experimental data needed to determine individual's hypotheses would likely have been much more difficult to collect and would have necessitated a smaller set of target concepts. The learning model assumes only that what they do can be modeled in aggregate as an approximation to the true posterior distribution.

To do this, we use is a form for the learning model that is a simple extension of Goodman, Tenenbaum et al. (2008). This model has two components: a prior $P(h \mid G, D_{**})$ specifying learner's estimate of how likely any hypothesis h is before any labeled objects have been observed. The prior is constructed using a version of a *probabilistic context-free grammar* (PCFG) G (see, e.g., Manning & Schütze, 1999), which specifies a distribution on parse trees, or here a distribution on compositions of the primitive functions. The nonterminals of this grammar are the *return types* of the primitive functions and the rules of the grammar state how those primitives can be combined. For instance, G might contain a

(a)

SIMPLE-FOL (SIMPLEBOOLEAN *rules not shown*)		
SET	→	*S*
	→	*(non-Xes S)*
BOOL	→	*(forall F SET)*
		(exists F SET)

(b)

FOL (SIMPLEBOOLEAN *rules not shown*)		
F	→	*(lambda x_i . BOOL)*
SET	→	*S*
	→	*(non-Xes S)*
BOOL	→	*(forall F SET)*
		(exists F SET)
		(size>= OBJECT OBJECT)
		(size> OBJECT OBJECT)
		(equal-size? OBJECT OBJECT)
		(equal-color? OBJECT OBJECT)
		(equal-shape? OBJECT OBJECT)

Figure 7. Two grammars for generating expressions with quantification. Both build on FULLBOOLEAN by adding primitives: (a) adds quantifiers, and (b) adds quantifiers and λ-abstraction, allowing for quantification over arbitrary predicates.

rule like *BOOL* → *and* (*BOOL, BOOL*), stating that a Boolean (BOOL) can be expanded to a conjunction of Booleans, which themselves are then expanded according to the rules of the grammar. This rule has an associated probability, which can be viewed as the prior probability of conjunction. These probabilities are formalized with parameters D_{**} (see Appendix A) which quantify the distribution of expansions for each nonterminal.

The second component specifies the *likelihood* $P(l_i \mid h, s_i, \alpha, \gamma, \beta)$, quantifying the probability that the set s_i was labeled l_i, if h were the true concept. There are three parameters of the likelihood: α quantifies the amount of noise, or probability that the labeling was done at random rather than according to the rule. γ represents a baseline preference for true-versus-false responses. β represents memory decay parameter specifying how much more important recent observations are for the model's inferences. We use actually use *two* values for β, so that $\beta = (\beta_+, \beta_-)$, respectively giving the memory for positive and negative instances of the concept, in case these differ. Note that a "pure" Bayesian model might not include factors like memory decay. By including these, we are giving the model the possibility of using them to capture human response patterns in case people do not engage in Bayesian inference with perfect memory.

Together, this prior and likelihood can be used to construct an inferential statistical model of rules (h) given the observed sets and labels ($s_1, s_1, \ldots, s_n$, and $l_1, l_2, \ldots, l_n$):

$$P(h \mid \vec{s}_n, \vec{l}_n, G, D_{**}, \alpha, \gamma, \beta) \propto P(h \mid G, D_{**}) \cdot \prod_{i=1}^{n} P(l_i \mid h, s_i, \alpha, \gamma, \beta). \tag{6}$$

where $\vec{s_n} = (s_1, s_2, \ldots, s_n)$ and $\vec{l_n} = (l_1, l_2, \ldots, l_n)$.

Using this model, Figure 8 shows the posterior probability of several hypotheses as the amount of training data increases for four target concepts. These plots were produced by running 250,000 steps of the Metropolis Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970; MacKay, 2003) on each number of labeled points, ranging through the 25 sets run in the experiment. At each point sequentially through the experiment, the top 250 hypotheses were stored, forming a large finite hypothesis space that was used for all further analysis. This means that any hypothesis that was found to be high probability (at any point in the experiment) was reevaluated on the entire set of labeled objects, producing the learning curves shown in Figure 8. These plots illustrate several important aspects of the learning dynamics. First, they show that in many cases the learning model can arrive at the correct concept. This is true even when the target concept is quite complex: for instance, in *the unique largest* (Figure 8b) the model correctly constructs a λ-expression that quantifies over all elements other than x and asserts that all other objects x_2 are strictly smaller than x. In this sense, the learning model "really works" and is capable of narrowing down a vast space of hypotheses using only a few labeled examples—in this case, around 30 labeled sets.

Second, these plots demonstrate the model's simplicity bias: the expressions that are learned early are often simplified approximations of the correct target concept. For instance, for *circle and not blue* (Figure 8a) the model initially learns *circle*; for *there exists a smaller blue object* the model first learns to pick out objects of size 3, the maximum size, then picks out objects that have a smaller object in the set, and finally it converges on the correct answer. Such learning patterns demonstrate that "errors" participants make in the experiment may be rational: the ideal learner does not immediately jump to *there exists a smaller blue object* when shown only two examples. Instead, simpler and thus more likely a priori hypotheses must be eliminated first. This behavior is a natural consequence of learners who have a simplicity bias.

These plots also illustrate the fact that for any given set of labeled training data, there are relatively few hypotheses relevant at any given time. Nearly all the probability mass in the model is split between at most the top 10 hypotheses—all the rest of the infinite space have very low probability. This means that a learner who only actively considered a handful of relevant hypotheses could well approximate the full ideal learner operating over the infinite space. This is a lucky fact for a theory of concept learning

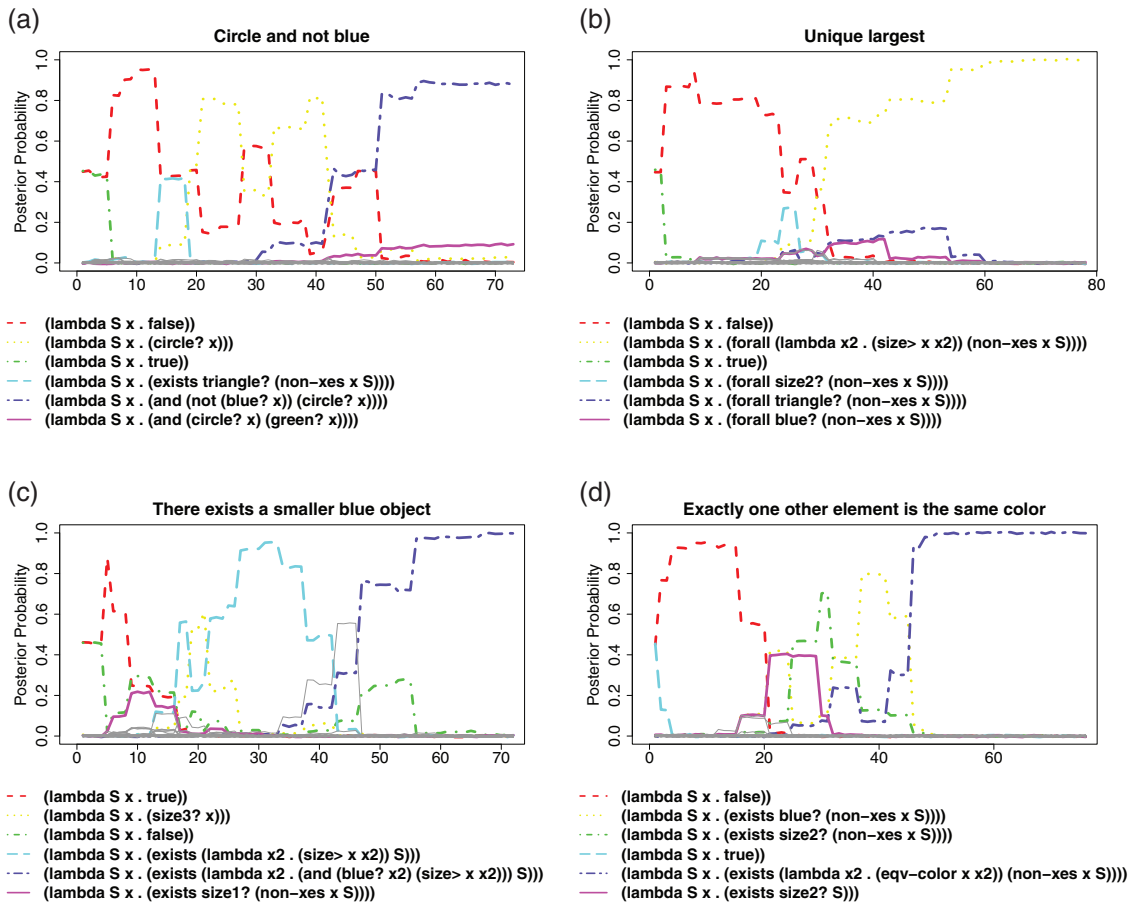

Figure 8. Learning curves with expressions from FOL-OTHER, with $\alpha = 0.75$, $\gamma = 0.5$, $\beta = -0.1$. The top six hypotheses are shown in color and all other hypotheses are in gray. This figure collapses (sums) across logically equivalent hypotheses. See the online article for the color version of this figure.

in such an unrestricted space because it means that our model implementations can do the same. Even though the model is stated as operating over an infinite space, we make an approximation to the full posterior distribution using only the top hypotheses at each point in time.

Figure 8d shows an interesting concept that is not easily expressed in the representation language FOL. Expressing *exactly one other . . .* requires two quantifiers in FOL and this intuitively should take a considerable amount of data to justify. Indeed, with this amount of data the model does not learn the correct concept, but comes to *there exists another object of the same color*. This shows that the representation language chosen may substantially influence what concepts are easily learnable.

A Parameter-Fitting Model

The learning model described in the previous section specifies a probability of any expression, given some set of labeled data. This is intended as our psychological theory of how human learners react to evidence, given the assumed structure of the model, choice of grammar, and choice of parameter values. However, we are really interested in the right representational system—what grammar G and grammar parameters D_{**} are most likely, given people's learning curves. We structure this problem as a Bayesian data analysis problem (Kruschke, 2010a; Gelman, Carlin, Stern, & Rubin, 2014), allowing us to infer likely parameter ranges for the grammars.

The basic setup of this analysis is to consider taking the human data we measured in our experiment and inferring the parameters of the model we do not know. For each item in each set (for a given concept and list in the experiment) we observe a number of counts of how often participants respond *true* and *false*. Let $r_n(x)$ be the number of participants who labeled set s_n with the set of labels x, and R the set of all human responses. In analyzing the data, we are interested in scoring the probability of any particular set of parameters given the participant responses:

$$P(D_{**}, \alpha, \gamma, \beta \,|\, R, \vec{s}_n, \vec{l}_n, G). \tag{7}$$

Here, the parameters of the learning model (G, D_{**}, α, γ, β) are being inferred from the measured behavioral responses (R) for a given grammar (G) and set of data ($\vec{s}_n$,$\vec{l}_n$,). Due to the difficulty of grammar induction, G does not appear on the left side of this equation—the grammar itself is not inferred from data. Instead, we used held-out data techniques described in the next section to compare fixed G, leaving the fitting to find only the parameters.

This data analysis model lets us determine what parameters for the learning model are most likely, given people's observed responses. This model is also Bayesian: intuitively, any setting of parameters will determine a learning curve. Bayes rule allows us to do inference from the empirically observed learning curves to determine statistically likely values of the parameters from the observed learning curves. The details of this model are in Appendix B. For this data analysis model, we also include a temperature parameter tuning the overall uncertainty in the model.

We note that this setup assumes that the distribution of subject responses is the "right" dependent measure, and that it should be modeled as a posterior predictive distribution. Thus, subject's concept at each point in the experiment can be viewed as a *sample* (Vul & Pashler, 2008; Goodman, Tenenbaum et al., 2008; Denison, Bonawitz, Gopnik, & Griffiths, 2013; Bonawitz, Denison, Griffiths, & Gopnik, 2014) of posterior distribution on rules. This framework allows us to abstract away from the specific algorithm individuals use. This reflects an important linking assumption, and one that is testable: if models with this assumption do not provide good fits, other linking functions should be explored.

Model Method

Ultimately, we are interested in determining which grammar (representational system) is most likely, given humans' responses. Full Bayesian model comparison would compute $P(G|R)$, the probability of any grammar given all responses, marginalizing over all the unknown parameters. Unfortunately this, like most such problems, is intractable. We therefore evaluate different grammars using held-out data on the maximum a posteriori (MAP) fitting parameters, an approach common in machine learning. We only train the model (fit the parameters) on one of the two lists for each concept. The held-out scores represent the ability of the model to predict human learning curves on entire sequences of data that it has received no training on. This indirectly penalizes models with too many degrees of freedom since overfitting will result in poor performance on new data. An overview of held-out evaluation and related methods can be found in Shiffrin et al. (2008).

This still leaves the issue of how to fit the model parameters. In the data analysis algorithm, this is a *doubly intractable* problem, with an infinite search over hypothesized expressions in the grammar for each of an infinite number of choices of the parameter values. We approximate a solution to this problem by first constructing a finite space of hypotheses to approximate the infinite one, and then using this finite space in our data analysis algorithm. To make the finite space, we run 100,000 Markov Chain Monte Carlo (MCMC) steps on each concept, list, and amount of data, using a version of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970; MacKay, 2003). These MCMC runs search over expressions using typical values of the likelihood parameters and $D_{**} = 1$, and produce a finite sample of hypotheses. Any hypothesis that occurs in the top 100 hypotheses for any amount of data on a particular concept and list is stored and added to the finite hypothesis space for the model. Thus, the finite space includes a large number of hypotheses that are high-probability at some point throughout the experiment. This is justified because the learning results above (see Figure 8) shows most hypotheses are very low probability at each amount of data, so the top 100 form a reasonable approximation to the infinite space.

