Cognitive Science 32 (2008) 68–107

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210701801974

Using Category Structures to Test Iterated Learning as a Method for Identifying Inductive Biases

Thomas L. Griffiths[a], Brian R. Christian[b], Michael L. Kalish[c]

[a]*Department of Psychology, University of California, Berkeley*
[b]*Department of English, University of Washington*
[c]*Institute of Cognitive Science, University of Louisiana at Lafayette*

Abstract

Many of the problems studied in cognitive science are inductive problems, requiring people to evaluate hypotheses in the light of data. The key to solving these problems successfully is having the right inductive biases—assumptions about the world that make it possible to choose between hypotheses that are equally consistent with the observed data. This article explores a novel experimental method for identifying the biases that guide human inductive inferences. The idea behind this method is simple: This article uses the responses produced by a participant on one trial to generate the stimuli that either they or another participant will see on the next. A formal analysis of this "iterated learning" procedure, based on the assumption that the learners are Bayesian agents, predicts that it should reveal the inductive biases of these learners, as expressed in a prior probability distribution over hypotheses. This article presents a series of experiments using stimuli based on a well-studied set of category structures, demonstrating that iterated learning can be used to reveal the inductive biases of human learners.

Keywords: Bayesian inference; Induction; Iterated learning; Mathematical modeling; Statistics

1. Introduction

Many of the problems that are explored in cognitive science including learning languages, categories, and causal relations are problems of induction, requiring people to evaluate underdetermined hypotheses in the light of limited data. Arguments from both philosophy (e.g., Goodman, 1955) and formal learning theory (e.g., Kearns & Vazirani, 1994; Vapnik, 1995) stress the importance of inductive biases—constraints on the hypotheses under consideration—in solving these problems. Inductive biases make it possible for a learner to choose between hypotheses that are equally consistent with the observed data. As a consequence, having the right inductive biases is the key to successfully solving inductive problems, as it is the only way

Correspondence should be addressed to Thomas L. Griffiths, Department of Psychology, University of California, Berkeley, 3210 Tolman Hall #1650, Berkeley, CA 94720-1650. E-mail: tom_griffiths@berkeley.edu

that learners can become more likely to select hypotheses that are true. Understanding how human learners solve these inductive problems is thus a matter of identifying their inductive biases.

Given the prevalence of inductive problems in cognitive science, it is not surprising that many of the theoretical debates about human learning can be expressed as debates about inductive biases. This is particularly apparent in the literature on language acquisition, where the opposing claims that constraints on the languages people can learn are strong, innate, and specific to language (e.g., Chomsky, 1965) or weak, and the result of general-purpose learning mechanisms (e.g., Elman et al., 1996) can both be interpreted as statements about human inductive biases. Likewise, the question of whether human category learning is better described by prototype models (Reed, 1972), exemplar models (Medin & Schaffer, 1978; Nosofsky, 1986), or something in between (Love, Medin, & Gureckis, 2004; Vanpaemel, Storms, & Ons, 2005) can be expressed as a question about constraints on the kind of category structures that people are willing to infer from data. Claims about preferences for simpler causal structures (Chater & Vitányi, 2003; Lombrozo, 2007) can also be viewed as a form of inductive bias relevant to causal learning. We can begin to shed light on these issues by developing methods for identifying human inductive biases.

Computational models have been a valuable tool for investigating human inductive biases, with traditional applications of these models taking two approaches. The first approach is to demonstrate that a particular set of inductive biases is sufficient to solve an inductive problem. This is a common strategy in debates about language acquisition, where models with weak inductive biases can be shown to be capable of learning aspects of language (Elman, 1990; Rumelhart & McClelland, 1986). The second approach is to compare computational models as accounts of human inferences, as reflected in laboratory data. By comparing models that differ in their inductive biases to the performance of human participants solving an inductive problem in the laboratory, it becomes possible to evaluate which set of inductive biases most closely matches those of human learners. This approach is common in many forms of computational modeling, and is dominant in the work on category learning mentioned earlier.

Recently, both of these approaches have been used with some success with probabilistic models of cognition, in which the inductive biases instantiated in the model are made particularly transparent (Anderson, 1990; Oaksford & Chater, 1998; Shepard, 1987). In these models, the beliefs of learners are expressed through subjective probability distributions. Assume that a learner has a set of hypotheses, H, and that his or her biases are encoded through a *prior* probability distribution, $P(h)$, specifying the probability the learner assigns to the truth of each hypothesis $h \in H$ before seeing some data d. Solving an inductive problem requires calculating the probability of each hypothesis h after seeing d, called the *posterior* probability $P(h|d)$. This can be done by using a principle of probability theory called Bayes's rule. This principle states that the posterior probability is

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')} \tag{1}$$

where $P(d|h)$, the *likelihood*, indicates the probability of the data d under hypothesis h. The posterior distribution can then be used to make decisions about which hypothesis was most likely to have been responsible for the observed data.

The prior probability distribution assumed by a learner encodes his or her inductive biases. A hypothesis with low prior probability will only be considered if it is supported by a great deal of evidence, reflected in having a higher likelihood than any other hypothesis. The transparency of the inductive biases of probabilistic models and the fact that Bayesian inference provides an optimal solution to inductive problems makes these models particularly useful for investigating human inductive biases. Probabilistic models of language (see Chater & Manning, 2006) can be used to explore what kinds of constraints are necessary in order to recover linguistic structure from observed data (e.g., Brent, 1999; Goldwater, Griffiths, & Johnson, 2006). Models of categorization, including those that use exemplar, prototype, and intermediate representations, can be given a probabilistic interpretation in which these representational differences amount to different assumptions about the nature of the probability distributions characterizing a category (Anderson, 1991; Ashby & Alfonso-Reese, 1995; Griffiths, Sanborn, Canini, & Navarro, 2007). Similarly, in causal learning, a learner who assigns lower prior probabilities to more complex hypotheses (cf. Chater, 1996; Feldman, 2000; Lombrozo, 2007) is expressing a bias to choose the simpler of two hypotheses that otherwise explain the observed data equally well. From this probabilistic perspective, identifying human inductive biases reduces to the problem of identifying the prior probability distributions assumed by human learners.

In this article, we explore a novel approach to using computational models to identify human inductive biases. This approach takes inspiration less from traditional computational models of cognition, and more from methods such as *mechanism design* in theoretical economics (Hurwicz, 1973). The basic idea behind mechanism design is to structure an interaction between agents in a way that provides them with incentives to produce a particular kind of behavior. For example, an auction can be structured so that every bidder should bid the true value that they assign to the prize, revealing information that might be concealed in a more conventional auction (Vickrey, 1961). Mechanism design proceeds from the assumption that the agents are rational utility maximizers, making it possible to predict their behavior in different situations. By analogy, assuming that learners are Bayesian agents can allow us to design tasks intended to reveal their inductive biases. More precisely, we can develop methods that will allow us to identify the prior distributions used by human learners, under the assumption that they solve inductive problems in a way that approximates Bayesian inference.

Pursuing this approach, we explore an experimental method based on a class of models of language evolution by "iterated learning" (Kirby, 2001). These models assume that every speaker of a language learns that language from another speaker, who had to learn it from somebody else in turn. Formally, we can imagine a sequence of learners, each of whom receives data from the previous learner, forms a hypothesis from those data and then generates new data that are passed to the next learner, as shown in Fig. 1. Griffiths and Kalish (2005, 2007) showed that iterated learning with Bayesian agents can be analyzed as a Markov chain, with the hypothesis selected by one learner depending only on the hypothesis selected by the previous learner. The stationary distribution of this Markov chain is the prior distribution on hypotheses assumed by the learners. This means that the probability that a learner selects a particular hypothesis converges to the prior probability assigned to that hypothesis as the number of iterations of learning increases. The rate of convergence is determined by the amount of information being passed from one learner to the next: If a large amount of information is

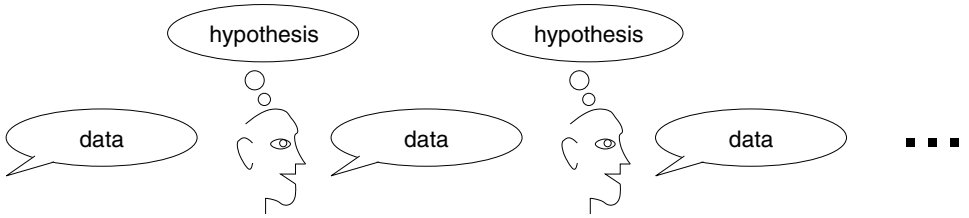

Fig. 1. In iterated learning, as described in models of language evolution, each learner chooses a hypothesis using data generated by the previous learner and generates data for the next learner based on that hypothesis.

being conveyed, then it strongly constrains the hypotheses considered by the next learner and convergence will be slow.

The convergence of iterated learning to the prior has significant implications for the connection between the biases of individual learners and linguistic universals (Griffiths & Kalish, 2007; Kirby, Dowman, & Griffiths, 2007), but goes beyond language evolution, applying to iterated learning with any kind of hypotheses and data. The fact that iterated learning converges to the prior suggests a simple procedure for investigating human inductive biases: implement iterated learning in the laboratory with human learners. As the number of iterations increases, the probability with which people select a hypothesis should converge to the prior probability of that hypothesis, providing a simple way to determine which hypotheses people are biased toward. This method for identifying human inductive biases has an important advantage over the traditional approaches to using computational models introduced earlier: It does not require the modeler to be able to postulate what an appropriate prior might be, or to implement a model instantiating this prior. The hypotheses involved can be extremely rich and complex, the priors might not be easily expressible in a compact mathematical formula, and iterated learning will still converge to an equilibrium that is informative about human inductive biases. Iterated learning does not require any kind of formal model of how people perform a task, allowing inductive biases to be revealed simply by inspecting the hypotheses that emerge.

Iterated learning also provides an alternative to standard experimental methods for investigating inductive biases. Much of the empirical work exploring human inductive biases focuses on the relative difficulty of learning or remembering hypotheses with different structures (e.g., in the domain of category learning; Feldman, 2000; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Shepard, Hovland, & Jenkins, 1961). This requires enumerating all structures of interest and determining how well people learn or remember each of these structures. This approach is typically applied in cases where there are only a few structures of interest, as the quantity of data that needs to be collected increases with the number of structures studied. Iterated learning offers a way to identify inductive biases in cases where there are more structures than can be easily enumerated. By producing samples from the prior, it can quickly reveal the structures that have high prior probability without needing to systematically explore all of the structures with low prior probability.

Our goal in this article is to demonstrate that iterated learning can be used to reveal the inductive biases of human learners. To do so, we need to use a task for which these inductive biases are well-known and quantifiable, so that we can establish that iterated learning produces the correct results. In previous work, we examined the outcome of iterated learning for

one-dimensional functions (Kalish, Griffiths, & Lewandowsky, 2007). The results were qualitatively consistent with the predicted outcome of iterated learning, converging to a distribution that favored positive linear functions—the functions that people find easiest to learn (Brehmer, 1971, 1974; Busemeyer, Byun, DeLosh, & McDaniel, 1997). However, even one-dimensional functions are relatively complex stimuli, and the development of a comprehensive model of human function learning that accurately characterizes people's inductive biases is an ongoing project (e.g., Kalish, Lewandowsky, & Kruschke, 2004). Here, we base our investigation on a different class of concepts: categories defined on stimuli characterized by three binary dimensions. Our experiments employ a generalization task, showing people objects labeled as to their category membership and asking them to identify the category from which those objects were drawn. This task is straightforward to model as a Bayesian inference, drawing on existing Bayesian models of generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001). Expectations about the relative prior probabilities of category structures are provided by the extensive literature on the difficulty that people have in learning and remembering different category structures defined on three binary dimensions (Feldman, 2000; Nosofsky et al., 1994; Shepard et al., 1961), and the prior can be estimated directly from generalization judgments. We can use the resulting Bayesian model to evaluate whether iterated learning produces the appropriate results, predicting not just its asymptotic consequences—convergence to the prior—but also the dynamics of this process.