Given this finite space of hypotheses we then do MCMC to approximately fit the parameters α, β, γ, and D_{**}. To do this, we run 6,000 iterations, each alternating between 10 MCMC steps over the likelihood parameters, and 10 MCMC steps over the prior parameters (In trial runs, most of the "burn in" time was used to increase a prior temperature parameter, which we initialized to a higher value of 3.5 for the full run). This, along with hand-tuned proposal distributions, means that the model mixes within several hundred of the outer-loops, or several thousand total MCMC steps. Because we do inference over the prior but our finite space was constructed with a particular prior, we also update the finite hypotheses by, every five steps, sampling 10,000 times from the prior and adding the top 25 hypotheses to the finite space. This keeps the finite approximation "current" to the inferred prior parameters and yields replicable results across multiple inference runs.

Modeling Summary

We have defined two statistical models. The first captures the behavior of an ideal learner over the space of λ-expressions, showing how for any particular choice of its prior and parameter values, one can compute model's expected learning curves. To do this, we formalized a prior probability on λ-expressions, and a likelihood measuring how well each λ-expression explains previously observed labels on sets. Bayes' theorem shows how to optimally combine the prior and likelihood, yielding an idealized cognitive model of λ-expression inference. We relate the concept learning model to human data through a data analysis model that infers which values of parameters most closely match human learning. Model comparison between LOTs is done by computing a held-out likelihood score, corresponding to performance of a fit model on independent (untrained) data. This gives an ability to assign each potential LOT a number representing how well it predicts human learning after its parameters have been fit.

Boolean Concept Analysis

In this section, we first apply this technique to simple Boolean representation languages, before moving on to languages with quantification. The Boolean concepts studied here are shown by the green lines in Figures 2 to 4. These target concepts involved operations limited to Boolean logical connectives. We first describe the Boolean languages compared by our model.

Languages Under Consideration

All languages we consider here—and in the next section—assume a fixed basis of perceptual features—namely that there are fixed Boolean predicates like *red?* and *circle?* that can take an element of a set and return true if it possesses the relevant feature. The assumed features are those fixed by our experimental design: features for shape, color, and size. The prior probabilities of these operations are fit in D_{**}. Beyond these assumed featural primitives, we are primarily interested in comparing sets of *predicates* (like *and* and *nor*).

First, we include the two grammars discussed earlier, SIMPLEBOOLEAN and NAND. The SIMPLEBOOLEAN grammar was the one used by

Feldman (2000), and in addition corresponds naturally to the way that these logical words are used in natural language. The NAND basis is natural because it corresponds to a minimal set of logical operations. A cognitive scientist who had strong expectations that the set of cognitive primitives was small, simple, and nonredundant might find this the most plausible basis. The NOR grammar, including only the operation *not-or* is similarly minimal and is also included for comparison. There are several natural extensions of SIMPLEBOOLEAN to consider. First, we might add logical operations such implication (*implies* or $\Rightarrow$) or the biconditional (*iff* or $\Leftrightarrow$). These operations are redundant in that they can be written using primitives in SIMPLEBOOLEAN: *implies* is $\lambda x\ y.\ (or\ (not\ x)\ y)$, and *iff* is $\lambda x\ y.\ (or\ (and\ x\ y)\ (and\ (not\ x)\ (not\ y)))$. The "claim" of a representational system including these primitives is that they are so simple for learners, they must be cognitive primitives rather than compositionally derived from other connectives. We include three additional languages, shown in Table 1: IMPLIES adds *implies*, BICONDITIONAL adds *iff*, and FULLBOOLEAN adds both.

All these languages allow for free-form recombination of logical connectives in that there are no restrictions on the compositional structure. However, there are ways of writing Boolean expressions that force everything into a structurally constrained format known as a *normal form*. Figure 9a shows one example: a DNF grammar for *disjunctive normal form*, in which all concepts are written as disjunctions of conjunctions. This might be natural if people paid attention to conjunctions of features, and preferentially stated concepts in terms of these conjunctions; indeed, this LOT was used by Goodman, Tenenbaum et al. (2008). Similarly, we can also consider a CNF grammar that expresses concepts as conjunctions of disjunctions.

Figure 9b shows a grammar for conjunctions of *Horn clauses* (Horn, 1951; McKinsey, 1943), which generate expressions of the form $x_1 \wedge x_2 \wedge \ldots \wedge x_k \rightarrow y$. Horn clauses are often used in artificial intelligence systems due to their desirable computational properties (e.g., Hodges, 1993; Makowsky, 1987; Russell & Norvig, 2009, section 7.5.3). In particular, they support efficient algorithms for inference and satisfiability (Dowling & Gallier, 1984; Russell & Norvig, 2009), and thus provide a plausible basis for Boolean reasoning in any computational system. Indeed, recent cognitive models have assumed the plausibility of Horn clauses in human learning of theories about the world (Katz et al., 2008; Kemp et al. 2008a).

For baseline measures, we include ONLYFEATURES, which corresponds to learners with no logical connectives, but only access to primitive features (i.e., a learner with no generative capacity). An even simpler base, RESPONSEBIASED corresponds to learners who only try to learn the correct response bias, which is equivalent here to a representation language who only expressions are *true* and *false*.

Finally, we evaluate several other types of models, described in Appendix C. First, an EXEMPLAR model measures each set's similarity to all previously observed sets and attempts to generalized previous labels based on this similarity. LOGISTIC performs a simple logistic regression within each concept and list, providing a type of "psychophysicist's baseline": if we can predict learning curves better than a freely fit logistic curve, then that provides good evidence for a real representational theory. Note that LOGISTIC is not a learning model in that it does not learn any mapping from the stimuli to concepts or representations, and it provides no kind of mechanistic or cognitive theory. Instead, it is only a curve fit to the observed subject accuracy. It therefore cannot generalize to new, held-out lists or concepts. We also include a version of the model that incorporates no prior: the UNIFORM model has an improper, flat prior on expressions, corresponding to no prior bias for simplicity. As with the LOT models, the free parameters (e.g., the distance metric for the EXEMPLAR model) are inferred from data.

Posterior Parameters

A plot of the posterior parameter values for α, β, and γ for each grammar can be found in Appendix D. These results generally make clear several intuitive results, and we do not undertake a detailed analysis or comparison here. The fit alpha values tend to be a little less than 50%, meaning that the model acts as though there is substantial noise, about half of the data. The base rate is near 50% meaning that models do not have a strong yes/no bias for noise data. The β values show that positive examples tend to be remembered better than negative. They fall around -1, meaning that the preference for recent data in making inferences is relatively strong, but not overwhelming (a data point 10 back has the effective weight in the likelihood of 1/10 a recent data point). The model temperature parameters are generally over 1, indicating that the models—likely the PCFG priors—are overconfident. This overconfidence is particularly apparent for models that are poor, like the ONLYFEATURES model.

Table 1
Summary of Boolean Languages Compared

Language	Description
SIMPLEBOOLEAN	*and*, *or*, *not*, used in any composition.
IMPLICATION	Same as SIMPLEBOOLEAN, but with logical implication ($\Rightarrow$).
BICONDITIONAL	Same as SIMPLEBOOLEAN, but a biconditional operation ($\Leftrightarrow$).
FULLBOOLEAN	Same as SIMPLEBOOLEAN, but with logical implication ($\Rightarrow$) and biconditional ($\Leftrightarrow$).
HORNCLAUSE	Expressions must be conjunctions of Horn clauses (e.g., *(implies (and (and a b) c) d)*).
DNF	Expressions are in disjunctive normal form (disjunctions of conjunctions).
CNF	Expressions are in conjunctive normal form (conjunctions of disjunctions).
NAND	The only primitive is *NAND* (not-and).
NOR	The only primitive is *NOR* (not-or).
ONLYFEATURES	No logical connectives; the only hypotheses are primitive features (*red?*, *circle?*, etc).
RESPONSEBIASED	Learners only infer a response bias on *true*/*false*.

(a)

DNF		
START	→	*lambda x . DISJ*
DISJ	→	*CONJ*
		(or CONJ DISJ)
CONJ	→	*BOOL*
		(and BOOL CONJ)
BOOL	→	*(F OBJECT)*
		(not (F OBJECT))
OBJECT	→	*x*
F	→	COLOR
		SHAPE
		SIZE
COLOR	→	*blue?*
		green?
		yellow?
SHAPE	→	*circle?*
		rectangle?
		triangle?
SIZE	→	*size1?*
		size2?
		size3?

(b)

HORN CLAUSE		
START	→	*lambda x . HORN-CONJ*
HORN-CONJ	→	*HORN-CLAUSE*
		(and HORN-CLAUSE HORN-CONJ)
HORN-CLAUSE	→	*(implies HORN-CONJ PRIM)*
HORN-CLAUSE	→	*(implies HORN-CONJ false)*
PRIM	→	*(F OBJECT)*
OBJECT	→	*x*
F	→	COLOR
		SHAPE
		SIZE
COLOR	→	*blue?*
		green?
		yellow?
SHAPE	→	*circle?*
		rectangle?
		triangle?
SIZE	→	*size1?*
		size2?
		size3?

Figure 9. Two additional bases for Boolean logic. The DNF grammar expresses concepts as disjunctions of conjunctions; the HORNCLAUSE grammar expresses concepts as conjunctions of Horn clauses.

Language Comparison Results

Table 2 shows the key model comparison of these representation languages' ability to predict human responses. This table shows the held-out likelihood score (H.O.LL) described above. Better model fit on the held-out likelihood corresponds to numbers closer to positive infinity. The "FP" column gives the model's number of free parameters, counting the several parameters in the likelihood and the D_{**} parameters of the grammar. The last two columns of Table 2 give two intuitive measures of the model's performance. $R^2_{response}$ gives the model's overall R^2 value to individual responses, quantifying the amount of variation in proportion of people who select *true* for each single response that can be explained by the model. R^2_{mean} gives the model's ability to predict each concept's average difficulty, across all concepts. These correlation measures provide a more intuitive way of understanding the relative performance of each model and are computed only on held-out data.

In this table grammars are sorted by the main measure, held-out likelihood. The worst performing models and grammars are ones that lack structured representation: the EXEMPLAR model, ONLYFEATURES grammar, and RESPONSE-BIASED grammar. The best of these, the EXEMPLAR model, can explain only around half of the variance in human responses. The failure of these models provides evidence that such unstructured approaches miss fundamental facts about people's patterns of generalization. The next worst model is the UNIFORM model that has no simplicity bias. The language used in this model was the best-performing language, FULLBOOLEAN. Again, this provides strong evidence for a simplicity bias in concept learning, in line with Feldman (2003c), among others.

Above these grammar, we have HORNCLAUSE, which still performs several hundred points worse than the top grammars. This indicates that this common representation for AI and machine learning research does not accurately capture human inductive biases. The next best languages are IMPLIES and SIMPLEBOOLEAN. SIMPLEBOOLEAN allows for free-form combination of *and*, *or*, and *not*; IMPLIES additionally includes logical implication. The fact that SIMPLEBOOLEAN performs worse than languages with more primitives like FULLBOOLEAN and BICONDITIONAL means that people likely have a richer set of logical connectives than just *and*, *or*, and *not*. In particular, the grammars that perform best according to Table 2 are the grammars that add *iff* to SIMPLEBOOLEAN, as well as the normal-form grammars, DNF and CNF. The largest differences between languages with and without *iff* appears to be in concepts that require exclusive-or (*XOR*), such as *red XOR circle*. Even for those concepts, differences in learning curves for languages with

Table 2
Model Comparison Results on All Languages for Boolean Target Concepts Only

Grammar	H.O.LL	FP	$R^2_{response}$	R^2_{mean}
FULLBOOLEAN	−16296.84	27	.88	.60
BICONDITIONAL	−16305.13	26	.88	.64
CNF	−16332.39	26	.89	.69
DNF	−16343.87	26	.89	.66
SIMPLEBOOLEAN	−16426.91	25	.87	.70
IMPLIES	−16441.29	26	.87	.70
HORNCLAUSE	−16481.90	27	.87	.65
NAND	−16815.60	24	.84	.61
NOR	−16859.75	24	.85	.58
UNIFORM	−19121.65	4	.77	.06
EXEMPLAR	−23634.46	5	.55	.15
ONLYFEATURES	−31670.71	19	.54	.14
RESPONSE-BIASED	−37912.52	4	.03	.04

Note. H.O.LL gives the held-out likelihood on data independent from that used to fit the parameters. FP gives the number of free parameters, dominated by the number of rules in the grammar (e.g. one for each primitive). R^2_{response} gives the correlation of proportion correct on raw responses between each model and humans. R^2_{mean} gives the correlation between model and humans on overall concept accuracy.

and without *iff* are not very large, resulting in very close held-out likelihood scores.

The best grammar, FULLBOOLEAN, scores about 8 points better on held-out likelihood than its closest competitor, BICONDITIONAL. Appendix E provides details on the inferred grammar parameters (D_{**}) for FULLBOOLEAN that were found by the data analysis model.

In one sense, we may take the held-out data scores as "final" measures of how well each grammar performs. However, we might also wonder if these differences between the top grammars are statistically significant—after all, we tested only finitely many concepts out of an infinity of possible concepts. Also, our inference algorithms make several approximations, and it would be good to know if these approximations are good enough to maintain sensitivity to 8-point differences in likelihood. To address this, we first computed a Wilcoxon signed-ranks test, a nonparametric paired comparison, on the likelihood that each pair of models assigned to each held-out data point. With Bonferroni correction for multiple comparisons,[3] this reveals no difference between the FULLBOOLEAN language and BICONDITIONAL, CNF, DNF, or SIMPLEBOOLEAN. However, it does show that FULLBOOLEAN performs significantly better than the others ($p < .05$, corrected)—in particular, IMPLIES, HORNCLAUSE, NAND, and NOR. We also ran the inference algorithms multiple times on the set of Boolean concepts (A single "run" takes about 50 processor days of CPU time, making gathering a large sample of runs impractical). This revealed some variation in the order of the top four grammars, but consistency in ranking these four better than the rest, typically with either DNF, BICONDITIONAL, or FULLBOOLEAN ranked first. Though this shows that the resolution of the present data set cannot distinguish between the top four grammars, it does indicate that FULLBOOLEAN, CNF, DNF, and BICONDITIONAL are better than the other languages for capturing people's inductive bias.

We note that the best grammars can explain an impressive amount of variation in the individual participant responses. This is especially compelling because this correlation is computed only on held-out data: with no parameters fit to the held-out data, the learning model described above can explain 88% of the variation in participants' response patterns. This is further demonstrated by Figure 10, which shows FULLBOOLEAN's probability of responding *true* compared to participants in the experiment. This shows substantial noise in the individual object (in a particular set, list, and concept) responses, shown in gray. The binned data for which we have much less measurement error shows a strong and almost perfectly linear relationship between model posterior predictive and human responses. This holds across both training and held-out data suggesting no overfitting in the model. The other well-performing grammars appear similar when plotted in this manner.

Learning Curves

Importantly, the model is capable of capturing many of the qualitative phenomena that learners exhibit. In particular, learners in the experiment tended to make systematic patterns of errors (see Figure 5). Because we have implemented a full learning model, we can see if the model makes similar errors. Figure 11 shows six typical learning curves, which also function as posterior model checks, demonstrating the model's qualitative fit. The *x*-axis here shows the response number for List 2 in the experiment, the held-out data. The *y*-axis shows human participants' proportion correct at this object, and the model's proportion correct. Thus, each y-value represents the accuracy of learners and the model, conditioning on having seen the correct labels for all previous sets. The dotted blue line here represents the base rate of the concept. This figure shows that through the experiment, both learners and the model make systematic patterns of errors, corresponding to dips in accuracy for the black and red lines.

Figures 11a to d show concepts where the agreement of the model and people is quite close, and Figures 11e to f show cases where the agreement is less good. The fact that the model and people tend to agree indicates that what people are doing is largely rational, generalizing in a way that is similar to our model based on previous labeled examples. In this sense, their "errors" are not really mistakes, but only cases where previous data has led them to hypotheses that give answers different from what the target concept says. Even in cases where the model posterior predictive curves differ from human participants, many of the differences tend to be in the magnitude of an error and not presence or absence of an error. For instance, many of the model "dips" in Figure 11f line up with places where people do have an increased rate of errors. It is important to emphasize that these model curves are not fit to this data. There is no parameter of the models, for instance, that makes the learning curves dip around item 20 of Figure 11a. This dip is caused by the model's learned prior (D_{**}) on independent training data, combined with the fact that at this particular item the observed data leads both models and people to make an incorrect generalization.

[3] It is not clear what the best statistical testing procedure is to use here, since we are selecting the main comparison grammar, FULLBOOLEAN by good performance, and it should be more difficult to find statistical differences between the top performing grammars. Bonferroni correction here is likely highly conservative.

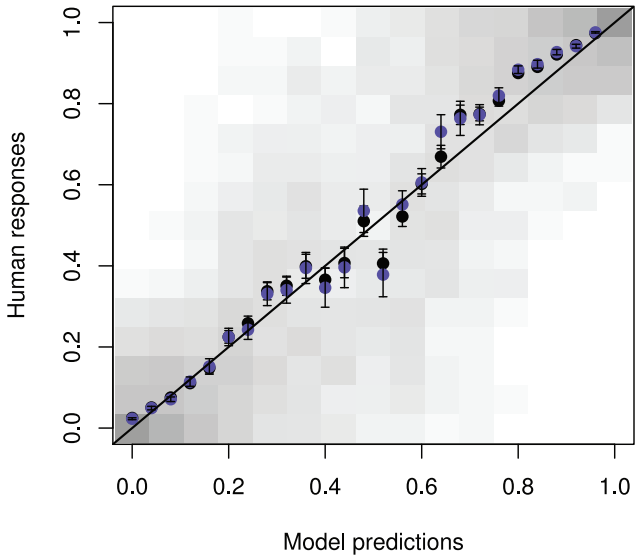

Figure 10. Relationship between model probability of responding *true* (*x*-axis) and participants' probability (*y*-axis). The gray background represents unbinned data, corresponding to raw responses on each object in each set, list, and concept, of the experiment. Black points are binned training data and blue (gray) are binned held-out data. See the online article for the color version of this figure.

Boolean Language Summary

These results demonstrate that models which treat learning as inference in a rich representational system can capture participants' detailed patterns of errors (see Figure 11) as well as their patterns of graded generalizations (see Figure 10). These best rule-like representations outperform other types of baselines, such as simple exemplar models, logistic curves, and response-biased models (see Table 2). Importantly, we are also able to provide evidence against intuitively implausible representations bases, such as the NAND basis, and even formalisms popular in AI like Horn clauses, although the amount of data at present does not distinguish between the best grammars.

We note that this experiment provides evidence about representational components, but does not establish definitively which primitives people are capable of manipulating. We are, after all, able to represent the concept of NAND. This experiment can be viewed as therefore informing our theory of the most natural set of primitives deployed by people's inductive machinery when faced with a feature-based rule learning task like our experiment. It is possible that in other tasks, different sets of data, or with appropriate priming, people would show NAND-based inferences. Our data indicates, however, that such representations would be less cognitively natural.

Languages Involving Quantification

Building off of previous experimental and computational studies (Kemp, 2009; Piantadosi, Tenenbaum, & Goodman, 2010), we extend the modeling results to the wider range of concepts from our experiment that involve quantification and relational terms. To model these concepts, we must consider spaces of representation languages that include these additional operations such as existential and universal quantification, or cardinality operations.

Languages Under Consideration

Like the Boolean analysis, we assumed a fixed set of primitives for accessing the shape, color, and size features of the objects. To study the compositional elements, we ideally might write a set of primitives and construct a grammar either including or excluding each possible primitive. The problem with this approach is that the number of possible grammars would then be exponential in the number of primitives. We might alternatively consider writing a large grammar including all of the primitives and positing that low-probability primitives are likely not components of the LOT. The problem with this is demonstrated by FULLBOOLEAN and SIMPLEBOOLEAN above, where low-probability operations (*iff* and *implies*) nonetheless improve the model fit. It is difficult to tell from the values of D_{**} which primitives should be considered "in" the grammar. We therefore take a more modest approach and construct several plausible *collections* of related primitives. Our language comparison either includes or excludes each set. Table 3 shows five different families of primitive functions. The primitives in each family can be added or not to the best Boolean language, FULLBOOLEAN, to form a new and more powerful LOT.

First we can consider adding first-order quantifiers, *exists* ($\exists$) and *forall* ($\forall$), as in the FOL grammar. These primitives strictly increase the expressive power of any of the Boolean representation languages, allowing for existential and universal quantification. We allow each type of quantification to operate either over the entire set *S*, or over the elements other than *x* in *S*, and the probabilities of each of these types of quantification are fit in the PCFG.

As described above, only adding *exists* and *forall* yields relatively impoverished quantificational abilities since the predicate mapping over the set must be another primitive. True quantification abilities would allow an arbitrary predicate *F* to range over a set, not just the primitive features. This can be accomplished by allowing *F* to expand to a new λ-expression using the rules in LAMBDA-AND-RELATIONAL. This introduces a rule for defining new *functions F*:

$$F \rightarrow \lambda x_i \,.\, BOOL. \tag{8}$$

This rule says that a nonterminal of type *F* can be expanded into a λ-expression λx_i followed by an expansion of *BOOL* (for $i = 1, 2, 3, \ldots$). For instance, *F* could expand to λx_2 . (*or (red? x_2) (blue? x_2)*). Quantifiers such as *exists* then can take this function and a set:

$$\lambda x\, S \;.\; (exists\ (\lambda x_2 \;.\; (or\ (red?\ x_2)\ (blue?\ x_2)))\ S). \tag{9}$$

Here, *exists* returns true iff the *function* λx_2. (*or (red? x_2) (blue? x_2)*) is true for some element of *S*.

The FOL operators correspond to those in classical logic. It has also been suggested, though, that other types of quantification actually provide a better account of people's inductive learning. For instance, Kemp, Goodman, and Tenenbaum (2008b) introduces two quantifiers, *exists-exactly one* and *exists-one-or-fewer*. Both of these quantifiers are analogous to *exists*, except that *exists-exactly one* is true if there is only one element of the set satisfying the predicate, and *exists-one-or-fewer* is true if there is at most one element of the set satisfying the predicate. These quantifiers can be written using the more standard *exists* and *forall*

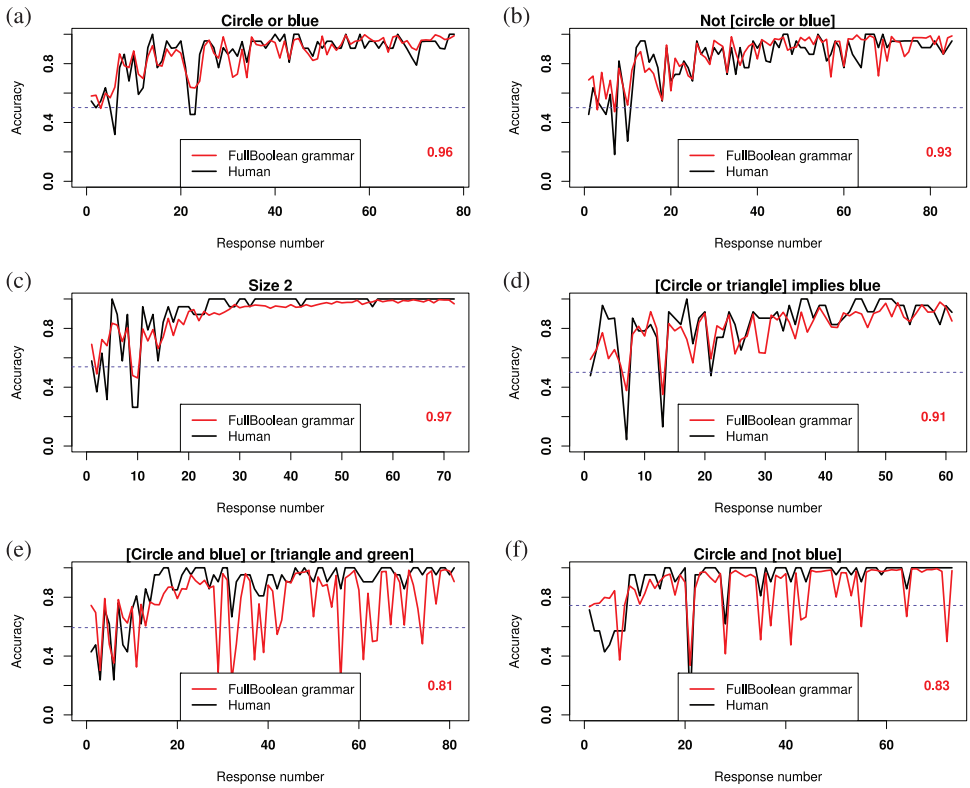

Figure 11. Human (black) versus model learning curves on four example concepts. The numbers in the lower right give R^2s between FULLBOOLEAN's posterior predictive accuracies and humans' observed accuracies. Note the human data for these sequences of data were held-out from training all models (a)–(d) were chosen as examples of "good" fit to the model (e)–(f) represent characteristic poor fits for the model. See the online article for the color version of this figure.

predicates. For instance, *exists-exactly one*, is a function of a function F and a set S, and can be written as,

$$\lambda\, F\, S\ .\ (\mathit{exists}\ \ (\lambda x_1\ .\ (\mathit{and}\ (F\ x_1)\ (\mathit{forall}\ \ (\lambda x_2\ .\ (\mathit{implies}\ (F\ x_2)\ (\mathit{equal?}\ \ x_1\ x_2)))\ \ S)))\ \ S). \tag{10}$$

In other words (*exists-exactly-one F S*) is true if there is one element x_1 in S satisfying F, and for each x_2 in S, if x_2 satisfies F it must be x_1. Importantly, these quantifiers are quite complex to express using *exists* and *forall*, so including them as primitives substantially changes the inductive bias of the model. More generally, we have argued elsewhere (Piantadosi et al., 2012, 2014) that small-set cardinalities (1, 2, and 3) should also be included as representation primitives, in line with very young children's abilities to manipulate small sets (Wynn, 1992). These novel quantifiers are included as ONE-OR-FEWER and SMALL-CARDINALITIES.

Finally, we also compare a simplified version of *second-order* quantification. In standard logic, second-order quantification allows for quantification over predicates, or equivalently subsets of the domain of discourse. For instance, a typical second-order expression is $\exists P \forall x.\mathrm{P}(x)$, which is true if there exists a predicate P such that $P(x)$ for all x. Here, it is difficult to allow for quantification over all predicates, but we can allow quantification over the primitive feature predicates, *red?*, *blue?*, *triangle?*, *size1?*, and so forth This is not formally powerful enough for capturing "real" second-order logic since this type of quantification can be expressed in first-order logic, but it does capture an intuitive sense of quantifying over predicates rather than objects (see also Kemp, 2009). Thus, an expression such as

$$\lambda x\ S\ .\ (\mathit{exists\text{-}color}\ (\lambda P\ .\ (\mathit{forall}\ (\lambda x_2\ .\ (P\ x_2))\ S))) \tag{11}$$

says that there exists a color predicate P (e.g., a predicate in *red?*, *green?*, *blue?*) such that $(P\ x_2)$ is true for all x_2 in S. In other words, all of the elements of S are the same color, no matter what that color happens to be. We note that—unlike second-order logic in general—this could be written using only disjunctions of *forall*, although it would be substantially more complex. Also, it is the case that these second-order predicates require an additional bound variable to be interesting: concepts such as $\lambda x\ S\ .\ (\mathit{exists\text{-}color}\ (\lambda\ P.\ (P\ x)))$ is true for all objects.