The plan of the article is as follows. Section 2 discusses the formal analysis of iterated learning in detail, outlining the connection between iterated learning and inductive biases. Section 3 summarizes previous results exploring inductive biases for category structures defined on three binary dimensions. Section 4 presents a Bayesian model for a generalization task based on these category structures. Section 5 summarizes the strategy behind our experiments. Sections 6, 7, 8, and 9 present a series of experiments implementing this strategy. Section 10 explores the implications of these experiments.

2. Iterated learning and inductive biases

Iterated learning has been discussed most extensively in the context of language evolution, where it is seen as a potential explanation for the structure of human languages. Kirby (2001) explored this idea using the *iterated learning model* in which several generations of one or more learners each learn from data produced by the previous generation. The case where each generation consists of a single learner, which has been the focus of much of this work, corresponds to the schematic illustration of iterated learning shown in Fig. 1. The first learner is exposed to some initial data, forms a hypothesis about the language it represents, and generates new data from that language. These new data are passed to the second learner, who infers a hypothesis and generates data from it that are provided to the third learner, and so forth. Through simulations, Kirby and colleagues have shown that languages with properties similar to those of human languages, such as compositionality, can emerge from iterated learning with simple learning algorithms (Kirby, 2001; Smith, Kirby, & Brighton, 2003).

Griffiths and Kalish (2005, 2007) provided a formal analysis of the consequences of iterated learning when the learners are Bayesian agents. In this case, each learner uses Bayes's rule

to compute a posterior distribution over the hypothesis of the previous learner, samples a hypothesis from this distribution, and generates the data provided to the next learner using this hypothesis. The probability that the nth learner chooses hypothesis h_n given that the previous learner chose hypothesis h_{n-1} is

$$P(h_n|h_{n-1}) = \sum_d P(h_n|d)P(d|h_{n-1}) \tag{2}$$

where $P(h_n|d)$ is the posterior probability obtained from Equation 1. The sequence of hypotheses selected by the learners forms a Markov chain because the hypothesis chosen by each learner depends only on that chosen by the previous learner. Equation 2 defines the transition matrix of the Markov chain, Q, being a square matrix where q_{ij} indicates the probability with which a learner chooses hypothesis i given the previous learner chose hypothesis j.

Griffiths and Kalish (2005, 2007) showed that when the learners share the same prior, $P(h)$, the stationary distribution of the Markov chain on hypotheses induced by iterated learning is simply the prior. The Markov chain will converge to this distribution provided it is *ergodic,* a simple condition that is easily tested (Norris, 1997). This means that the probability that the last in a long line of learners chooses a particular hypothesis will be equal to the prior probability of that hypothesis, regardless of the data provided to the first learner. The rate of convergence of this Markov chain is determined by the second eigenvalue of the transition matrix, λ_2, with higher values of λ_2 being associated with slower convergence.[1] The value of λ_2 can be computed easily for any finite matrix using standard tools from linear algebra (e.g., Strang, 1988), and simulations show that one of the main determinants of the rate of convergence is the quantity of data passed from one learner to the next (Griffiths & Kalish, 2005, 2007).

The prediction that iterated learning will converge to the prior of the learners suggests that we may be able to investigate the inductive biases of human learners by implementing a similar process in the laboratory. The results on the rate of convergence provide guidelines for designing experiments that will converge to the prior. In the remainder of the article we develop a method for demonstrating that iterated learning can be used to reveal the prior distributions assumed by human learners. To do so, we need to design an inductive task that involves hypotheses with two properties. First, we need to have expectations about the prior probabilities of these hypotheses, allowing us to test whether the outcome of iterated learning is consistent with the prior. Second, we need a set of hypotheses that is sufficiently constrained to allow us to specify a Bayesian model from which we can obtain an estimate of the prior, allowing us to compute transition matrices and make quantitative predictions about the asymptotic outcome and dynamics of iterated learning. Category structures defined on three binary dimensions are one standard set of hypotheses that has both of these properties.

3. Inductive biases for category structures

Shepard et al. (1961) conducted a series of experiments exploring the relative difficulty of learning different kinds of category structures defined on objects that vary along three binary dimensions such as shape, color, and size. Categories are defined in terms of which subsets of

the eight possible objects they contain. In principle, there are 256 different category structures, but if we restrict ourselves to categories with four members, this number is reduced to 70. If we do not distinguish structures that differ only in the assignment of physical features to the binary dimensions, this number is reduced still further, giving us a total of 6 different types of category structures. Examples of categories belonging to these 6 types are shown in Fig. 2.

Shepard et al. (1961) found that there is great variation in the ease with which people learn and remember different types of category structures. Type I, in which membership is defined along a single dimension, is easiest to learn; followed by Type II in which two dimensions are sufficient to identify members. Next come Types III, IV, and V, which all correspond to a one-dimensional rule plus an exception and are about equally difficult to learn. Type VI, in which no two members share a value along more than one dimension, is hardest to learn. Similar results have been obtained by Nosofsky et al. (1994) and Feldman (2000).

Because difficulty in learning a hypothesis is an indication that it may be inconsistent with the inductive biases of the learner, these category structures provide a way to test the predictions of our theoretical account of iterated learning: We can examine whether iterated learning of category structures defined on three binary dimensions converges to a distribution over hypotheses consistent with the results of Shepard et al. (1961). We chose to study these inductive biases using a simple generalization task in which people infer the structure of the category from a small number of objects labeled as to their category membership. Based on the results of Shepard et al., we should expect to see category structures of different types being selected with different probabilities, with Type I and Type II structures dominating responses.

4. Modeling generalization and iterated learning

The set of category structures on three binary dimensions identified by Shepard et al. (1961) is relatively small, consisting of only 70 structures. This means that we can enumerate all of the hypotheses entertained by a learner and use a Bayesian model to obtain an estimate of people's priors from a source that is independent of iterated learning, allowing us to check that iterated learning results in convergence to this prior. Having an estimate of the prior also allows us to make precise quantitative predictions about the structures that we expect people to choose at each iteration of learning. Specifically, we can compute the transition matrix between hypotheses that we would expect to result from using this prior. We can then use this transition matrix to evaluate the probability that people will select a particular hypothesis in each iteration, and use the second eigenvalue of this transition matrix, λ_2, as a measure of the rate of convergence.

Our Bayesian model of the generalization task used in our experiments is based on the account of generalization developed by Shepard (1987) and extended by Tenenbaum (1999) and Tenenbaum and Griffiths (2001). The basic format of the generalization task is that people are shown some examples of objects that possess a particular property, such as belonging to a category, and then asked which other objects they think possess that property. No feedback is provided for these judgments—the task just explores how people reason from a set of examples to new instances. Shepard analyzed the case where a property is generalized from a

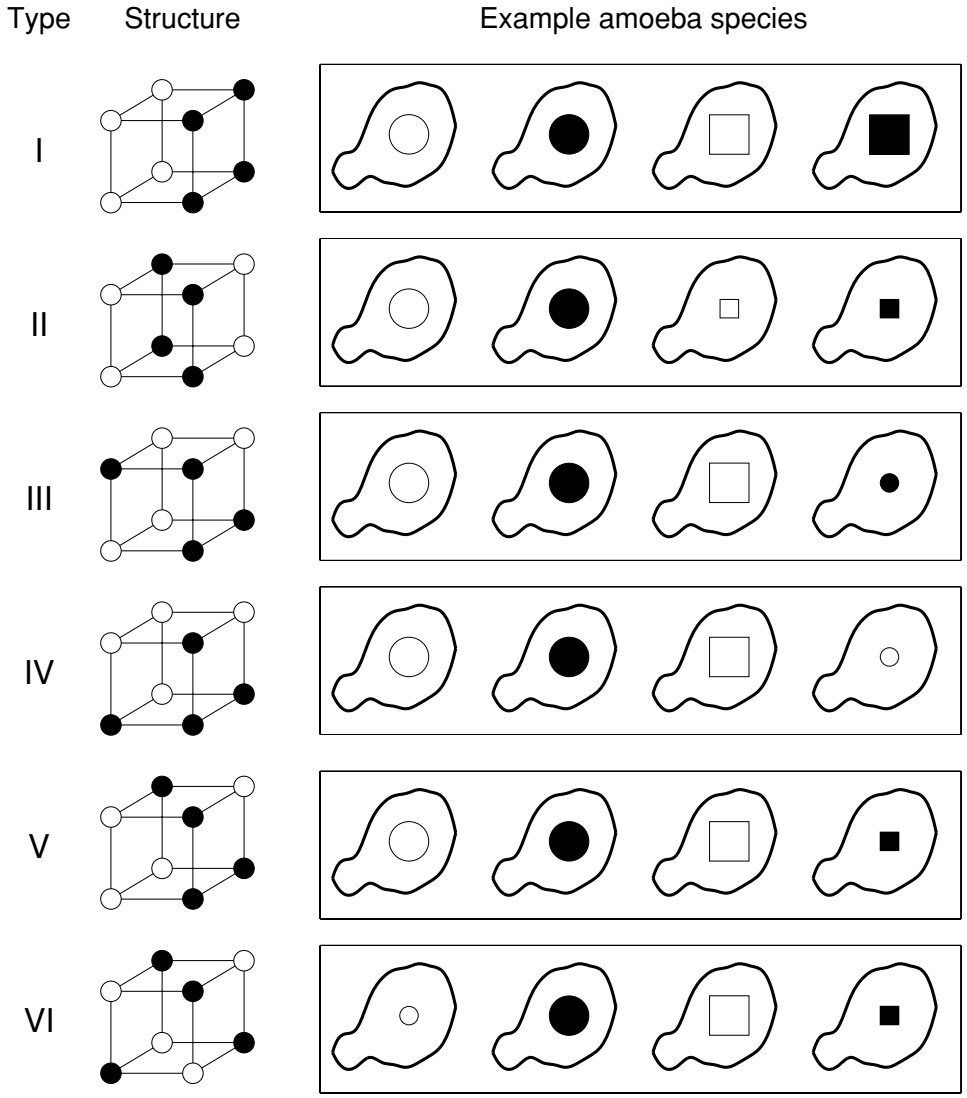

Fig. 2. Types of category structures containing four objects defined on three binary dimensions. The 70 possible category structures collapse to just six different types, corresponding to the numbers shown in the first column (Shepard, Hovland, & Jenkins, 1961). The second column shows one structure of each type. The three axes of each cube correspond to the three binary dimensions. The vertices are objects, with black dots corresponding to the members of the category. The third column shows the "species" of amoeba, of the kind used by Feldman (2000); and, in our experiments, matching each category structure shown in the second column. The nuclei of the amoebae vary along three binary dimensions: color (white or black), shape (circle or square), and size (small or large).

single example, yielding his famous universal law of generalization. Tenenbaum and Griffiths extended this analysis to the case where multiple examples are provided—a task that more closely resembles that used in "concept learning" experiments (e.g., Bruner, Goodnow, & Austin, 1956).

The generalization task differs in several ways from the task used in most studies of category learning. In a standard category learning experiment, participants are taught to discriminate between two categories of objects by guessing the category label of a series of example objects and receiving feedback on the accuracy of their guesses. In contrast, the generalization task uses only a single category and examples are presented directly, with neither presentation nor response yielding feedback. The task is thus more similar to that used in studies of property induction (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975) in which premises provide information about a set of objects that possess a particular property and people evaluate conclusions indicating whether other objects possess that property. Bayesian models similar to the account of generalization we give here have also been applied to property induction (e.g., Heit, 1998; Kemp & Tenenbaum, 2003).