To summarize, we have introduced several sets of primitive functions, each of which may or may not be included in a hypothesized LOT. Because the learning model we developed is powerful enough to handle learning in any representation system, we can apply the same methods as the previous Boolean section to see which combination of these primitives best captures people's learning curves.

Table 3
Five Sets of Primitives Which Can Each Be Independently Included or Not to Form a Space of Possible Grammars

	FOL
(exists F SET)	There exists some $x \in S$ such that *(F x)*
(forall F SET)	For all $x \in S$, *(F x)*
	LAMBDA-AND-RELATIONAL
(lambda x_i BOOL)	Lambda abstraction (also introduces a new bound variable x_i)
(equal? x y)	x and y are the same object
(equal-shape? x y)	x and y are the same shape
(equal-color? x y)	x and y are the same color
(equal-size? x y)	x and y are the same size
(size $>$ x y)	x is larger than y
(size $\geq$ x y)	x is large than or equal to y
	ONE-OR-FEWER
(exists-one-or-fewer F SET)	There exists one or zero $x \in S$ such that *(F x)*
	SMALL-CARDINALITIES
(exists-exactly-one F SET)	There exists exactly one $x \in S$ such that *(F x)*
(exists-exactly-two F SET)	There exists exactly two $x \in S$ such that *(F x)*
(exists-exactly-three F SET)	There exists exactly three $x \in S$ such that *(F x)*
	SECOND-ORDER-QUANTIFIERS
(exists-shape P)	There a shape predicate $s \in$ {*circle?, rectangle?, triangle?*} such that *(P s)*
(exists-color P)	There a color predicate $s \in$ {*blue?, green?, yellow?*} such that *(P s)*
(exists-size P)	There a size predicate $s \in$ {*size1?, size2?, size3?*}such that *(P s)*

Note. All grammars include expansions mapping *SET* $\rightarrow$ *S* and *SET* $\rightarrow$ *(non-Xes S),* respectively, *the context set S and the set S*\{*x*}.

Posterior Parameters

Figure D2 shows the posterior parameters in the quantifier analysis. The posterior parameters yield similar values to the Boolean analysis, including temperatures above one, noise and base rate parameters near 0.5, and memory parameters near -1.

Language Comparison Results

Grammars without the LAMBDA-AND-RELATIONAL operations generally performed poorly, so all grammars compared here include these primitives. Thus, by including or excluding each of four sets of primitives, we form a hypothesis space that encompasses a total of $2^4 = 16$ different grammars. We additionally include the top-performing Boolean grammars on this wider space of concepts to test whether people's inductive machinery goes beyond these simple Boolean predicates. Appendix F shows the posterior parameter values found for these grammars. The general ranges and values are similar to the Boolean case above.

Table 4 shows the results of the model comparison on all languages. Beyond the primitives in FULLBOOLEAN, the best grammar here includes only primitives from FOL and ONE-OR-FEWER. This grammar performs substantially better than the Boolean languages, across all measures. Using a Wilcoxon signed-ranks test on held-out likelihoods, the top grammar is significantly better than the second place grammar and all others, conservatively correcting for 16 comparisons ($p < .01$, corrected). This provides strong evidence for quantification in the LOT, in line with Kemp (2009); the superiority of a grammar with multiple types of quantifiers indicates that, like the Boolean results, quantificational operations in the LOT do not make use of a "minimal" basis of operations (such as just FOL). The details of the grammar parameters (D_{**}) found by the data analysis algorithm for FOL are shown in Appendix D.

These results suggest that SMALL-CARDINALITIES are potential primitives since the second-place grammar includes them; note that the concepts studied here do not include many operations on small cardinalities. Most concepts here required checking only for the existence of single elements, which is a cardinality operation captured by FOL. Grammars with these operations might do better if more of the target concepts require them. Additionally, as with the Boolean results, we do not have a precise ability to distinguish between the top hypotheses, as their held-out performance depends (as always) somewhat on the randomness of the held-out data split.

However, these results do provide strong evidence against SECOND-ORDER-QUANTIFIERS: for every other choice of primitives, addition of SECOND-ORDER-QUANTIFIERS reduced the model fit. Indeed addition of *only* SECOND-ORDER-QUANTIFIERS to FULLBOOLEAN resulted in a language that performed worse than any of the Boolean languages. This provides some evidence that people tend not to quantify over properties, consistent with Kemp (2009). This result contrasts with SMALL-CARDINALITIES, which improves over Boolean LOTs, though not as much as the inclusion of other quantifiers.

The ability of the quantificational grammars here to predict human responses on all the concepts is substantially worse than the previous analysis on solely Boolean concepts. The $R^2_{response}$ values show that the grammars explain around 66% of the variance, compared to the capability of the Boolean grammars

Table 4
Model Comparison Results on All Languages With Quantifiers

FOL	One-Or-Fewer	Small-Cardinalities	2nd-Ord.-Quan.	H.O. LL	FP	$R^2_{response}$	R^2_{mean}
✓	✓	·	·	−79023.25	41	.66	.78
✓	✓	✓	·	−79096.47	44	.66	.78
✓	·	·	·	−79329.61	40	.65	.78
✓	✓	·	✓	−79347.52	46	.65	.77
·	✓	·	·	−79463.06	39	.64	.80
✓	·	✓	·	−79518.84	43	.65	.77
·	✓	✓	·	−79863.95	42	.63	.77
✓	·	·	✓	−79908.25	45	.64	.78
✓	✓	✓	✓	−79997.35	49	.64	.74
·	✓	·	✓	−80261.60	44	.63	.77
✓	·	✓	✓	−80366.78	48	.63	.75
·	·	✓	·	−80392.52	41	.63	.72
·	✓	✓	✓	−80435.33	47	.62	.76
·	·	✓	✓	−80604.70	46	.63	.71
		BICONDITIONAL		−81790.49	26	.59	.72
		FULLBOOLEAN		−81844.53	27	.58	.71
		SIMPLEBOOLEAN		−82134.87	25	.58	.73
·	·	·	·	−82342.76	38	.57	.72
		DNF		−82380.87	26	.59	.73
		CNF		−82597.55	26	.58	.73
·	·	·	✓	−82745.12	43	.56	.72

Note. H.O.LL gives the held-out likelihood on data independent from that used to fit the parameters. FP gives the number of free parameters, dominated by the number of rules in the grammar (e.g., one ror each primitive). $R^2_{response}$ gives the correlation of proportion correct on raw responses between each model and humans. R^2_{mean} gives the correlation between model and subjects on overall accuracy aggregated by concept.

to explain around 88% of the variance for Boolean concepts. This could indicate that the representation languages we consider here do not as accurately model people's conception of quantificational concepts, or it could be that people give more variable responses on such complex concepts. Interestingly, the ability of the model to predict each concept's mean difficulty (R^2_{mean}) is actually higher, around 78% of the variance compared with 60–70% on Boolean concepts. This is potentially due to greater and more systematic variance in the concept mean difficulties. As with the Boolean concepts we can plot the model performance versus the participants' actual performance, collapsing across all concepts. Figure 12 shows this relationship and demonstrates the model's ability to predict fine gradations in human response probabilities.

Learning Curves

Again, like the Boolean analysis, the quantificational model is capable of predicting detailed patterns of human learning curves. Figure 13 shows eight different learning curves: Figure 13a–f show well-fit concepts and Figure 13g–h show relatively poorly fit concepts. Some plots show concepts in which the best grammar with quantifiers and other operations is substantially better than FULLBOOLEAN. In Figure 13a, for instance, FULLBOOLEAN is incapable of expressing *exists another object with the same color*, yet people learn this relatively quickly. Interestingly, on this concept both grammars fit equally well for the first few sets, during which people would not have observed enough data to justify using quantifiers and so therefore would respond with simple Boolean expressions. Once enough data has been seen to cause people to learn the concept (around 20–40 sets), the behavior of the quantifier language and the Boolean one diverge substantially. This is exactly the type of concept for which we find strong quantitative evidence in favor of a representation language that is capable of quantification. However, these types of clear and intuitive cases are relatively

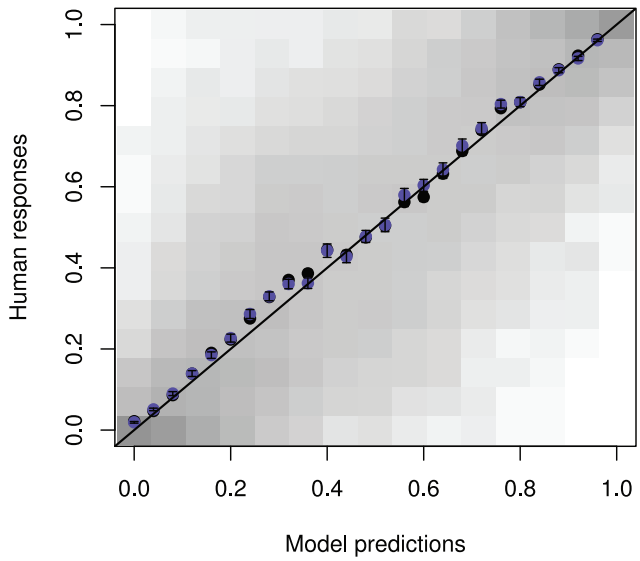

Figure 12. Relationship between model's probability of responding *true* (*x*-axis) and participants' probability (*y*-axis). The gray background represents unbinned data, corresponding to raw responses on each object in each set, list, and concept, of the experiment. Black points are binned training data and blue (gray) are binned held-out data. See the online article for the color version of this figure.

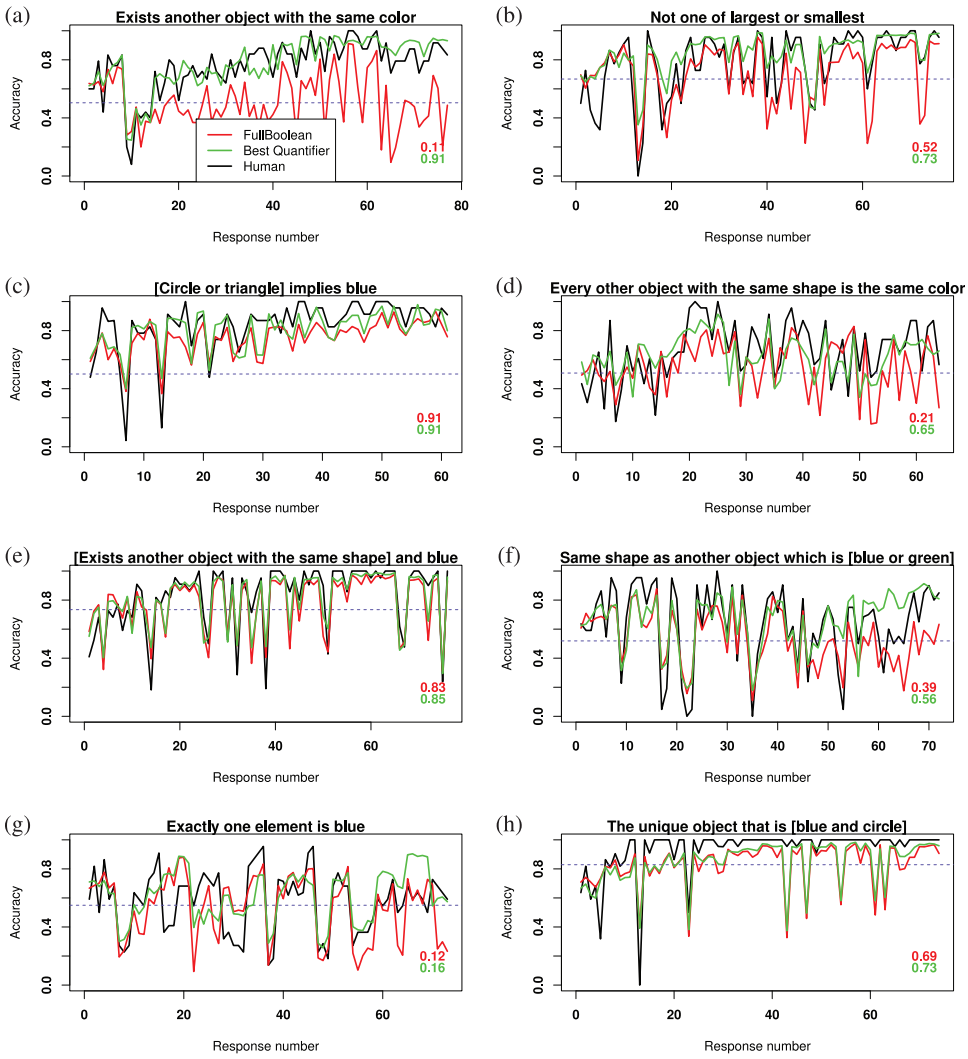

Figure 13. Human (black) versus learning curves according to the best grammar in Figure 4 and FULLBOOLEAN. The numbers in the lower right give R^2s between each language's model-based accuracies and humans' observed accuracies. Note the human data for these sequences of data were held-out from training all models. See the online article for the color version of this figure.

uncommon; most of the target quantifier concepts are difficult for people to learn. Importantly, Figure 13c demonstrates that for the Boolean concepts, both grammars are capable of performing equally well, here with an R^2 of about .91. By adding quantifiers, we do not decrease the model's ability to fit simpler concepts.

Figure 13e to 13f show two concepts that are not simple Boolean predicates, but for which the qualitative fits for the Boolean and quantificational grammars are approximately the same. Indeed, most of the concepts studied here are like this, and do not strongly distinguish between these types of grammars. The reason is likely that in these concepts people do not appear to learn the target concept, as evidenced by the fact that they make systematic patterns of mistakes even near the end of the experiment. The fact that FULLBOOLEAN can capture these mistake patterns indicates that people learn Boolean concepts that are similar to the target concepts, yet not the target concept. For instance, in Figure 13e ([exists an object with the same shape] and blue) they might learn the concept *blue* since it will often be the case that there is an object of the same shape, and so *blue* provides a good approximation to the target. In Figure 13f (*same shape as another object which is* [blue or green]), people may eventually learn *same shape as another object*, which can only be expressed as quantifiers. Using people's response patterns to infer what concept they may have learned is an important future direction for this work.

The curves shown in Figure 13g–h are particularly poor fits for the model. Both grammars seem to mischaracterize learning late in Figure 13g, yielding low correlations. Even though the correlation is high for Figure 13h, both grammars predict patterns of mistakes later that are not observed in human partici-

pants. This latter example suggests potential for improvement in how the model handles uniqueness and small cardinalities.

Quantifier Language Summary

These results generally indicate that people deploy quantificational primitives in rule-based learning tasks like our experiment. These primitives likely include first-order representations with special operations for small cardinalities like exactly one. Resources for second-order quantification appear unnatural. However, as in the case of Boolean languages, these results must be interpreted in the context of our task and experimental items. It is possible that other settings may bring out the kinds of operations as more natural inferences.

Discussion

These results have begun to elaborate the representational systems that support rule-based concept learning. We have shown that it is possible to take human learning curves and infer likely representational systems. Our approach is based on the fact that learners prefer concepts that are representationally simple and that distinct LOTs give different measures of simplicity even if they have the same expressive capability. This allowed comparisons to determine which LOTs best capture human learning curves. We found that systems with rich sets of Boolean connectives and quantifiers best described human learning. Importantly, we also showed that a single, unitary analysis can distinguish across levels of computational power (e.g., predicate logic, first-order logic, second-order logic). Our results rule out intuitively implausible bases like the NAND-basis, and provide quantitative evidence supporting representations systems with quantification over objects.

It may seem obvious that human conceptual systems involve first-order quantification since we are able to think thoughts like "Some dog adored Lindsay." It may seem equally obvious that we can think second-order thoughts, despite the fact that LOTs with this computational ability did not better explain human performance. However, note that our experiments did not test what it is possible to think, but rather that it is natural to learn. More precisely, our results have characterized whether each of these kinds of quantification are a natural part of people's *inductive* machinery. We believe that the presence of these abilities in the context of high-level inductive learning is not at all obvious; it could be the case that our mechanisms of learning operate only over very simple representations like conjunctions of features or continuous spaces. What we have shown here adds to a growing body of work that demonstrates Bayesian induction in rich, cognitive representational systems. This is not to say that there are no situations in which participants might demonstrably learn concepts involving higher-order quantification; in fact, such situations likely exist. Our findings only characterize the inductive mechanisms at play when subjects are asked to learn novel concepts in a sequence of examples from feedback, without the support of rich communicative context or interaction or common sense background knowledge. Pushing this approach to capture the full range of possible human concept induction across learning paradigms is an ambitious and important project.

The results here have examined learning in adults, and therefore is likely to be useful in working out the primitives and processes that support children's learning. However, our results are not about children, and there exist several possibilities for how these findings may relate to processes like language acquisition and cognitive development. In the strongest case, it is possible that the inductive biases guiding adults are very similar to children, suggesting a high degree of continuity in logical, rule-based learned. Alternatively, it is likely that adults have more operations than are available to children, either as a consequence of learning, maturation, or language acquisition. Differences could be in inductive bias (e.g., adults have "cached" some compositions of primitives that children have not, as in Dechter, Malmaud, Adams, & Tenenbaum, 2013) or computational power (adults may have operations available that are qualitatively unlike those for early learners, as in Carey, 2009). Finally, it is possible that young children's inductive systems behave very little like adults in our experiment, as might be the case if the acquisition of natural language is the key cognitive step in achieving the kinds of quantificational inference abilities shown in our experiment. Our data does not speak to these possibilities, although the potential to now extend this simple paradigm to children and other populations is clear.

Our work complements Kemp (2012), who demonstrated a close model fit between grammar-based logical rule learning and human behavior across 11 domains with varying logical structures. Our approach and formalism is deeply related to Kemp's, but different in several key details. Our data set contrasts with Kemp's in that it gathers detailed time courses of concept learning, and our methods realize model comparisons within the context of a Bayesian analysis that infers both distributions on concepts from data, and parameters of the learning model like those in the probabilistic grammar. Kemp's formalism focuses on minimal concept descriptions, with close connections to Minimum Description Length methods (Grünwald, 2007) like those studied previously for Boolean concepts in Feldman (2000). Our results do not exhaustively examine concepts; instead we created a set of concepts that we believed, a priori, were interesting, in part for their concise yet semantically complex descriptions in natural language. Our formalism of λ-calculus is only superficially different than Kemp's, as both primarily function to formalize compositions over primitive operations. Our choice of λ-calculus was motivated by connections to semantics and an eye toward what we believe should be a primary aim for LOT models in the future: inference of arbitrarily rich computational systems. We find it compelling that, despite these differences, we find qualitatively similar results, including a tendency for quantification over objects but not features (Kemp, 2009, 2012).

Our results have also provided quantitative evidence in support of a capacity for rule-like inference. Even the worst rule-based theories have substantially higher correlations with human responses than alternatives like the exemplar model. It is possible that the types of (e.g., gradient) effects typically offered in support of non-rule-based approaches are compatible with the types of rules used in the model—for instance, from "averaging" over rules (Tenenbaum, 2000), a probabilistic rule-based system (Stuhlmüller, Tenenbaum, & Goodman, 2010), or a model that incorporates both rule-like behavior and exemplar behavior (Nosofsky, 1991). Indeed, participants' success with

many of these rule-based concepts illustrates that rule-learning may be a viable developmental theory across domains. At least for our task, rule induction is not extremely difficult for people or models: humans could learn these concepts in a few dozen examples, and the model could using a very simple—indeed very general—stochastic algorithm that may even be developmentally plausible (Ullman et al., 2012). Forming compositions of primitives in order to explain data is not intractably difficult, in theory, practice, or people.

This work has drawn on the capacity to run a new kind of experiment on the Internet—one that involves a large number of participants, conditions, and data points. We expect too that, upon releasing the data with this publication, further techniques and analysis tools will be employed to develop and test rich theories of the statistical learning the experiment captures. Our approach gained much of its power by aggregating results across concepts. Intuitively, a human's probability of using a primitive like *and* should be found by seeing how well grammars performs on all concepts, not simply on conjunctions. If *and* is high probability, other possible concepts like *or* and *not* must be lower probability, and so the predicted learning on these concepts are affected by the probability of *and*. One could imagine simpler "pairwise" comparisons for instance comparing average learning rates on concepts like *red and circle* and *red or circle*. The difficulty with this approach is that it seems difficult or impossible to control other variables like the competing hypotheses and the informativeness of the data with respect to the target concept, so comparisons of average accuracy and learning rate may not be meaningful. The alternative used here is full implementation of a learning model which can support precise quantitative fits. The learning model allowed different representational systems to be "plugged in" without requiring any modification to the inductive mechanisms. We then could recover a single score corresponding to how well the best-fitting parameters of the model generalized to unseen human response patterns. This provides a quantitative standard and allows for an effective comparison of representations that cannot be directly observed in behavior. In principle—due to the computational power of λ-calculus—this approach can be extended to any type of representational system or learning task.

The quantitative results in this paper motivate a challenge to the other paradigmatic approaches to cognitive science—such as connectionism (e.g., Rumelhart & McClelland, 1986; Smolensky & Legendre, 2006) or cognitive architectures (e.g., Anderson, 1993; Newell, 1994)—to provide a model that quantitatively outperforms theories based on near-ideal statistics and explicit representations, or to explain how these types of statistical learning competencies may be implemented. We believe several design features of our experiment make this especially difficult for, for example, connectionist models: the concepts learned are rule-like and relational, the set sizes are variable, participants learn the concepts from relatively little data, and revise hypotheses from a single data point. The important aspect of this challenge is that we have an explicit and principled quantitative measure of performance, likelihood on held-out data. Other fields such as natural language processing find standardized data sets critical for comparing approaches and measuring progress. This present work provides one standardized data set that we hope will prove useful in refining debates about what types of representations and architectures support the richness of human cognition.

Conclusion

The primary goal this work has been to firmly ground the language of thought hypothesis in an empirical domain. We have shown through the use of learning experiments that we can behaviorally distinguish different formal languages for concepts, considering exclusion and exclusion of many sets of primitive operations. LOTs should be considered psychological theories that are subject to empirical evaluation, rather than merely theoretical or philosophical possibilities. Overall, our results demonstrate that intuitively implausible bases cannot explain human learning patterns, and we have provided tentative, early answers to what types of representational systems best capture people's psychological notion of simplicity in the context of inductive rule learning. In particular, rich (nonminimal) Boolean languages including some first-order quantificational concepts appear to best describe human learning. Our results with other baseline models showed that they performed substantially worse than virtually all theories with explicit, compositional representations.

The work has provided evidence for the theory that learners combine a compositional representational system with an approximately ideal statistical inference mechanism. When effective inference is combined with rich representations, we are able to model key phenomena including patterns of errors, graded responses, and eventual learning of complex, compositional concepts. More generally, inductive inference combined with a compositional representational system provides a working theory for how elementary conceptual abilities might be elaborated into complex systems of structured concepts over the course of learning and development.

References

Anderson, J. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00174.x

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 18327–18332. http://dx.doi.org/10.1073/pnas.1306572110

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43,* 800–813. http://dx.doi.org/10.3758/s13428-011-0081-0

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20,* 351–368. http://dx.doi.org/10.1093/pan/mpr057

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences, 18,* 497–500. http://dx.doi.org/10.1016/j.tics.2014.06.006

Boole, G. (1854). *An investigation of the laws of thought: On which are founded the mathematical theories of logic and probabilities*. London, UK: Walton and Maberly.

Bourne, L. (1966). *Human conceptual behavior*. Boston, MA: Allyn & Bacon.

Bruner, J., Goodnow, J., & Austin, G. (1956). *A study of thinking*. New Brunswick, NJ: Transaction Publishers.

Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences, 7,* 19–22.

Chen, M., & Shao, Q. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics, 8,* 69–92.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics, 58,* 345–363.

Conklin, D., & Witten, I. (1994). Complexity-based induction. *Machine Learning, 16,* 203–225.

Dechter, E., Malmaud, J., Adams, R. P., & Tenenbaum, J. B. (2013). Bootstrap learning via modular concept discovery. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence* (pp. 1302–1309). Palo Alto, CA: AAAI Press.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition, 126,* 285–300. http://dx.doi.org/10.1016/j.cognition.2012.10.010

Domingos, P., & Richardson, M. (2007). Markov logic: A unifying framework for statistical relational learning. In L. Geetor & B. Taskar (Eds.), *Introduction to Statistical Relational Learning* (pp. 339–371). Cambridge, MA: MIT Press.

Dowling, W., & Gallier, J. (1984). Linear-time algorithms for testing the satisfiability of propositional Horn formulae. *The Journal of Logic Programming, 1,* 267–284.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407,* 630–633. http://dx.doi.org/10.1038/35036586

Feldman, J. (2003a). A catalog of Boolean concepts. *Journal of Mathematical Psychology, 47,* 75–89.

Feldman, J. (2003b). Simplicity and complexity in human concept learning. *The General Psychologist, 38,* 9–15.

Feldman, J. (2003c). The simplicity principle in human concept learning. *Current Directions in Psychological Science, 12,* 227–232.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. (2008). *LOT 2: The language of thought revisited*. New York, NY: Oxford University Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. In S. Pinker & J. Mehler (eds.), *Connections and symbols: Cognition special issue,* pp. 3–71. Cambridge, MA: MIT Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Boca Raton, FL: CRC Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155–170.

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning & Memory, 9,* 408–418.

Goodman, N. (1955). *Fact, fiction and forecast*. Cambridge, MA: Harvard University Press.

Goodman, N., Mansinghka, V., Roy, D., Bonawitz, K., & Tenenbaum, J. (2008). Church: A language for generative models. In S. McAllester & P. Myllymaki (Eds.), *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence* (pp. 220–229). Corvallis, OR: AUAI Press.

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

Goodman, N., Tenenbaum, J., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–655). Cambridge, MA: MIT Press.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17,* 767–773.

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: Most versus more than half. *Natural Language Semantics, 17,* 63–98. http://dx.doi.org/10.1007/s11050-008-9039-x

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition, 65,* 197–230.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65,* 137–165.

Hampton, J. A. (2006). Concepts as prototypes. *Psychology of Learning and Motivation, 46,* 79–113.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97–109.

Haygood, R. (1963). *Rule and attribute learning as aspects of conceptual behavior* (Unpublished doctoral dissertation). University of Utah, Salt Lake City, UT.