Our experiments test generalization from two different kinds of data. In some conditions participants are shown *positive* examples—objects sampled at random from those that belong to the category, whereas in other conditions they are shown *mixed* examples—objects sampled at random from the set of all objects, and labeled according to their category membership. Both cases can be viewed as problems of Bayesian inference, where the data d are the set of labeled objects and each hypothesis h is a possible category structure with four members. Using both positive and mixed examples allows us to evaluate the effect of the amount of information that data d provides about hypotheses h on the rate of convergence of iterated learning, as these different kinds of data convey different amounts of information. In outlining our Bayesian model, we consider these two forms of data in turn.

Positive examples are objects sampled from the category. For instance, if we were learning about a new species of amoebae, a few examples of members of that species would constitute positive examples. We assume that these objects are sampled without replacement, meaning that the first example has equal probability of being any of the $|h|$ objects in the category h,the second example has equal probability of being any of the $|h|-1$ remaining objects, and so forth.[2] Following this assumption, the probability of a particular sequence of m positive examples is obtained by multiplying together $1/|h|, 1/(|h|-1), \ldots, 1/(|h|-m+1)$, to give

$$P(d|h) = \begin{cases} (|h|-m)!/|h|! & d \subset h \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $d \subset h$ indicates that all m objects in d are members of h (and $m < |h|$). We will refer to a hypothesis h such that $h \supset d$ as being *consistent* with the data d.

Equation 3 provides the likelihood function that we can use to define a Bayesian model indicating how a rational learner should evaluate the posterior probability of any hypothesis h given some set of objects d. Applying Bayes's rule (Equation 1) with this likelihood and a prior on hypotheses, $P(h)$, we obtain

$$P(h|d) = \frac{P(h)\,(|h|-m)!/|h|!}{\sum_{h' \supset d} P(h')\,(|h'|-m)!/|h'|!} \tag{4}$$

for each hypothesis h consistent with d, and 0 otherwise. In our case, all categories are of the same size ($|h| = 4$), so all hypotheses consistent with d have posterior probability:

$$P(h|d) = \frac{P(h)}{\sum_{h' \supset d} P(h')} \tag{5}$$

which is simply the prior normalized over the set of consistent hypotheses (normalization ensures that the resulting posterior probabilities sum to 1). This Bayesian model thus gives a very simple prescription for how a rational agent should update his or her beliefs about which category structures are likely to be present upon observing a set of examples drawn from a category: Assign each category structure that is consistent with the observed examples a posterior probability proportional to its prior probability.

Mixed examples are objects sampled at random and then labeled as to their category membership. For instance, one could obtain mixed examples for a species of amoebae by collecting a number of amoebae and asking a teacher to label each as to its membership of the species. Assuming that the objects that will receive labels are sampled without replacement from a set of k objects (e.g., all possible amoebae), and each object is equally likely to be chosen, the probability of a particular ordered sequence of objects is $1/k$ for the first object multiplied by $1/(\mathrm{k} - 1)$for the second object, down to $1/(k - m + 1)$for the mth object, to give

$$P(d|h) = \begin{cases} (k-m)!/k! & d \dashv h \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $d \dashv h$ indicates that the labels for the objects in d match the category structure h (i.e., all objects labeled as belonging to the category are in h,and all objects labeled as not belonging to the category are not in h). As with positive examples, we will say that any hypothesis h such that $h \vdash d$ is *consistent* with the data d.

Once again, we can substitute the likelihood function given by Equation 3 into Bayes's rule (Equation 1) to obtain the posterior probability of each hypothesis. Because $P(d|h)$ is the same for all hypotheses h consistent with d, it follows that the posterior probability of h given d is

$$P(h|d) = \frac{P(h)}{\sum_{h' \vdash d} P(h')} \tag{7}$$

for all consistent hypotheses, and 0 otherwise, for any choice of k and m. As with positive examples, the posterior distribution is simply the prior, normalized over the hypotheses consistent with the data. The posterior probability of each category structure that is consistent with the data is thus proportional to its prior probability.

Equations 5 and 7 indicate how individual Bayesian learners should update their beliefs. To turn the generalization task into a form of iterated learning, we simply generate the data seen by the next learner in a way that depends on the hypothesis selected by the previous learner. In the simplest of our experiments, participants select a hypothesis consistent with the labeled set of objects that they observed, and the data seen by the next participant are generated by a computer sampling objects and labels using the selected hypothesis. Specifically, the computer follows the generative processes implicit in the likelihood functions given by Equations 3 and 6. For positive examples, m objects are sampled

without replacement from the set picked out by h. For mixed examples, m objects are sampled without replacement from the set of all objects, and then labeled in the way dictated by h.

The likelihoods given by Equations 3 and 6 and posterior distributions from Equations 5 and 7 can be substituted into Equation 2 to find the transition matrix of the Markov chain on hypotheses induced by iterated learning with positive and mixed examples respectively. The result is a square matrix, Q, where the number of rows and columns is equal to the number of hypotheses. For the category structures studied by Shepard et al. (1961), this is a 70×70 matrix because there are only 70 category structures that can be considered as hypotheses. Given an initial distribution over hypotheses, represented as a column vector p where p_i is the probability of hypothesis i, the distribution over hypotheses at each subsequent iteration can be computed by multiplying this vector by the transition matrix, with the distribution after n iterations being $Q^n p$ (Norris, 1997). This can be used to make predictions not just about the asymptotic distribution over hypotheses, which we know to be the prior, $P(h)$, but also the dynamics of iterated learning.

Although the asymptotic distribution over hypotheses will not be affected by the quantity of data seen by the learners, the dynamics will depend on the degree to which the data constrain the choices of the learners. The posterior distributions given in Equations 5 and 7 are both simply the prior normalized over the set of hypotheses consistent with the data. The extent to which the data constrain the choices of the learners is thus directly reflected in the number of hypotheses that are likely to be consistent with the set of examples comprising the data. This will differ for positive and mixed examples and vary with the number of examples provided. With three positive examples, three members of the category are mixed, so there are five possibilities for the fourth member of the category (i.e., the 5 objects that remain from the 8 produced by considering all possible values of three binary dimensions), resulting in five hypotheses. An illustrative set of three positive examples and the five consistent hypotheses are shown in Fig. 3a.

Seeing three examples drawn from a category provides fairly strong constraints, narrowing 70 hypotheses down to 5. The strength of the constraints provided by a set of examples is simply a logical consequence of the nature of the category structures we are exploring and not a property of our model. A similar calculation shows that 15 hypotheses are consistent with any set of two examples. There are six candidates for the third member of the category, and five candidates for the fourth member. Because we consider all categories with the same set of members equivalent, we can ignore the two orders in which these two new members could be chosen, giving a total of $(6 \times 5)/2 = 15$ hypotheses. One set of two positive examples and its corresponding hypotheses are shown in Fig. 3b. With three mixed examples, the cases where the three examples are all positive or all negative result in 5 consistent hypotheses (corresponding to the 5 options for the fourth member or non-member), whereas the cases where the three examples are genuinely mixed result in 10 hypotheses (5 candidates times 5 candidates divided by 2 orders). A case with three mixed examples and 10 hypotheses appears in Fig. 3c. With four mixed examples, four positive or four negative examples perfectly determine a single hypothesis, three positive and one negative (or vice versa) allow 4 hypotheses (being the candidates for the fourth member or non-member), and two positive and two negative allow 6 hypotheses (4 candidates times

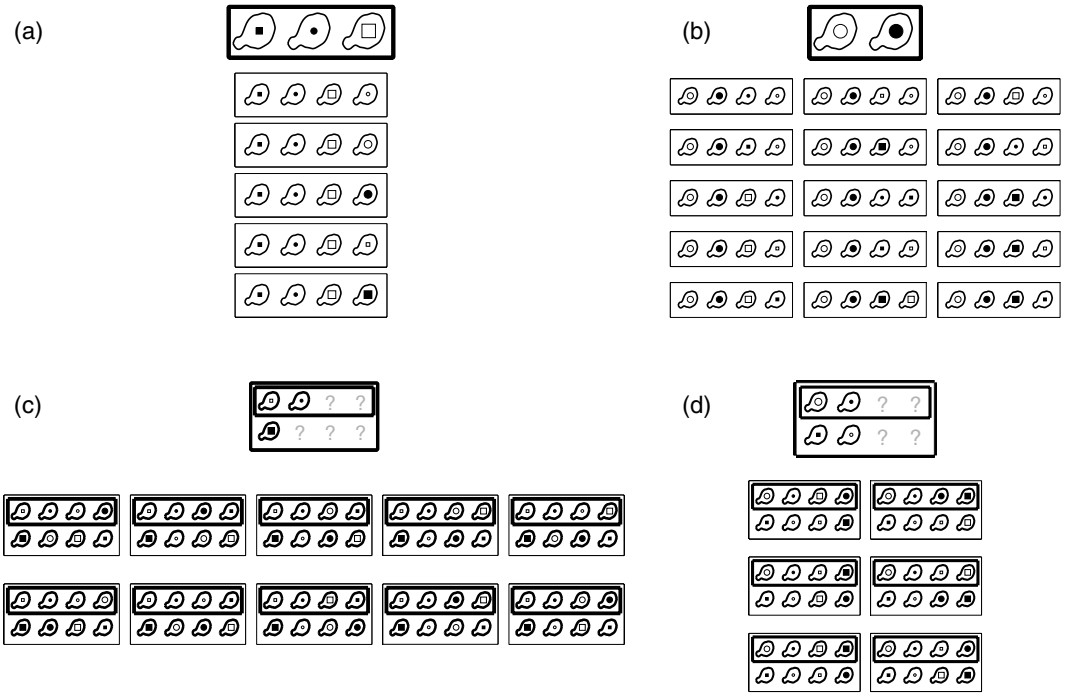

Fig. 3. Sample displays showing sets of examples and the corresponding consistent hypotheses. The heavy box at the top of each panel contains a set of positive or mixed examples of members of a "species" (i.e., category). For positive examples, only the known members of the species are shown. For mixed examples, known members are shown inside a smaller box, and known non-members are shown outside this box. The lighter boxes below show all possible completions of this species—the hypotheses that are consistent with these data. These displays were also those used in the experiments described later in the article. (a) With three positive examples, there are 5 hypotheses consistent with each set of examples. (b) With two positive examples, there are 15 hypotheses consistent with each set of examples. (c) With three mixed examples, either 5 or 10 hypotheses are consistent (with 10 being consistent for this particular sample). (d) With four mixed examples, either 1, 4, or 6 hypotheses are consistent (with 6 being consistent for this particular sample).

3 candidates divided by 2 orders). Four mixed examples and 6 hypotheses are shown in Fig. 3d.

The results in the previous paragraph indicate that the number of positive or mixed examples transmitted between people determines how strongly the data constrain the hypotheses under consideration. One of the outcomes of our formal analysis of iterated learning is thus that the rate of convergence will be affected by this quantity. If we can estimate a prior distribution over hypotheses, we can compute the transition matrix for the Markov chain and make precise predictions about the dynamics of iterated learning. We can also get a general measure of the convergence rate from the second eigenvalue of this transition matrix, λ_2. Values of λ_2 closer to 1 will indicate a slower rate of convergence, providing a simple quantitative measure of the effect of the amount of information transmitted between people on the convergence of iterated learning.

5. Design of experiments with human learners

In the remainder of the article, we present three experiments. The goal of these experiments is to demonstrate that iterated learning can be used to reveal the inductive biases of human learners, as reflected in the prior probability distribution that we estimate from people's responses using our Bayesian model. As a secondary objective, we analyze whether the fine-grained dynamics of this process are consistent with the Bayesian framework presented earlier. Analyzing the dynamics of iterated learning is primarily of interest as a means of gaining insight into the factors that influence convergence, which will be important for designing experiments that are informative about people's priors.