Hemmer, P., Tauber, S., & Steyvers, M. (2014). Moving beyond qualitative evaluations of Bayesian models of cognition. *Psychonomic Bulletin & Review, 22,* 614–628. http://dx.doi.org/10.3758/s13423-014-0725-z

Hindley, J., & Seldin, J. (1986). *Introduction to combinators and λ-calculus*. Cambridge, UK: Press Syndicate of the University of Cambridge.

Hodges, W. (1993). Logical features of Horn clauses. *Handbook of Logic in Artificial Intelligence and Logic Programming, Logical Foundations, 1,* 449–503.

Horn, A. (1951). On sentences which are true of direct unions of algebras. *Journal of Symbolic Logic, 16,* 14–21.

Huszár, F., Noppeney, U., & Lengyel, M. (2010). Mind reading by machine learning: A doubly Bayesian method for inferring mental representations. In *Cognition in flux* (pp. 2810–2815). Austin, TX: Cognitive Science Society.

Johnson, M., Griffiths, T., & Goldwater, S. (April, 2007). Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Proceedings of the North American Conference on Computational Linguistics* (pp. 139–146). Rochester, NY: Association for Computational Linguistics.

Katz, Y., Goodman, N., Kersting, K., Kemp, C., & Tenenbaum, J. (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of 30th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kemp, C. (2009). Quantification and the language of thought. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Cullotta (Eds.), *Advances in neural information processing systems, 22*. Retrieved from http://papers.nips.cc/paper/3816-quantification-and-the-language-of-thought.pdf

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review, 119,* 685–722.

Kemp, C., Goodman, N., & Tenenbaum, J. (2008a). Learning and using relational theories. *Advances in Neural Information Processing Systems, 20,* 753–760.

Kemp, C., Goodman, N., & Tenenbaum, J. (2008b). Theory acquisition and the language of thought. In *Proceedings of 30th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kruschke, J. (2010a). *Doing Bayesian data analysis, 2nd ed.: A tutorial introduction with R, JAGS, and Stan*. San Diego, CA: Academic Press.

Kruschke, J. K. (2010b). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14,* 293–300.

Lee, M. D. (2011). In praise of ecumenical bayes. *Behavioral and Brain Sciences, 34,* 206–207.

Lee, M., & Sarnecka, B. (2010). A model of knower-level behavior in number concept development. *Cognitive Science, 34,* 51–67. http://dx.doi.org/10.1111/j.1551-6709.2009.01063.x

Lee, M., & Sarnecka, B. W. (2011). Number-knower levels in young children: Insights from Bayesian modeling. *Cognition, 120,* 391–402.

Levesque, H., Pirri, F., & Reiter, R. (1998). Foundations for the situation calculus. *Linköping Electronic Articles in Computer and Information Science, 3*(18). Retrieved from http://www.cs.toronto.edu/kr/publications/Levesque98.pdf

Levine, M. (1966). Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology, 71,* 331–338. http://dx.doi.org/10.1037/h0023006

Levinson, S. C. (2013). Recursion in pragmatics. *Language, 89,* 149–162.

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications*. New York, NY: Springer-Verlag.

Liang, P., Jordan, M., & Klein, D. (2010). Learning programs: A hierarchical Bayesian approach. In *Proceedings of the 27th International Conference on Machine Learning* (pp. 639–646). Retrieved from http://machinelearning.wustl.edu/mlpapers/paper_files/icml2010_LiangJK10.pdf

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. New York, NY: Cambridge University Press.

Makowsky, J. (1987). Why Horn formulas matter in computer science: Initial structures and generic examples. *Journal of Computer and System Sciences, 34,* 266–292.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 59). Cambridge, MA: MIT Press.

Margolis, E., & Laurence, S. (1999). *Concepts: Core readings*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman & Company.

McKinsey, J. (1943). The decision problem for some classes of sentences without quantifiers. *Journal of Symbolic Logic, 8,* 61–76.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469–1481.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & W. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). New York, NY: Cambridge University Press.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology, 35,* 113–138.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21,* 1087. http://dx.doi.org/10.1063/1.1699114

Milch, B., Marthi, B., & Russell, S. (2004). Blog: Relational modeling with unknown objects. In *ICML 2004 workshop on statistical relational learning and its connections to other fields* (pp. 67–73). Retrieved from https://www.cs.berkeley.edu/~russell/papers/ijcai05-blog.pdf

Montague, R. (1973). The proper treatment of quantification in ordinary English. In J. Kulas, J. H. Fetzer, & T. L. Rankin (Eds.), *Philosophy, language, and artificial intelligence* (pp. 17–34). New York, NY: Springer.

Muggleton, S., & De Raedt, L. (1994). Inductive logic programming: Theory and methods. *The Journal of Logic Programming, 19,* 629–679.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT press.

Neisser, U., & Weene, P. (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology, 64,* 640–645.

Newell, A. (1994). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Nosofsky, R. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition, 19,* 131–150.

Nosofsky, R., Palmeri, T., & McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological review, 101,* 53–79.

O'Donnell, T., Tenenbaum, J., & Goodman, N. (2009). Fragment grammars: Exploring computation and reuse in language (MIT Computer Science and Artificial Intelligence Laboratory technical report series MIT-CSAIL-TR-2009–013).

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411–419.

Pearl, J. (1998). Graphical models for probabilistic and causal reasoning. In P. Smets (Ed.), *Quantified representation of uncertainty and imprecision* (pp. 367–389). Dordrecht, Netherlands: Springer-Science and Business Media.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31,* 109–130.

Piantadosi, S. T. (2011). *Learning and the language of thought* (Unpublished doctorial dissertation). MIT, Cambridge, MA.

Piantadosi, S. T., Goodman, N., Ellis, B., & Tenenbaum, J. (2008). A Bayesian model of the acquisition of compositional semantics. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Piantadosi, S. T., Goodman, N., & Tenenbaum, J. (2014). *Modeling the acquisition of quantifier semantics: A case study in function word learnability*. Manuscript under review.

Piantadosi, S. T., Tenenbaum, J., & Goodman, N. (2010). Beyond Boolean logic: Exploring representation languages for learning complex concepts. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language, 24,* 554–585.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning, 62,* 107–136.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach*. Essex, UK: Pearson.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1632–1637). Austin, TX: Cognitive Science Society.

Shapiro, S., Pagnucco, M., Lespérance, Y., & Levesque, H. J. (2011). Iterated belief change in the situation calculus. *Artificial Intelligence, 175,* 165–192.

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75*(13), 1–42. http://dx.doi.org/10.1037/h0093825

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248–1284.

Smith, B. J. (2007). boa: An R Package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software, 21*(11), 1–37.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind*. Cambridge, MA: MIT Press.

Smullyan, R. (1985). *To mock a mockingbird: And other logic puzzles including an amazing adventure in combinatory logic*. New York, NY: Oxford University Press.

Stuhlmüller, A., Tenenbaum, J. B., & Goodman, N. (2010). Learning Structured Generative Concepts. In *Proceedings of 32nd Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Tenenbaum, J. (2000). Rules and similarity in concept learning. *Advances in Neural Information Processing Systems, 12,* 59–65.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331,* 1279–1285.

Ullman, T., Goodman, N., & Tenenbaum, J. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development, 27,* 455–480. http://dx.doi.org/10.1016/j.cogdev.2012.07.005

Visser, I., Jansen, B. R., & Speekenbrink, M. (2010). A framework for discrete change. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual pathways of change: Statistical models for analyzing learning and development* (pp. 109–123). Washington, DC: American Psychological Association.

Visser, I., Raijmakers, M. E., & van der Maas, H. L. (2009). Hidden Markov models for individual time series. In J. Valsinger, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 269–289). Dordrecht, Netherlands: Springer-Science and Business Media.

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science, 19,* 645–647.

Wolf, F., & Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics, 31,* 249–287.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology, 24,* 220–251.

Zettlemoyer, L. S., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence* (pp. 658–666).

Appendix A

The Formal Learning Model

We introduce the learning and data analysis models with Figure A1, a graphical model (Pearl, 1998) that describes the relationships between the variables used. The blue nodes in this figure denote variables that are known to learners; they may not be known in the data analysis model. Let s_i and l_i respectively denote the ith set of objects observed and their corresponding labels. For $i < n$, the sets and the labels are both known to learners since they have been provided feedback on previous sets. The main variable of interest, h, is a λ-expression in some representation language, G. The true value of h is the λ-expression that generates the *true/false* labels for each past set and the current set s_n. It is assumed that learners know a grammar G that generates expressions h, as well as variables that parameterize the likelihood (α, β, γ), and the prior (D_{**}), both of which are discussed later. Thus, learners must take their grammar and the previously observed labels to infer a hypothesis h, and apply this to the current set s_n to find l_n.

For convenience, denote the sequence of sets $(s_1, s_2, \ldots, s_i)$ by $\vec{s}_i$ and the corresponding sequence of sets of labels $\vec{l}_i$. We are interested in scoring the probability of h given these previously observed sequences, $\vec{s}_{n-1}$ and $\vec{l}_{n-1}$, and the other known variables. Using Bayes rule, this probability is given by

$$P(h \mid \vec{s}_{n-1}, \vec{l}_{n-1}, G, D_{**}, \alpha, \gamma, \beta) \propto P(h \mid G, D_{**}) P(\vec{l}_{n-1} \mid h, \vec{s}_n, \alpha, \gamma, \beta)$$
$$= P(h \mid G, D_{**}) \prod_{i=1}^{n-1} P(l_i \mid h, s_i, \alpha, \gamma, \beta). \quad (12)$$

Equation (12) makes use of several natural conditional independences shown in Figure A1, for instance, the fact that l_n is independent of G once the hypothesis h is known, and that the l_i are independent once h is known. This equation has two parts, a prior and likelihood, which we now address in turn.

Priors on Expressions

The prior $P(h \mid G, D_{**})$ embodies the core assumptions that learners bring to learning. Here we choose the prior to capture the assumption that learners should prefer representationally simple hypotheses. What is simple may depend on several factors: simplicity might depend on an expression's description length, corresponding to the number of primitives used in the expression h. Or, it may not be the case that all primitives are equally costly, meaning that the prior might depend on which primitives are used, not just how many. Additionally, Goodman et al. (2008a) show that a model that prefers re-use can capture selective attention

(*Appendices continue*)

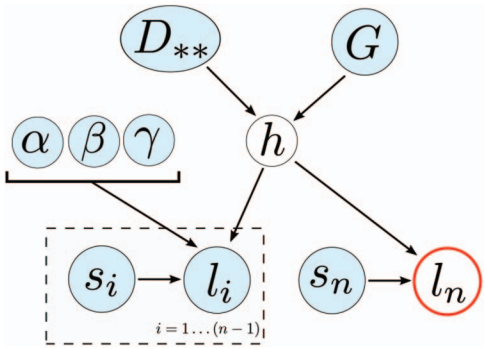

Figure A1. Graphical model representing the variables of the learning model. Here, the expression for the target concept h depends on Dirichlet parameters D_{**} and the grammar G. The specific labels observed for the ith object of the nth set depend on the hypothesis, set, and likelihood parameters, α and γ. In responding, the labels for the nth set of objects are not observed, but the nth set is. See the online article for the color version of this figure.

effects in concept learning, where participants prefer concepts that use the same dimension (e.g., color) multiple times to those that use different dimensions. Thus, *circle or square* is easier than *circle or red* since the former references two shape dimensions, and our prior should potentially incorporate notions of re-use.

One way to capture all of these factors is to first imagine converting one of the above context-free grammars (e.g., Figure 6a) to a *probabilistic context-free grammar* (PCFG). This amounts to assigning a probability that each nonterminal will be expanded according to each of its rules (see Manning & Schütze, 1999). For instance one could make all rule expansions equally likely, meaning that learners would have equal preferences for using any primitive (given a nonterminal type). However, we might also assign the probabilities nonuniformly, corresponding to varying expectations about the probability of any particular expansion or primitive. Any such choice of probabilities induces a distribution on expressions, with the probability of any expression given by product of the probabilities of each of its expansions. Following Goodman et al. (2008a) and Johnson, Griffiths, and Goldwater (2007) we use a variant of PCFGs that potentially encourage reuse of rules: a Dirichlet-Multinomial PCFG (see also O'Donnell, Tenenbaum, & Goodman, 2009). This is best understood as integrating over the rule production probabilities, using a Dirichlet prior on the (multinomial) rule expansions. Suppose that $C_{AB}^{(h)}$ is the count of how many times an expression h uses the rule $A \rightarrow B$, and that C_{A^*} is a vector of counts of how many times the nonterminal A expands to *each* B. Then, for a given grammar G,

$$P(h \mid D_{**}, G) \propto \prod_{\text{nonterminals } A} \frac{\beta(C_{A*}(h) + D_{A*})}{\beta(D_{A*})}, \tag{13}$$

where D_{A^*} is a vector of parameters of the same length as C_{A^*}, and D_{**} is the set of all Dirichlet parameters (for each A). Here, β is the multinomial beta-function, which is given in terms of the Gamma function:

$$\beta(c_1, c_2, \ldots c_n) = \frac{\prod_{i=1}^{n} \Gamma(c_i)}{\Gamma(\sum_{i=1}^{n} c_i)}. \tag{14}$$

This prior uses a single Dirichlet-multinomial for each set of rule expansions for each nonterminal A. This Dirichlet-multinomial is parameterized by a set of real numbers, D_{A^*}. If we renormalize D_{A^*}, we get the expected probability of each rule expansion from A, using the basic properties of the Dirichlet-distribution. Importantly, however, the Dirichlet parameters also characterize reuse: if the Dirichlet parameters are small in magnitude, then observing a rule used once will substantially increase its probability of being re-used in the future. In contrast, if the magnitude of D_{A^*} is large, then adding additional rule counts does not change the probability an expansion will be reused, so the model does not prefer reuse strongly. When $D_{A^*} = 1$ for all A, this prior recovers the rational rules model of Goodman et al. (2008a). By doing inference over the D_{**} we are therefore able to infer both the relatively probabilities of each rule expansion, and how much their probabilities of use in the future are influenced by whether or not they were used already in an expansion.