The basic element used in the design of these experiments is a set of 12 "chains" of trials, as shown in Fig. 4a. Each trial consists of the presentation of data (a set of examples from a category) and selection of a hypothesis (a category structure) by a learner. The chains include six *dependent* chains, one for each type of category structure. These chains are initialized by randomly selecting a hypothesis of the appropriate type and generating data in accordance with that hypothesis. In the first iteration, the learner chooses the hypothesis he or she believes to have produced these data. In the second iteration, the data seen by the learner are generated in accordance with the hypothesis selected in the first. If the data are positive examples, they are sampled without replacement from the members of the category structure chosen in the first iteration. If the data are mixed examples, a set of objects is sampled without replacement and those objects are then labeled as members or non-members in accordance with the category structure chosen in the first iteration. The learner selects the hypothesis he or she believes generated these data, and this hypothesis is used to generate the data seen in the third iteration. This procedure continues until all chains are complete. As a control, we also include six *independent* chains in which data are generated at every iteration using the initialization procedure for the dependent chains—sampling a category structure of one of the six types (depending on the number of the chain) and generating data from that hypothesis. Consequently, the hypotheses selected at every iteration are independent. These independent chains provide us with a way to measure any time-varying effects on responding, allowing us to establish that changes in the distribution over hypotheses over time in the dependent chains are the consequence of iterated learning and not simply practice or fatigue. Finally, aggregating responses across dependent and independent chains provides us with a large number of individual decisions that can be used to estimate a prior over category structures for comparison to the outcome of iterated learning.

The variable of interest in our experiments is the proportion of trials on which people select hypotheses corresponding to category structures of the six different types. The quantities that we manipulate within each experiment are the iteration of learning and whether the chains are dependent or independent. Our formal analysis of iterated learning indicates that as the number of iterations of learning in the dependent chains increases, the proportion of trials on which people select hypotheses corresponding to each type of category stucture will come to reflect the prior distribution over types. Our design allows us to examine the extent to which this is the case both by aggregating across all dependent chains, and thus examining how the probabilities of the different types change from the uniform distribution used in initialization, and by looking at how the probabilities of the different types change

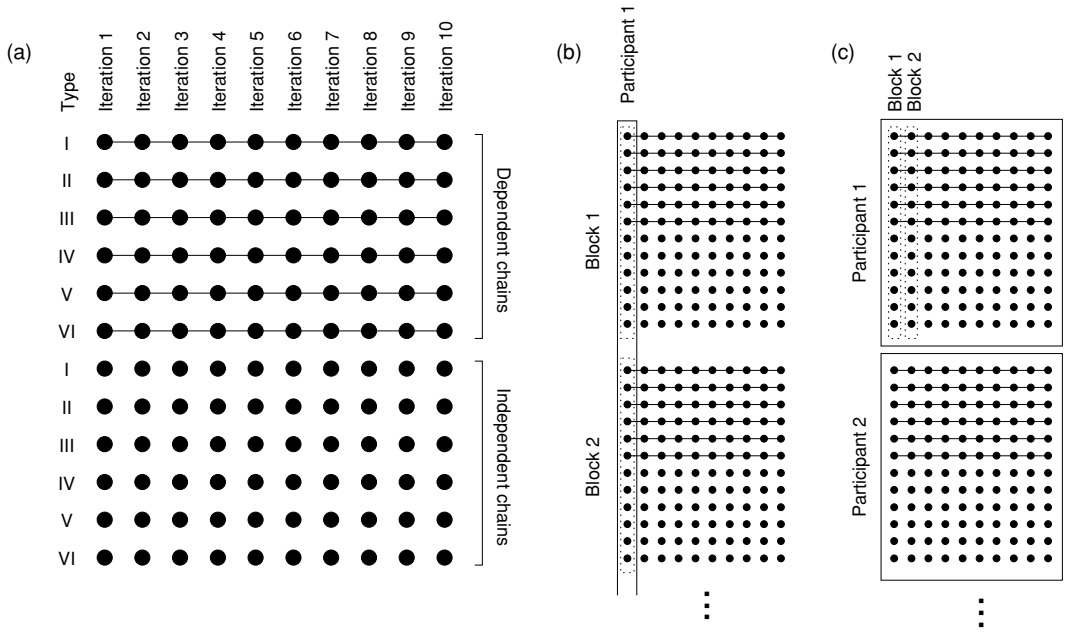

Fig. 4. Design of experiments with human learners. (a) Basic element used in the design of iterated learning experiments. Each dot represents a trial, and links between dots indicate that the hypothesis selected on one trial is used to generate the data provided on the next trial. Trials are divided into 12 chains, each consisting of 10 iterations of learning. Each chain is assigned to one of the six types of category structures identified by Shepard, Hovland, and Jenkins (1961). At initialization, a hypothesis is selected at random from the set of category structures matching the type of that chain and data are generated from that hypothesis, as described in the text. The learner chooses a hypothesis on the basis of those data. For dependent chains—those with links between hypotheses—this hypothesis is used to generate the data seen on the next trial. For independent chains—those without links—the data seen on the next trial is generated via the same procedure as initialization. (b) In a between-subject experiment, each iteration involves a different learner, with data being passed between learners in the dependent chains. Each learner receives 10 blocks of trials, where each block consists of one trial from a dependent and an independent chain of each type, and their responses in each block are used to generate the data for the dependent chains in that block for the next participant. Blocks are illustrated by dotted boxes, participants by solid boxes. (c) In a within-subjects experiment, each participant completes a full set of 10 iterations on each of 12 chains, and data are not passed between participants.

within a chain initialized with a particular type of category structure. Because convergence to the prior should hold regardless of the initial data provided to the learners, we can check for convergence by comparing across chains initialized with category structures of different types. Finally, by manipulating the number and nature of the examples provided to the learners, we can examine how the strength of the constraints implied by these data influence the rate at which convergence occurs.

In addition to these variables, we explore two different ways of conducting experiments using iterated learning. Our discussion so far has focused on cases where iterated learning occurs between-subject, with each learner seeing data generated by a previous learner, as shown in Fig. 4b. This is the natural way to think about iterated learning in the context of language evolution, and was the focus of our previous experiments (Kalish et al., 2007).

However, if our main interest is in using iterated learning as a method for revealing inductive biases, we might want to use a different design. Our formal analysis is equally valid for iterated learning with a within-subjects design, where a single learner responds to data that are generated from his or her own previous responses, as shown in Fig. 4c. This means replacing successive generations of learners with successive blocks of learning. Each block contains a single trial from each of the 12 chains, and responses influence the data provided in the next block for the dependent chains. By completing a sequence of blocks, each participant produces his or her own set of chains, with no need for data to be passed between participants.

These two designs push on different assumptions behind our formal analysis. In the between-subjects case, the assumption that all learners share the same prior may seem like an oversimplification, but it is clear that participants will select hypotheses in a way that depends only on the data produced by the previous participant and not on any previous data or hypotheses (i.e., that we have a Markov chain). In the within-subjects case, a chain of hypotheses is clearly being generated based on the same prior because the same participant is selecting the hypotheses, but it is possible that previous decisions influence the current decision, violating the Markov assumption. However, the within-subjects design has a major advantage over the between-subject design for any large-scale exploration of iterated learning, as it does not require coordinating responses between participants: each participant produces a complete set of independent and dependent chains.

Our experiments are not intended to test whether people are Bayesian, or to evaluate specific hypotheses about what human inductive biases are like for category structures defined on three binary dimensions. Provided people behave in a way that is consistent across judgments, selecting hypotheses with a fixed stochastic preference ordering, these judgments will be well-approximated by a Bayesian model with an appropriate prior. If this is the case, we would expect the distribution over hypotheses produced by iterated learning to converge to this prior. However, there are still plenty of reasons why this might not occur, due to the violations of the assumptions behind our formal analysis mentioned earlier—variation in priors in a between-subject experiment, and possible failure of the Markov assumption in a within-subjects experiment. Our goal is simply to demonstrate that these experimental designs actually do produce distributions over hypotheses that reflect human inductive biases, supporting the use of this method for this purpose in the future.

Based on these considerations, Experiments 1A and 1B compare the results of iterated learning of category structures from three positive examples using both within- and between-subjects designs. The aims of these experiments are to provide a basic test of the prediction that both forms of iterated learning should result in convergence to the prior, and to establish a correspondence between the results of the two designs. Experiment 2 uses the within-subjects design to explore the dynamics of iterated learning, examining the correspondence between human responses and model predictions for iterated learning of category structures from two positive examples, three mixed examples, and four mixed examples. The aim of the experiment is to investigate the dynamics of iterated learning, and to establish that convergence to the prior is observed in all cases. In all experiments, it is expected that the prior distribution on category structures will exhibit the bias in favor of Type I and Type II structures seen in previous research, and that the responses people produce will converge toward this prior.

6. Experiment 1A: Iterated learning within-subjects

6.1. Method

6.1.1. Participants

There were a total of 73 participants, consisting of 53 undergraduates from the University of Louisiana, Lafayette participating in exchange for course credit, and 20 members of the Brown University community who were compensated at a rate of $7 per hour.

6.1.2. Stimuli

Following Feldman (2000), stimuli were presented as "amoebae" with a wavy cell wall and an internal nucleus. Nuclei varied along three binary dimensions: shape (round or square), color (black or white), and size (large or small). Categories were "species" of amoebae, and each possible species was illustrated using a set of four amoebae in a black box. Examples of species of the six types appear in Fig. 2. Each trial of the experiment involved the presentation of three positive examples from a category, with participants selecting one of the five category structures that were consistent with those examples. To remove memory demands, examples and hypotheses were presented simultaneously on a computer screen, as shown in Fig. 3a. The examples appeared in a heavy box at the top of the screen, with the consistent hypotheses shown in lighter boxes below them. Each consistent hypothesis was represented by a complete species of four amoebae, of which the first three matched the examples. Participants selected the category structure they considered most likely to have generated the observed examples by clicking on one of the lighter boxes using a mouse.

6.1.3. Procedure

The design of the experiment followed the structure outlined in Section 5 and shown in Fig. 4c. Each participant completed 120 trials divided into 10 blocks of 12 trials. Each block corresponded to a new iteration of learning, resulting in a total of 10 iterations of learning across the 12 chains. Within each block, 6 trials belonged to dependent chains, with the objects being generated at random from the hypothesis selected on the previous trial in that chain. The other 6 trials within each block were part of independent chains, with the objects being generated from a randomly selected category structure. There was one dependent and one independent chain for each of the 6 types of category structures. Each dependent chain was initialized by sampling a category structure of the appropriate type and then generating data from that category structure. The same procedure was used to generate the data shown to the participant on each trial in the independent chain. Trials were randomized within blocks, reducing the opportunity for participants to notice any correspondence between trials in the dependent chains.

Each participant completed the experiment in a sound-attenuated booth, with the entire experiment being administered by computer. On starting the experiment, participants received the following instructions:

> In this experiment, you will be asked to infer the structure of different species of amoebae from sample members of that species.

> Each species will be presented on a single screen. At the top of the screen, you will see three amoebae from that species. An amoeba has a nucleus, which is either black or white, large or small, and round or square, and amoebae vary only in their nuclei. Every species has exactly four members.
>
> Your task is to select your best guess at the identity of the entire species based on the three you see at the top of your screen. The five boxes shown on the bottom of the screen represent the possible species consistent with the sample members.
>
> Clicking once on a box of four amoebae will highlight it, and a second click will submit that as your answer. If you are sure about your answer straight away, you can just double-click on a box.
>
> Each time the screen refreshes a new species is generated, and you have to decide what you think the structure of that species might be.