We also include one more parameter, a *temperature* T, which controls the strength of this prior by raising it to the $1/T$th power. As $T \rightarrow 0$ the prior assigns most probability mass to short expressions, and as $T \rightarrow \infty$ the prior approaches a uniform distribution. For notational simplicity, this is left out of our equations.

The Likelihood of Data Given an Expression

The *likelihood* $P(\vec{l}_n \mid h, \vec{s}_n, \alpha, \gamma, \beta)$ in Equation (12) quantifies how well each expression h explains previously observed labels. Following the set-up of the experiment, we can consider sets of objects and learners who have observed *true* and *false* labels on some collection of previously observed data points. Given h, we assume labels are noisily generated for the current set by choosing the correct label (according to h) for each item with high probability α, and with probability $(1 - \alpha)$ choosing from a baseline distribution on labels, parameterized by γ. Thus, when the labeling is not done according to h, *true* is chosen with probability γ and *false* is chosen with probability $(1 - \gamma)$. This captures the intuition that the labels typically come from h, but occasionally noisy labels are generated from some baseline distribution. This process is therefore parameterized by two variables, α and γ. This process gives that

$$P(l_i = x \mid h, s_i, \alpha, \gamma) = \begin{cases} \alpha + (1-\alpha)\cdot\gamma \text{ if } h \text{ returns } x \text{ for } s_i \text{ and } x = \textit{true} \\ \alpha + (1-\alpha)\cdot(1-\gamma) \text{ if } h \text{ returns } x \text{ for } s_i \text{ and } x = \textit{false} \\ (1-\alpha)\cdot\gamma \text{ if } h \text{ returns } y \neq x \text{ for } s_i \text{ and } y = \textit{true} \\ (1-\alpha)\cdot(1-\gamma) \text{ if } h \text{ returns } y \neq x \text{ for } s_i \text{ and } y = \textit{false}. \end{cases} \tag{15}$$

Equation (15) simply adds up all the ways that x could be generated for s_i. When x is the correct label generated by h, then

(*Appendices continue*)

x could be generated by labeling from h with probability α, or by choosing from the baseline distribution with probability $(1 - \alpha)$. This choice from the baseline depends on whether x is *true* (probability γ) or x is *false* (probability $1 - \gamma$). If x is not the label returned by applying h to s_i, then it has to have been generated from the baseline distribution. This equation embodies the assumption that learners reason about the statistical process that generates their observed data, allowing them to imagine how likely any particular hypothesis h would make the observed data, given the noisy labeling process.

Equation (15) allows us to score the likelihood of the label for any particular labeled set of objects. But in the experiment, participants see a sequence of labeled sets and objects. It is likely that learners have better memory for more recent examples, so we include a *memory-decay* on the likelihood, so that learners prefer more strongly to get more recent examples correct. Motivated by power law decays in memory (Anderson & Schooler, 1991), this takes the form of a power law decay on the *log* likelihood:

$$\log P(\vec{l}_n \mid h, \vec{s}_n \alpha, \gamma, \beta) = \sum_{i=1}^{n} (n - i + 1)^{-\beta} \log P(l_i \mid h, s_n, \alpha, \gamma) \tag{16}$$

Here, we have weighted the likelihood of an individual set from (15) by a power-law term, $(n - i + 1)^{-\beta}$, which makes earlier data less important. This introduces one free parameter, $\beta > 0$, which controls the amount of memory-decay in the model. For $\beta > 0$, learners prefer hypotheses that explain more recent data, but as $\beta \to 0$, this preference is reduced.

Appendix B

Inference for Data Analysis

Let $r_n(x)$ be the number of participants who labeled set s_n with the set of labels x (one true/false response for each item in s_n), and R the set of all human responses. In analyzing the data, we are interested in scoring the probability of any particular set of parameters given the participant responses. By Bayes rule, the probability

$$P(D_{**}, \alpha, \gamma, \beta \mid G, R, \vec{s}_n, \vec{l}_n) \propto P(R \mid G, D_{**}, \alpha, \gamma, \beta, \vec{s}_n, \vec{l}_n) P(D_{**}, \alpha, \gamma, \beta). \tag{17}$$

The first term here is the likelihood of the human responses for any given setting of the parameters. Under the assumption that participants choose labelings according to the output of the learning model, this term is a multinomial likelihood,

$$P(R \mid G, D_{**}, \alpha, \gamma, \beta, \vec{s}_n, \vec{l}_n) = \prod_{i=1}^{n} \prod_{x} \left[P(l_i = x \mid \vec{s}_{i-1}, \vec{l}_{i-1}, G, D_{**}, \alpha, \gamma, \beta) \right]^{r_i(x)} \tag{18}$$

where product over x runs over all possible labelings x of s_i (all values of l_i). Equation (18) says that the probability of all responses according to the learning model is a product over all sets observed, and for each set a product over all possible labelings raised to the number of times that labeling is observed in participants. The key term of (18) is the probability of any labeling given all previous data, $P(l_i = x \mid \vec{s}_{i-1}, \vec{l}_{i-1}, G, D_{**}, \alpha, \gamma, \beta)$, since this is the model's predicted distribution of responses to set i. This term is important because it characterizes the model generalizations: the model works well if it generalizes like people do, labeling new data from typically ambiguous evidence in the same way as our study participants. The target expression h does not appear here because the right h is not known to participants; however, it can be computed using the previously defined prior (13) and likelihood (15):

$$P(l_i = x \mid \vec{s}_{i-1}, \vec{l}_{i-1}, G, D_{**}, \alpha, \gamma, \beta) = \sum_{h \in G} P(l_n = x \mid h, s_i, \alpha, \gamma, \beta) P(h \mid \vec{s}_{n-1}, \vec{l}_{n-1}, G, D_{**}, \alpha, \gamma, \beta). \tag{19}$$

Intuitively, participants' distribution of guesses at l_n is given by the probability of l_n given each hypothesis h, times the probability that h is correct according to all previously labeled data.

The second term in Equation (17) is the prior on parameters. We choose these priors to have a very simple form: D_{**} is chosen according to a *gamma* (1, 2) prior and the priors on α, β, and γ are taken to be uniform. In practice, the amount of data the model is fit to makes these priors largely irrelevant.

Together, these parts of Equation (17) specify a probabilistic data analysis model that allows us to use the learning model to infer a distribution on unknown parameters that characterize the grammar and the likelihood. Until this point, we have presented this model as though there is only one concept and list; this was for notational convenience since otherwise all variables would have to be indexed by a concept (and potentially a list). The main parameters of interest—D_{**}, α, γ, β, and most importantly G—are psychological variables that are true *across* concepts. Most of the power of our analysis comes from finding parameters that work for a wide variety of concepts. So in reality, (18) involves a product over concepts.[4]

[4] And so $\vec{s}_n$ and $\vec{l}_n$ must also be indexed by concept and list correspondingly.

(*Appendices continue*)

Appendix C

Formal Description of Alternative Models

Here we describe several additional models compared in the Results sections:

Uniform

This model assigns all hypotheses a uniform prior:

$$P(h) \propto 1. \tag{20}$$

This prior is also improper and is an interesting baseline that corresponds to no substantive expectations about the form of concepts.

Response-Biased

This model corresponds to a simple response-biased model which uses labeled data to do inference over the proportion of time the hypothesis is true. This can be interpreted as a special representation language where there are only two possible expressions: one that is always true and one that is always false.

Logistic

This model provides another baseline which fits a logistic learning curve within each concept. This model therefore has no interesting representational capacities or predictive abilities, but comparison to it reveals the degree to which LOT models can surpass how a statistician or psychophysicist might model performance in this task.

Exemplar

This is an exemplar model on the set-based stimuli. It is difficult to know exactly how exemplar models might be applied to sets of objects, since such models are generally stated in terms of similarity of object features, not similarities of collections of objects. Here, we begin by defining an object-wise distance metric:

$$\begin{aligned} d(x, y) = {} & \delta_{shape(x)=shape(y)} \cdot W_{shape} \\ & + \delta_{color(x)=color(y)} \cdot W_{color} \\ & + \delta_{size(x)=size(y)} \cdot W_{size}, \end{aligned}$$

where *shape*, *color*, and *size* are functions that map objects to their shapes, colors, and sizes, and W_{shape}, W_{color}, and W_{size} are free parameters. Given two sets, we can then consider all possible ways of aligning their objects. This is necessary because if the next set is similar to a previously observed set, we need to know how objects in the current set correspond to objects in the previous one. Since their orders may change, this can only be accomplished by finding an alignment between the sets. For convenience, let $d^*(s_j, s_k)$ be the total distance according to d of the best alignment of elements of *sets* s_j and s_k. If the sets are different sizes, then some elements may be dropped. Then we define a distance metric on sets by

$$D(s_j, s_k) = abs(|s_j| - |s_k|) \cdot W_{length} + d^*(s_j, s_k). \tag{21}$$

Intuitively, this says that sets are penalized W_{length} for differences in cardinality, and then according to the distance of their elements in the best alignment via $d(x, y)$. D is used to define the probability of generalizing labels from a previously labeled s_j to the next set, s_n, according to the best alignment between the two. This *log* probability is proportional to

$$-\beta^{n-j+1} \cdot \log\ D(s_j, s_n). \tag{22}$$

Like the Bayesian models, this includes a power law memory-decay parameter, β.

(*Appendices continue*)

Appendix D

Posterior Parameter Fits

Figures D1 and D2 show posterior HPD intervals for the grammar parameters of each model, giving the estimate for how likely each primitive operation is to be used in people's conceptual representations. The values can be interpreted as probabilities by re-normalizing the values in each row (e.g. all BOOL to . . . rows).

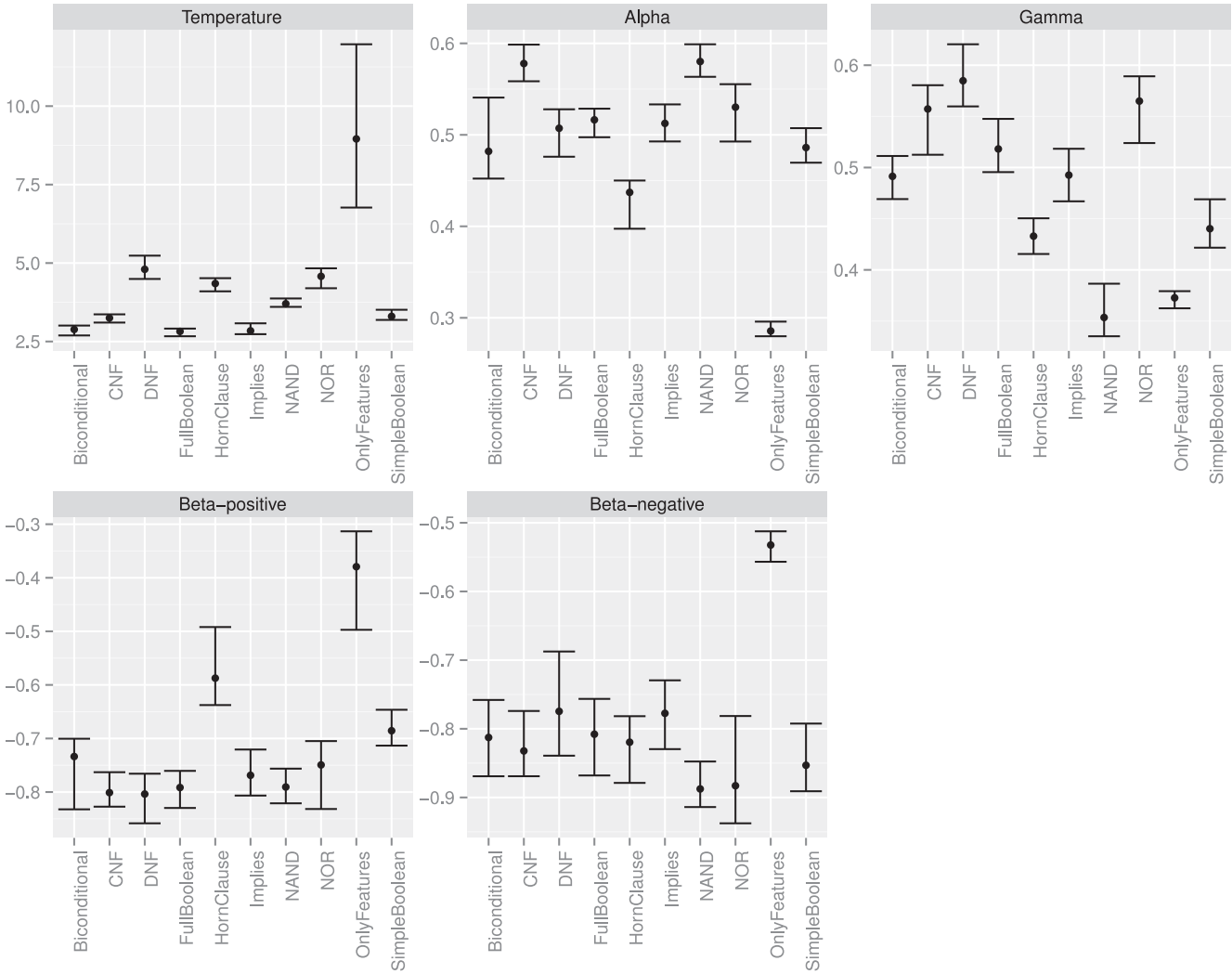

Figure D1. Posterior parameter ranges for each Boolean model. Dots show medians and error bars show posterior 95% quantiles.

(*Appendices continue*)

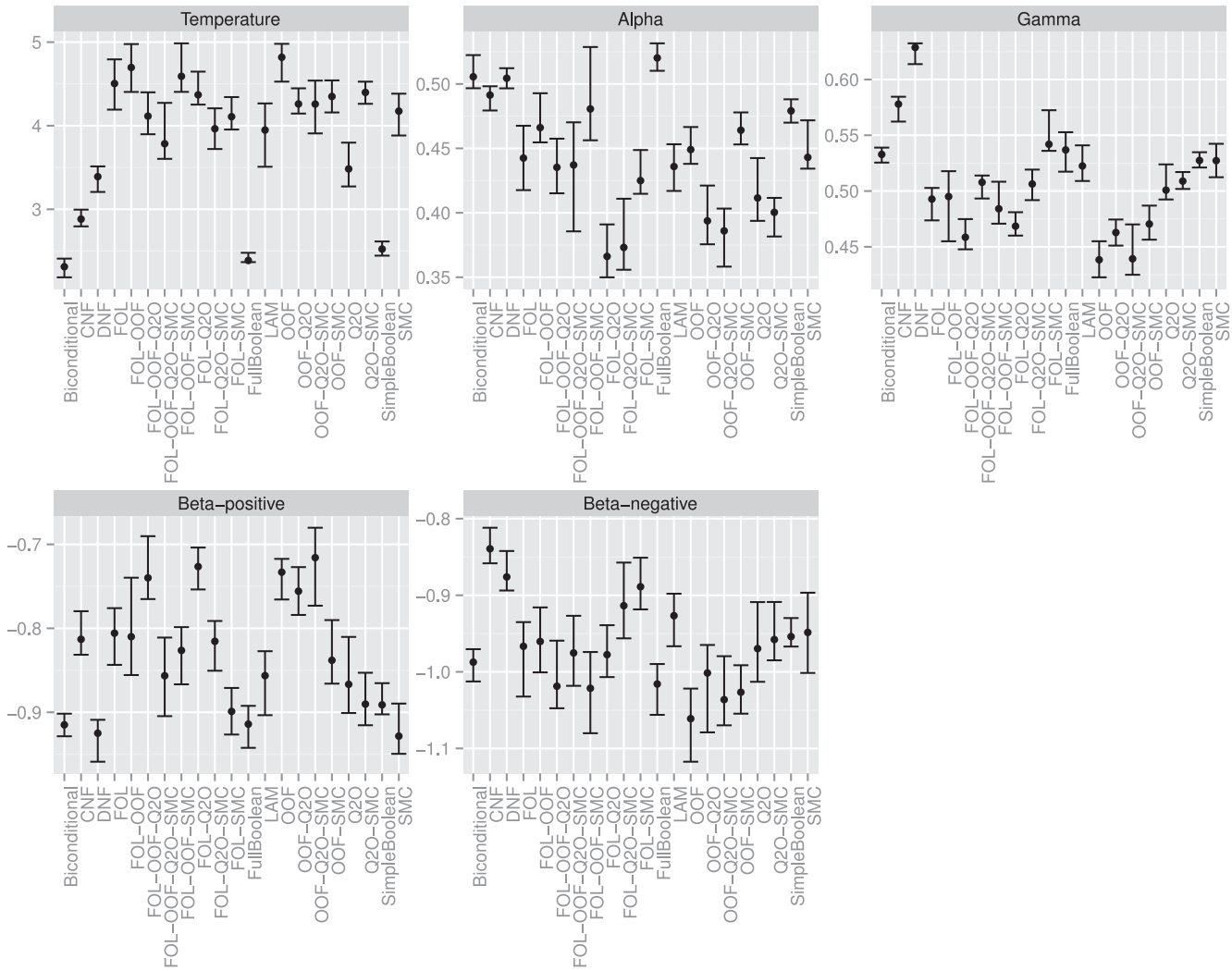

Figure D2. Posterior parameter ranges for each language. The *x*-labels give which primitives are included in each language (FOL = first order logic; OOF = one or fewer; SMC = small cardinalities; 2OQ = second order quantifiers).

(*Appendices continue*)

Appendix E

The Inferred Boolean Grammar

A more detailed picture of the grammars inferred from the experimental results is shown in Figure E1. This shows the D_{**} parameters found by the data analysis model for FULLBOOLEAN. The red points correspond to MAP estimates of the parameters, and the intervals are *highest-posterior density* ranges, using the Chen and Shao (1999) algorithm from the R package *boa* (Smith, 2007). These numbers can, roughly, be interpreted by re-normalizing for each nonterminal type to yield a PCFG. Thus, *COLOR* expands to *yellow?* and *green?* with approximately equal probability. *blue?* is much more salient—it is more likely to be used in a concept. Similarly, for *SHAPE* expansions, participants are roughly twice as likely to expand to *circle?* as the others, indicating a bias in the prior for concepts using circular shapes.

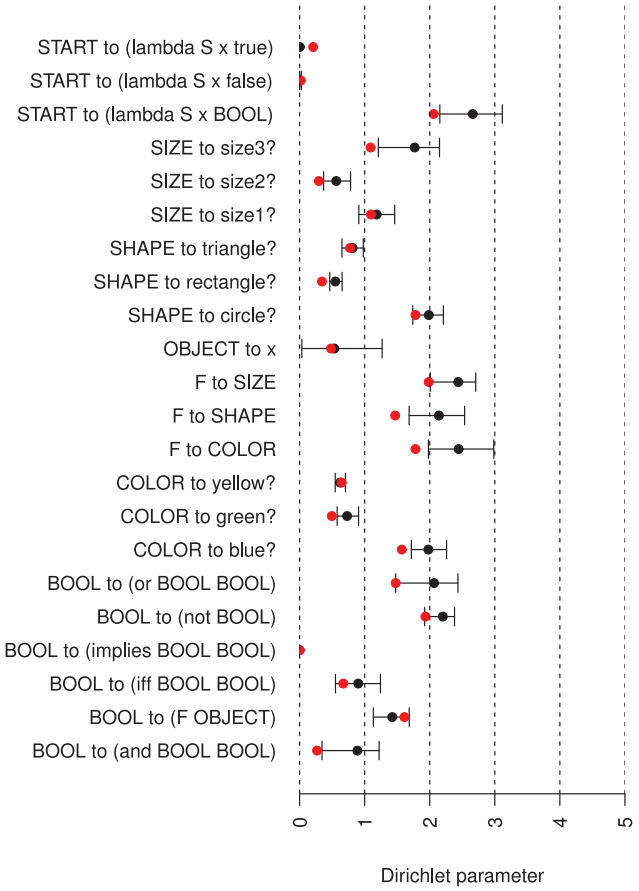

Figure E1. Posterior parameters D_{**} found by the inference algorithm for the FULLBOOLEAN grammar. The red (gray) dots are MAP grammar parameters and the intervals are 95% HPD intervals computed using the Chen and Shao (1999) algorithm. See the online article for the color version of this figure.

The (unnormalized) magnitude of these numbers shows the role of reuse in an expression: roughly, each time a rule is used in creating an expression, its parameter value is increased by 1 for later expansions in the same expression, and the nonterminals are renormalized. For example, *F* is equally likely to expand to a *SIZE*, *SHAPE* or *COLOR* since all of these have equal magnitudes ($\approx$ 2.0). Roughly, once *F* is expanded one way once—say to *SHAPE*—in an expression, *SHAPE* is preferentially re-used with probability $(2.0 + 1)/(2.0 + 1 + 2.0 + 2.0) = 0.43$ next time a *F* expansion is followed. Here, since *SHAPE* was used once, 1 has been added to its initially unnormalized probability of 2.0, and the expansions renormalized. Thus, as the magnitude of these parameters gets large, reuse of rules is less preferred; the general low values of these parameters indicate that reuse is likely preferred by learners, consistent with Goodman et al. (2008a).

Figure E1 reveals several interesting trends. First, the *START* symbol is expanded with very high probability to an expression involving *BOOL* instead of true or false hypotheses, indicating a prior bias against truth-functionally trivial expressions. Plausibly, *SIZE* is preferentially expanded to the most salient sizes, *size3?* and *size1?*. Not surprisingly, this grammar assigns substantial probability to *iff*, the primitive that only the top two grammars, FULLBOOLEAN and BICONDITIONAL include. Lower probability is assigned to *implies*. However, *implies* is not zero probability—otherwise these probabilities for FULLBOOLEAN would essentially yield the grammar BICONDITIONAL. The MAP probability of *implies* is about 0.001, so use of *implies* would yield a prior about 7 log points lower, requiring getting 10 to 15 additional example objects (not sets) correct throughout the experiment for typical values of α. Said another way, this difference in the prior could easily be overcome in the likelihood with just a handful of examples; this is why FULLBOOLEAN can outperform BICONDITIONAL despite the fact that the only primitive FULLBOOLEAN additionally has (*implies*) is low probability.

Examination of grammars for CNF and DNF reveal trends of some preference for re-use, especially of feature primitives. These grammars also tend to set probabilities to generate primarily conjunctive concepts, rather than disjunctive concepts, leading to a stronger prior conjunction bias than FULLBOOLEAN; this may be a hallmark of over-fitting, potentially explaining why CNF and DNF performed better on training data but trended worse on held-out data.

It is important to note that the prior on these parameters was *Gamma* (2, 1), which means that if the data was not informative about the parameter values, we could expect them to be approximately 1 with a variance of 1. Thus, values that are far from 1 represent parameters that have changed from the prior in order to better explain the data. Values near 1 with a different variance (e.g. *AND*) may have a true value close to 1 and the data has increased our confidence relative to the prior in this value.

(*Appendices continue*)

Appendix F

The Inferred Full Grammar

A better understanding of the grammar that is inferred with quantifiers is provided by the probabilistic version for the grammar that is parameterized by D_{**}. As with the Boolean grammar, the relative size of each of these parameters characterizes the probability of using each primitive, and the magnitude of these values inversely relates to the degree that reuse is preferred. Values far from the *Gamma* (2, 1) prior (with mean 1 and variance 1) should be interpreted as values that the experimental data is highly informative about.

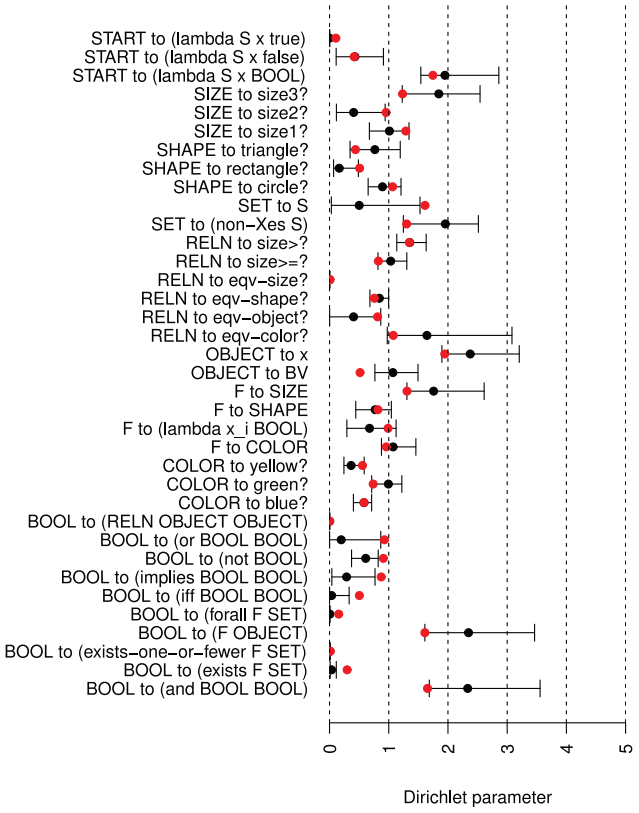

Figure F1. Posterior parameters D_{**} found by the inference algorithm for the best grammar in Figure 4, including only FOL operations. The red (gray) dots are MAP grammar parameters and the intervals are 95% HPD intervals computed using the Chen and Shao (1999) algorithm. See the online article for the color version of this figure.

Figure F1 shows the D_{**} parameters for the full data set and illustrates similar patterns to Figure E1: for instance, *size2?* is a low-probability operation, *F* is about equally likely to expand to *SIZE*, *SHAPE* and *COLOR*. Unlike the Boolean results, this shows that *and* is higher probability than *or*, agreeing with concept learning asymmetries between conjunctive and disjunctive concepts.

Most interestingly, however, are the probabilities assigned to the more complex primitives. For instance, the quantifiers *exists* and *forall* are given relatively low probability, and the low magnitude of all expansions of *BOOL* in general indicate a stronger preference for reuse. The relational terms in LAMBDA-AND-RELATIONAL vary substantially in probability, indicating preferences to compare equality of shapes, colors, and objects, not sizes. As with *implies* in the FULLBOOLEAN above, the primitive *exists-one-or-fewer* is given a MAP probability of 0.013, which is low in the prior, but easily overcome with a few examples, allowing it to improve model fit. Additionally, *SET*s are more likely to be expanded to (*non-Xes S*) than *S*, indicating that quantification tends to occur over all *other* objects in the set; this makes it more natural to express concepts such as *exists another object with the same color* and less natural to express *everything iff there is a triangle in the set*, concepts which people find easy and hard respectively.

Received November 24, 2014
Revision received September 18, 2015
Accepted September 20, 2015 ■