Participants then completed 4 practice trials before going on to the main experiment. No feedback was provided during the task.

The species selected by the participants were recorded on each trial. Because each species corresponds to a category structure of a particular type, these responses can be used to determine the proportion of trials on which structures of different types were selected, for each chain and iteration, allowing us to examine how this proportion changes over time.

6.2. Results

The variable of interest in this study was the proportion of trials on which people selected category structures of each type, with the key prediction being that these proportions would converge toward the prior as the number of iterations in the dependent chains increased. Although previous work provides qualitative guidelines about the relative prior probabilities of the six different types of category structure, the model presented in Section 4 allows us to estimate these quantities directly, to use them as the basis for evaluating whether the outcome of iterated learning reflects the prior, and to make quantitative predictions about its dynamics. We defined the prior over hypotheses, $P(h)$, by assuming that the prior probability was affected only by the type of category structure to which a hypothesis corresponds, being uniform within types. Because there are six such types, the prior can be completely specified by five parameters, giving the probabilities of structures of Types I through V (the probability of structures of Type VI follows from the fact that probabilities sum to 1). By estimating this prior from people's responses, we can make quantitative predictions about the dynamics of iterated learning.

The parameters were estimated from the frequencies with which participants selected hypotheses given different sets of examples, aggregated across dependent and independent chains. According to the Bayesian model, people's choices should follow the distribution given by Equation 5. Parameters were found using maximum-likelihood estimation for this distribution. Preliminary analyses indicated that a subset of the participants were responding at random, so a variant of the Expectation–Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) was used to simultaneously estimate the prior and probabilistically classify participants as either responding at random or in accord with the model (see the appendix for details). The resulting parameter estimates gave individual structures of Types I through

VI prior probabilities of 0.109, 0.032, 0.002, 0.001, 0.003, and 0.022, respectively.[3] These probabilities are consistent with previous findings concerning the relative difficulty of learning different types of category structures, with the only exception being the relatively high probability of Type VI structures, which we consider in more detail in Section 10.1. Using these parameters and the procedure outlined in the appendix, we computed the probability that each participant was responding at random. The remainder of the analyses reported here use only the 64 participants for whom this probability was less than 0.5. A more detailed analysis of the responses of the other participants appears in the appendix.

The left-most panel of Fig. 5a shows how the proportion of participants selecting category structures of each type varies as a function of the number of iterations, aggregating over all six dependent chains. The same quantities for the independent chains are shown in Fig. 5c. To evaluate whether iterated learning influenced responses, we ran a chi-square test comparing the proportions of the six types across the dependent and independent chains at each iteration. The results of these tests are shown in the first row of Table 1. There was a statistically significant difference between the dependent and independent chains on every iteration after the third, indicating that iterated learning influenced the distribution over the category structures being selected. Fig. 5 shows that this difference was due to an increase in the probability of Type I structures and a decrease in the probability of Type VI structures in the dependent chains, consistent with a prior favoring Types I and II over Type VI.

Fig. 5b shows the predictions of the Bayesian model outlined in Section 4 for the dependent chains. These predictions depend only on the prior, and were generated using the prior estimated from the aggregated generalization judgments from the dependent and independent chains rather than directly optimizing the parameters of the model to produce the best fit to the overall dynamics. As can be seen from the figure, there is a close correspondence between

Table 1
Results of Chi-Square Tests for Effects of Iterated Learning and Design in Experiments 1A and 1B

	Iteration									
Variable	1	2	3	4	5	6	7	8	9	10
Dependent vs. Independent										
Within-subjects (Expt. 1A)	0.17	8.06	10.91	20.88*	23.53*	27.60*	21.44*	26.80*	47.38*	31.31*
Between-subjects (Expt. 1B)	7.39	1.55	10.35	2.10	8.88	18.12*	24.23*	25.39*	33.38*	30.78*
Within- vs. Between-subjects (comparing Expts. 1A and 1B)										
Dependent	14.50	9.46	13.25	11.16	15.79	11.95	9.38	10.35	18.92*	2.55
Independent	28.04*	5.70	4.53	2.52	6.26	2.21	7.64	3.41	4.85	5.17

Note. Asterisks indicate statistical significance for $\chi^2(5)$ at $p < .005$, using the Bonferroni correction to maintain an overall false positive rate below .05 for each sequence of tests.

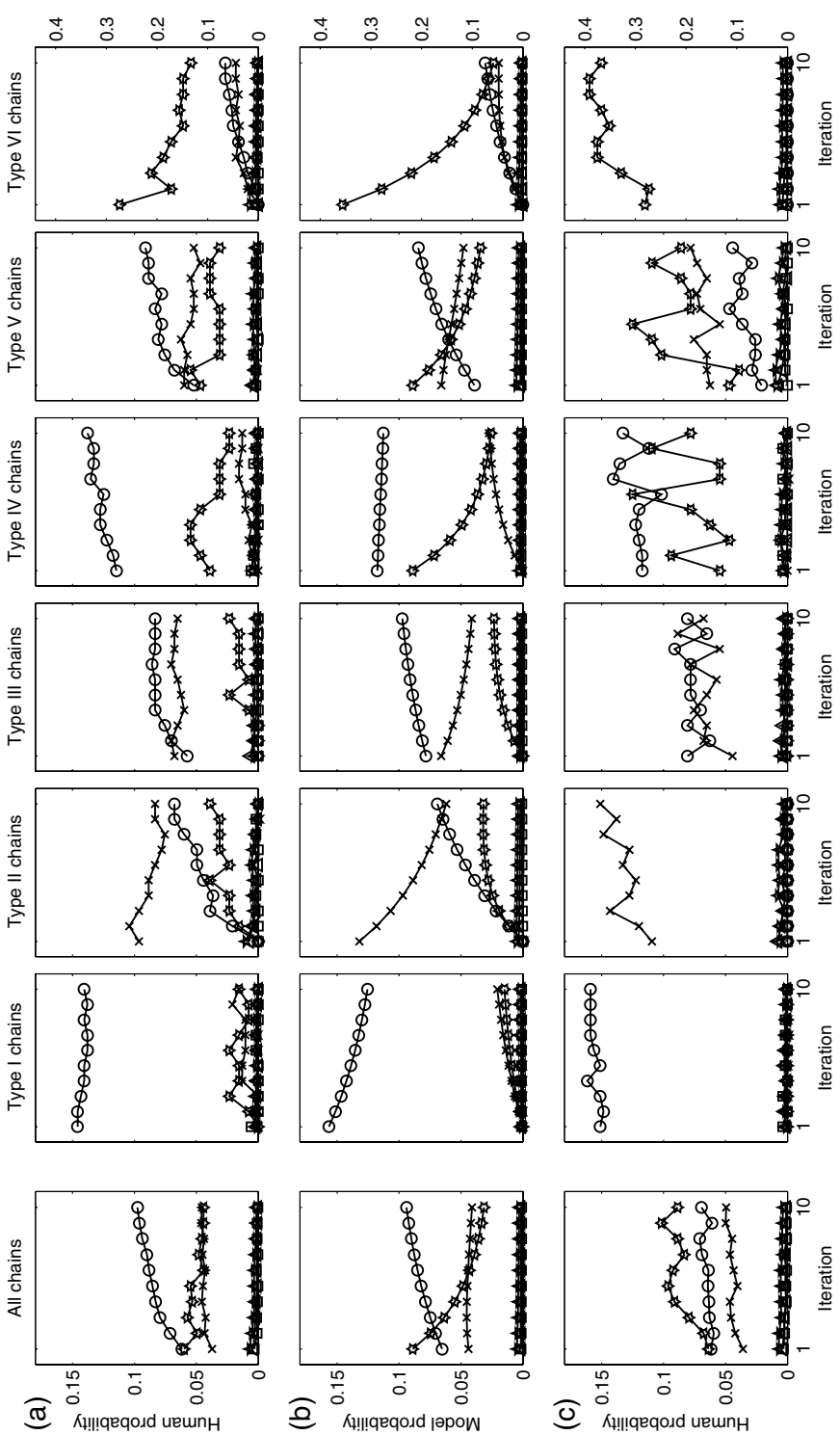

Fig. 5. Results of Experiment 1A, in which iterated learning was done within-subjects with three positive examples, for (a) dependent chains, (b) the Bayesian model, and (c) independent chains. Each plot shows the probability that a participant selected a particular category structure of a given type as a function of the number of iterations of learning. Types are distinguished by the number of strokes in the markers: Type I is a circle, Type II a cross, Type III a triangle, Type IV a square, Type V a five-pointed star, and Type VI a six-pointed star. The left-most plots collapse across all chains, whereas the six plots on the right break these data up based on the type of the category structure used to initialize the chain. The Bayesian model predicts response probabilities across all iterations of the independent chains that match the first iteration of the dependent chains. The vertical axis uses the same scale for all plots except that for Type VI chains.

the predictions of the model and the human data, with a linear correlation of $r(58) = 0.981$. In particular, both the model and the human data move toward an asymptotic distribution over hypotheses consistent with the prior. The model predictions for the independent chains correspond to the probabilities for the first iteration of the dependent chains, held constant across all iterations. The mean probabilities exhibited in the independent chains are consistent with these predictions, and the variation around the mean provides a guide to the amount of noise in the data.

Despite moving toward the prior, the Markov chains defined by the human responses and the Bayesian model in the dependent chains do not actually converge. As a simple test of convergence, we can compare the distributions over types of category structures obtained from the six dependent chains initialized with different types. Because the stationary distribution is the same regardless of the initialization of the chains, the distributions over types of category structure in the final iteration should be the same if the chains have converged. A chi-square test for independence on the human data shows that this is not the case—these distributions are statistically significantly different: $\chi^2(25) = 96.69$, $p < .0001$. The remaining panels of Fig. 5 show a more detailed analysis of the data and model predictions, breaking down the chains by the type of category structure used in their initialization. The lack of convergence is apparent in the differences in the distributions over types in the final iteration. However, these individual chains still show evidence of movement toward the prior, with Type I structures increasing in prevalence while Type VI structures decrease, and the model and data still exhibit a strong correlation, with $r(358) = 0.955$.

7. Experiment 1B: Iterated learning between-subjects

7.1. Method

7.1.1. Participants

There were a total of 75 participants, consisting of 29 undergraduates from the University of Louisiana, Lafayette participating in exchange for course credit, and 46 members of the Brown University community who were compensated at a rate of $7 per hour.

7.1.2. Stimuli

Stimuli were the same as those in Experiment 1A.

7.1.3. Procedure

The design of the experiment followed the structure outlined in Section 5 and shown in Fig. 4b. Participants were divided into seven "families" of learners. Each participant completed 120 trials, corresponding to 10 blocks of 12 chains. Within each block, 6 trials belonged to dependent chains, with the objects being generated at random from the hypothesis selected by the previous participant in the family. The other 6 trials within each block were part of independent chains, with the objects being generated from a randomly selected hypothesis corresponding to 1 of the 6 types. Trials were randomized within blocks, and there was no contact between participants except through the data passed along the dependent chains.

Based on the results of Experiment 1A, it was apparent that some participants responded to the task at random. These participants were screened out by computing the probability that each participant was responding at random using the parameters for the prior estimated in Experiment 1A, and removing those participants for whom this probability was greater than 0.5. A total of 5 participants were removed for responding at random and were replaced before further participants were run from that family, resulting in a total of 7 complete families of 10 learners.

7.2. Results

The variable of interest in this study was the proportion of trials on which people selected category structures of each type, with the key prediction being that these proportions would converge toward the prior as the number of iterations in the dependent chains increased. As in Experiment 1A, the prior probabilities of category structures of each of the six types were estimated by maximum-likelihood estimation, assuming that participants were making their choices in the dependent and independent chains by sampling from the distribution given in Equation 5. The estimated probabilities of structures of the six types were 0.097, 0.040, 0.002, 0.001, 0.003, and 0.029 for Types I through VI respectively, very close to the corresponding values in Experiment 1A. The proportion of trials on which each type of category structure is shown as a function of iteration in Fig. 6. The second row of Table 1 gives the results of a series of chi-square tests for differences between the distributions over types of structures in the dependent and independent chains at each iteration. There is a statistically significant difference between the dependent and independent chains at all iterations after the fifth, indicating an effect of iterated learning. As in Experiment 1A, the proportion of Type I structures increases over time and the proportion of Type VI structures decreases over time for the dependent chains, consistent with a prior favoring Types I and II over Type VI. Fig. 6b shows the predictions of a Bayesian model using the estimated prior for the dependent chains. The model predictions and human data are in close correspondence, with the correlation between the two being $r(58) = 0.991$.

Consistent with the results of Experiment 1A, the proportion of responses corresponding to each type of category structure gradually moved toward the prior, but did not converge by the end of the experiment. A chi-square test comparing the distributions on types at the end of the experiment showed a statistically significant difference between the six dependent chains: $\chi^2(25) = 47.07$, $p < .005$. The proportion of trials on which category structures of different types were selected in these six chains is shown as a function of iteration in Fig. 6, together with the corresponding model predictions. The difference in the distributions across these chains in the final iteration reflects the lack of convergence. The model predictions correlate well with the human responses, with the correlation across all chains being $r(358) = 0.985$.

Comparison of Figs. 5 and 6 shows that the results produced using the between-subject design were extremely close to those of the within-subjects design. The responses shown in the figures correlated at $r(58) = 0.966$ overall and $r(358) = 0.951$ when broken down across the six types of dependent chain. To test for a difference between the two experiments, we conducted a series of chi-square tests comparing the distribution of types of category structure selected at each iteration, for both the dependent and the independent chains. The results are

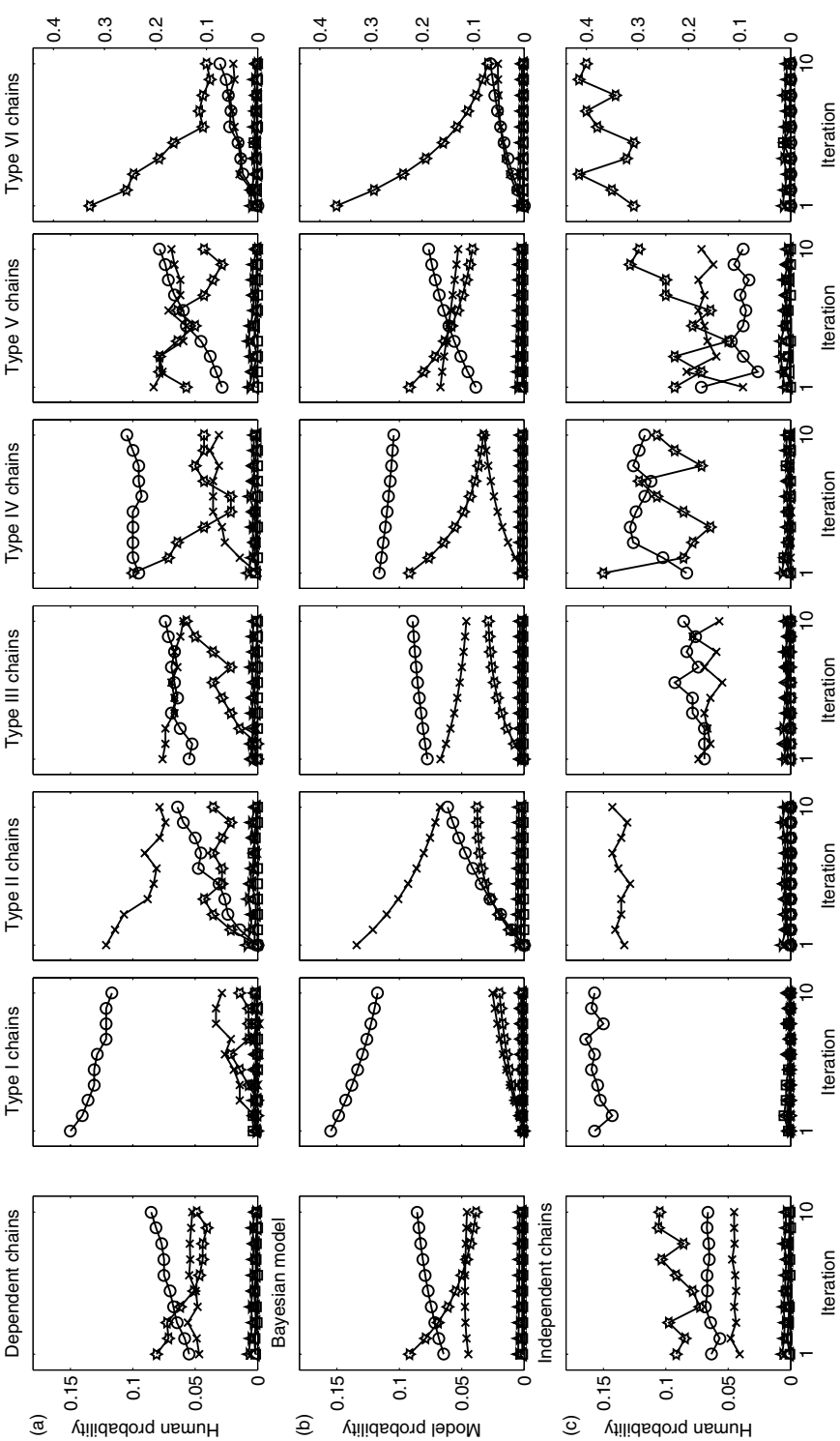

Fig. 6. Results of Experiment 1B, in which iterated learning was done between-subject with three positive examples, for (a) dependent chains, (b) the Bayesian model, and (c) independent chains. Each plot shows the probability that a participant selected a particular category structure of a given type as a function of the number of iterations of learning. Plot markers are those used in Fig. 5. The left-most plots collapse across all chains, whereas the six plots on the right break these data up based on the type of the category structure used to initialize the chain. The Bayesian model predicts response probabilities across all iterations of the independent chains that match the first iteration of the dependent chains. The vertical axis uses the same scale for all plots except that for Type VI chains.

shown in the last two rows of Table 1. Statistically significant differences were only observed in the ninth iteration for the dependent chains and the first iteration for the independent chains, but there were no systematic patterns of difference between the two designs.

8. Discussion of Experiments 1A and 1B

The results of Experiments 1A and 1B illustrate that experiments based on iterated learning can be used to explore the inductive biases of human learners. The correspondence between the proportion of trials on which a given hypothesis was selected and the estimated prior probability of that hypothesis increased with successive iterations for both the within- and between-subject designs, with Type I and Type II structures increasing in prevalence over the course of the experiment, and there were no systematic differences in the results obtained from using these two designs. These results suggest that we can use the simpler within-subjects design for further investigation of the consequences of iterated learning, and provide support for the claim that iterated learning should converge to the prior. However, Experiments 1A and 1B did not provide a strong test of this claim, as the dependent chains had not converged to their stationary distributions by the end of the experiments.

The slow convergence of the dependent chains is a consequence of the strong constraints that three positive examples place on consistent hypotheses. With only five consistent hypotheses for each set of three positive examples, there can only be slow drift in the hypotheses selected in a given chain. Although the rate of drift prevented the chains from converging to the prior in the 10 iterations included in the experiment, the results were paralleled by the predictions of the Bayesian model. Using the priors estimated from the human data, the second eigenvalues of the transition matrices for Experiment 1A and Experiment 1B were $\lambda_2 = 0.949$ and 0.948, respectively. These eigenvalues are both close to 1, indicating that the corresponding Markov chains will converge slowly.

Because faster convergence is important for iterated learning to be able to produce samples from the prior in a short experiment, Experiment 2 set out to manipulate the rate of convergence by varying the strength of the constraints provided by the data. Provision of different numbers of positive and mixed examples should affect the dynamics of iterated learning. In particular the Bayesian models outlined in Section 4 make precise predictions about the rate of convergence resulting from these different regimes. By demonstrating that we can manipulate the rate of convergence we have a further opportunity to test these predictions, and by showing that it is possible to construct chains that converge to their stationary distribution in the course of an experiment we can establish one of the prerequisites for using iterated learning as a method for studying inductive biases in the laboratory.

9. Experiment 2: Manipulating the rate of convergence

9.1. Method

9.1.1. Participants

There were a total of 238 participants, consisting of 171 undergraduates from the University of Louisiana, Lafayette participating in exchange for course credit, and 67 members of the

Brown University community who were compensated at a rate of $7 per hour. These participants were divided across three conditions: 117 (including 20 from Brown) saw two positive examples, 57 (including 26 from Brown) saw three mixed examples, and 64 (including 21 from Brown) saw four mixed examples.

9.1.2. Stimuli

Stimuli were the same as those in Experiment 1A, although varying the quantity of data provided to the participants required using slightly different displays. With two positive examples, the two examples were shown in a black box and participants selected from the 15 consistent hypotheses, as shown in Fig. 3b. A similar display was used with mixed examples, with negative examples being shown outside the box. Three mixed examples would yield either 5 or 10 consistent hypotheses, as illustrated in Fig. 3c, and four mixed examples would yield either 4 or 6 hypotheses, as illustrated in Fig. 3d. If all four examples were either positive or negative, only 1 consistent hypothesis existed. These trials were included in the chains for the purpose of our analyses, but participants were not shown the corresponding displays or required to respond to them.

9.1.3. Procedure

The procedure was the same as that of Experiment 1A, with the only modification being the number and nature of the examples (and hypotheses) shown to participants on each trial. The instructions in the two positive examples condition were identical to those used in Experiment 1A, replacing information regarding the number of examples and the number of hypotheses with 2 and 10, respectively. The instructions in the three mixed examples condition were similar, but the third paragraph was replaced with:

> The three amoebae you see will appear either inside or outside a black box. The amoebae inside the box are members of the species, and the amoebae outside the box are not members. Your task is to select your best guess at the identity of the entire species based on the membership of these three amoebae. The boxes shown on the bottom of the screen represent the possible species consistent with the sample members: the amoebae inside the box are in the species, the amoebae outside the box are out of the species. The number of boxes will be either five or ten, depending on the specific examples you see.

Participants in the four mixed examples condition received the same instructions, with "three" replaced by "four" and "five or ten" replaced by "four or six."

9.2. Results and Discussion

The variable of interest in this study was the proportion of trials on which people selected category structures of each type, with the key predictions being that these proportions would converge toward the prior as the number of iterations in the dependent chains increased, and that the rate of convergence would be affected by the strength of the constraints imposed by the data. As in Experiment 1A, the EM algorithm was used to find a maximum-likelihood estimate of the prior probability of category structures of the six different types and simultaneously classify participants as responding according to this prior or at random. The estimated probabilities of structures of Types I through VI were 0.111, 0.022, 0.003, 0.002, 0.004, and 0.017,

respectively, providing a close match to the estimates produced in Experiments 1A and 1B. Participants assigned a probability greater than 0.5 of responding at random were not used in the analyses appearing in the remainder of this section, but the data produced by these participants are analyzed further in the appendix. This left a total of 69 participants receiving two positive examples, 29 participants receiving three mixed examples, and 39 participants receiving four mixed examples. The higher rates of random responding in this experiment may be due to the increased difficulty of reasoning about larger sets of consistent hypotheses and the implications of both positive and negative examples, which many of the participants reported as being challenging during debriefing.[4]

Fig. 7 shows the proportion of trials on which a category structure of each of the six types was selected as a function of iteration for the three different conditions. In all three conditions, the probability of selecting a category structure of a given type in the final iteration closely matched the prior probability of that type in the final iteration, confirming the basic prediction of convergence to the prior. Fig. 7 also shows the predictions of the Bayesian model using the prior given earlier. These model predictions correlate extremely well with the human responses, giving $r(58) = 0.996$, 0.993, and 0.984 for two positive examples, three mixed examples, and four mixed examples, respectively.

The results of a series of chi-square tests comparing the distributions over types of category structures at each iteration of the dependent and independent chains are given in Table 2. There was a statistically significant difference at every iteration after the third for two positive examples, at iterations five and seven for three mixed examples, and at iterations four through nine for four random examples. The results for two positive examples and four mixed examples clearly indicate an effect of iterated learning, with the strong temporal trend seen for four random examples suggesting that the last non-significant result was simply due to chance. The weaker results for three mixed examples may be due to the combination of a relatively fast rate of convergence, meaning that the distribution after just one iteration (which is measured

Table 2
Results of Chi-Square Tests for Effects of Iterated Learning in Experiment 2

	Iteration									
Dependent Versus Independent	1	2	3	4	5	6	7	8	9	10
Two positive examples	12.40	8.09	16.51	22.48*	25.23*	20.68*	28.83*	36.03*	35.04*	46.86*
Three mixed examples	11.17	8.30	13.58	5.78	19.75*	12.79	25.34*	13.44	7.87	14.57
Four mixed examples	9.80	4.08	11.12	20.94*	21.01*	20.47*	35.38*	33.91*	33.59*	11.13

Note. Asterisks indicate statistical significance for χ^2 (5) at $p < .005$, using the Bonferroni correction to maintain an overall false positive rate below .05 for each sequence of tests. The weak pattern of significant differences for three mixed examples is due to rapid convergence and small sample size, as discussed in the text.

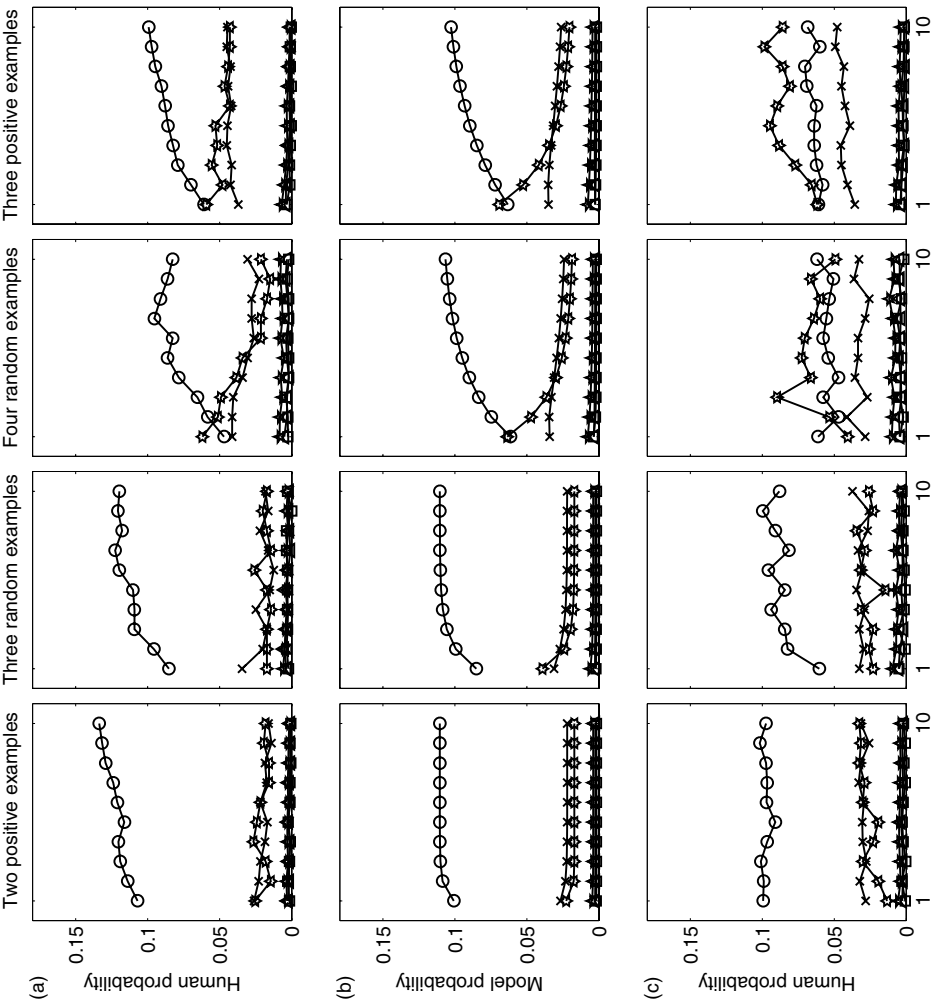

Fig. 7. Results of Experiment 2, in which iterated learning was run within-subjects and the type of data transmitted between iterations was varied, for (a) dependent chains, (b) the Bayesian model, and (c) independent chains. Each plot shows the probability that a participant selects a particular category structure of a given type as a function of the number of iterations of learning. Plot markers are those used in Fig. 5. All plots collapse across all chains. The Bayesian model predicts response probabilities across all iterations of the independent chains that match the first iteration of the dependent chains. The right-most panel reproduces the results of Experiment 1A, for comparison.

by the independent chains) will be close to the distribution after several iterations in the dependent chains, and the fact that this condition had fewer participants than any other.

For completeness, Figs. 8, 9, and 10 show a more detailed analysis of the data and model predictions, breaking down the chains by their initialization, as was done in Figs. 5 and 6 for Experiments 1A and 1B, respectively. In contrast to Experiments 1A and 1B, these individual chains show evidence of convergence, with the distribution over types in the final iteration converging to the same value regardless of initialization. Chi-square tests comparing the distribution over types of category structure in the final iteration across the six dependent chains initialized with different types showed no statistically significant difference for any of the three conditions, giving $\chi^2(25) = 25.87$, $p = .41$, for two positive examples; $\chi^2(25) = 24.19$, $p = .51$, for three mixed examples; and $\chi^2(25) = 31.65$, $p = .17$ for four mixed examples. The distribution to which the chains converge is the prior, with Type I structures increasing in prevalence while Type VI structures decrease. The model and data still exhibit a strong correlation under this more detailed analysis, with $r(358) = 0.986$, 0.954, and 0.953, respectively.

The qualitative prediction that the strength of the constraints on hypotheses produced by the data should affect the rate of convergence is borne out by these results. Two positive examples provide only weak constraints on possible hypotheses, and convergence to the prior is rapid. Three mixed examples produce stronger constraints and slower convergence, and convergence is slowest with the strong constraints provided by four mixed examples. This trend also matches the quantitative predictions of the Bayesian model. The second eigenvalues of the transition matrices produced by using the prior estimated from people's responses are $\lambda_2 = 0.508$, 0.702, and 0.893 for these three conditions. For comparison, Fig. 6 reproduces the results of Experiment 1A in the same format as the results of the three conditions from this experiment. The model predictions shown for these results using the prior estimated from the three conditions of Experiment 2, providing a single set of parameters that accounts for the entire set of results and reinforcing the consistency of these parameters across experiments. A total of 66 participants were classified as responding non-randomly, yielding an overall correlation of $r(58) = 0.961$ between the model predictions and the data. The second eigenvalue of the transition matrix of the Bayesian model using this prior was $\lambda_2 = 0.936$, confirming that convergence to the prior should be slowest when participants are provided with three positive examples.

10. General discussion

The results of our experiments demonstrate that iterated learning can be used to investigate the inductive biases of human learners. In each experiment the distribution over category structures selected by people converged toward an equilibrium broadly consistent with previous research on the difficulty of learning and remembering category structures (Feldman, 2000; Nosofsky et al., 1994; Shepard et al., 1961), with Type I structures being most prevalent, followed by Type II, and then the other four types, and matching the prior estimated using our Bayesian model. Experiments 1A and 1B demonstrated that this effect persists whether iterated learning occurs within-subjects, as is more convenient in a laboratory setting, or

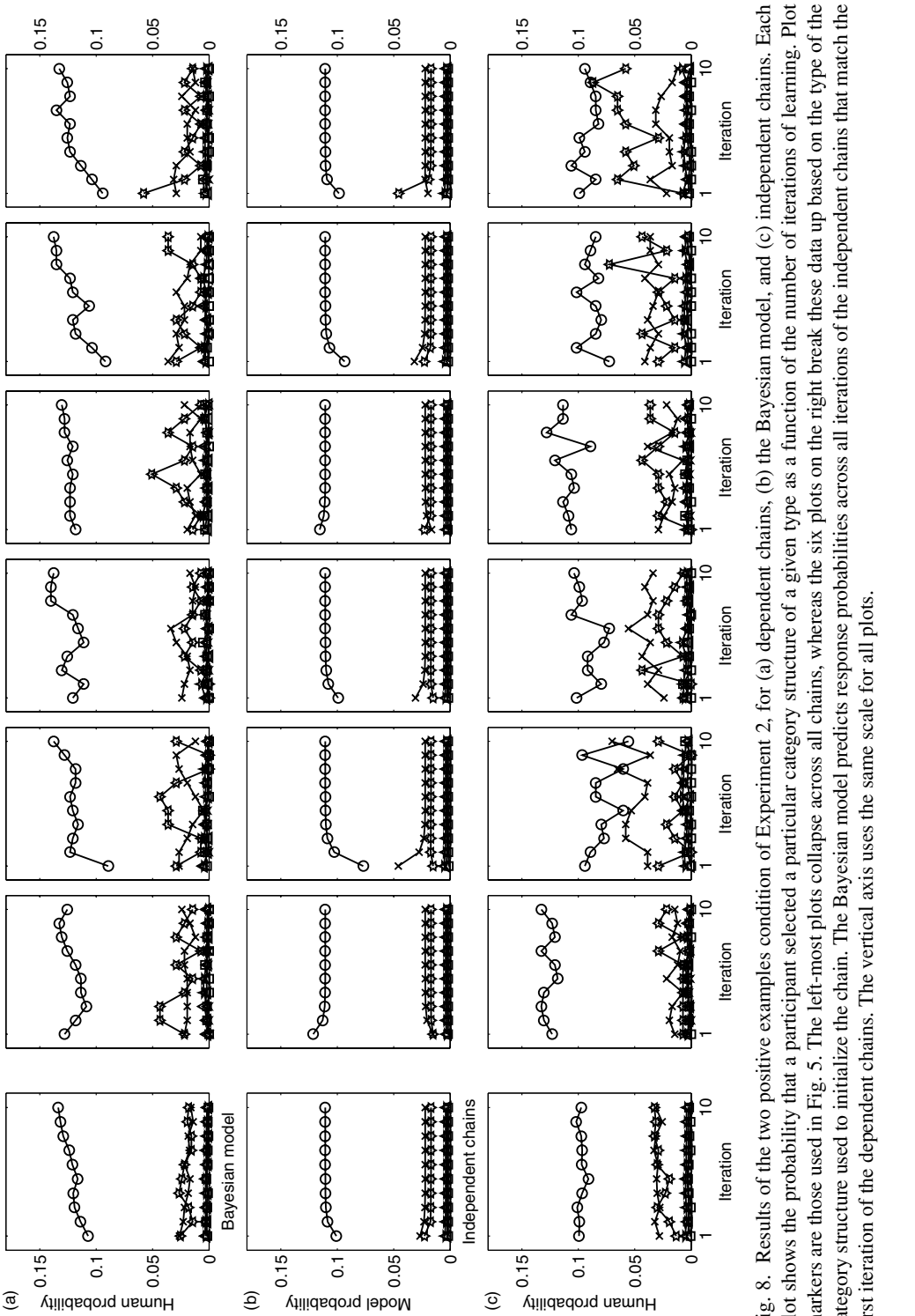

Fig. 8. Results of the two positive examples condition of Experiment 2, for (a) dependent chains, (b) the Bayesian model, and (c) independent chains. Each plot shows the probability that a participant selected a particular category structure of a given type as a function of the number of iterations of learning. Plot markers are those used in Fig. 5. The left-most plots collapse across all chains, whereas the six plots on the right break these data up based on the type of the category structure used to initialize the chain. The Bayesian model predicts response probabilities across all iterations of the independent chains that match the first iteration of the dependent chains. The vertical axis uses the same scale for all plots.

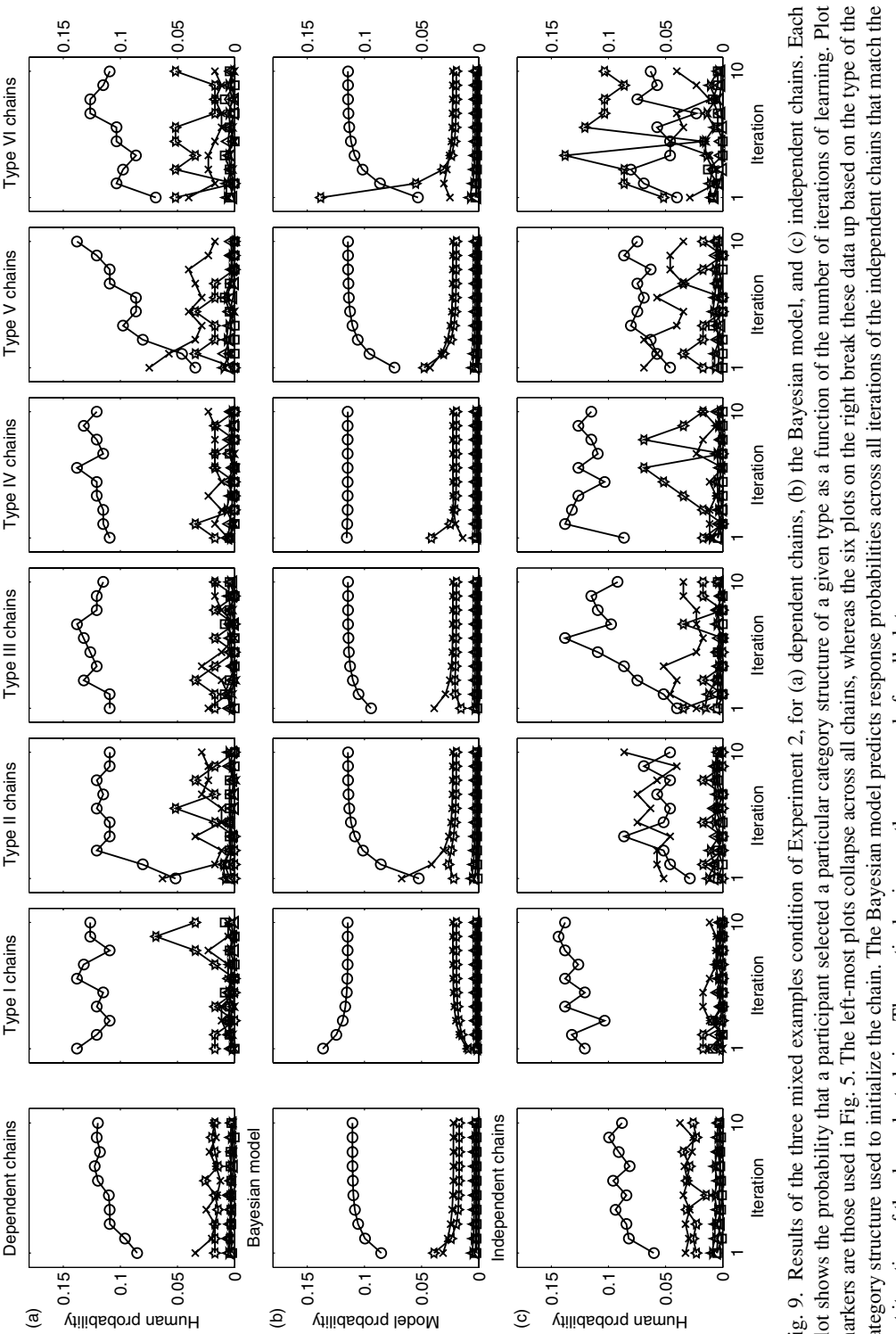

Fig. 9. Results of the three mixed examples condition of Experiment 2, for (a) dependent chains, (b) the Bayesian model, and (c) independent chains. Each plot shows the probability that a participant selected a particular category structure of a given type as a function of the number of iterations of learning. Plot markers are those used in Fig. 5. The left-most plots collapse across all chains, whereas the six plots on the right break these data up based on the type of the category structure used to initialize the chain. The Bayesian model predicts response probabilities across all iterations of the independent chains that match the first iteration of the dependent chains. The vertical axis uses the same scale for all plots.

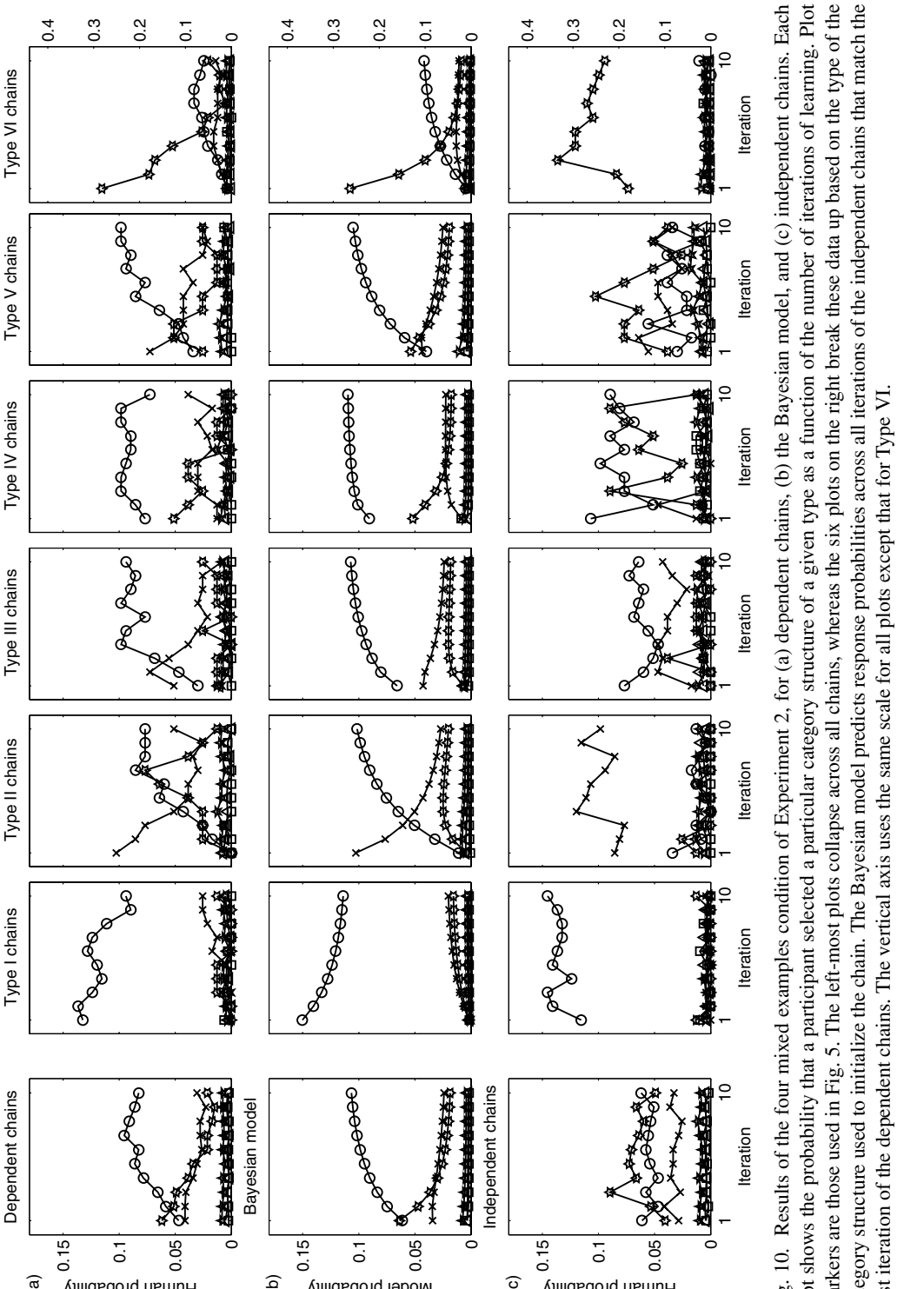

Fig. 10. Results of the four mixed examples condition of Experiment 2, for (a) dependent chains, (b) the Bayesian model, and (c) independent chains. Each plot shows the probability that a participant selected a particular category structure of a given type as a function of the number of iterations of learning. Plot markers are those used in Fig. 5. The left-most plots collapse across all chains, whereas the six plots on the right break these data up based on the type of the category structure used to initialize the chain. The Bayesian model predicts response probabilities across all iterations of the independent chains that match the first iteration of the dependent chains. The vertical axis uses the same scale for all plots except that for Type VI.

between-subjects, as used in previous work and in formal analyses of language evolution. The dynamics of iterated learning, represented by changes in the distribution over category structures at each iteration, were strongly in accord with the predictions of our Bayesian model. In particular, Experiment 2 showed that the rate of convergence was reduced when the data provided to participants placed stronger constraints on hypotheses. In the remainder of the article, we discuss the implications of these results for understanding human inductive biases for category structures, the robustness of our formal analyses to different assumptions about the way in which people select their responses, and the circumstances under which iterated learning will be a particularly valuable tool for estimating the priors of human learners.

10.1. Priors and category structures

Although the prior distribution on category structures estimated from the responses of our participants and revealed as the outcome of iterated learning is generally consistent with the results of previous studies, there is one notable difference: the persistently high probability of Type VI category structures. The previous research mentioned earlier suggested that Type VI structures are hardest to learn, but the prior that seemed to characterize the inferences of participants in our experiments gave these structures higher probability than Types III through V. One possible explanation for this difference is the lack of memory demands in our task. The task used in our experiments was explicitly intended to remove memory demands, allowing us to study the inductive biases involved in generalization directly. However, most research on category learning uses a task that has a memory component, training participants on the category labels of a set of objects and then asking them to rely on their memory for these data when guessing the labels for new objects. The experiments that suggest Type VI structures are hard to learn required participants to remember a set of examples from the category, whereas in our experiments participants could see both the examples and the full set of consistent category structures. Type VI structures actually have far greater symmetry and simplicity than Types III through V, being describable as the structures for which every two members have the same value on exactly one dimension. These properties are quite apparent in the sample species shown in Fig. 2. Our presentation format and task could thus have resulted in a stronger preference for Type VI.

The inductive biases of human learners solving particular problems will reflect the cumulative biases inherent in representing and remembering stimuli as well as forming generalizations. We believe that our experiments provide an accurate characterization of the inductive biases of human learners in a pure generalization task, with the problems that people typically experience with Type VI being a consequence of memory demands. This has interesting implications for the role of complexity in memory and generalization. Feldman (2000) has argued that the relative difficulty of learning different types of category stuctures can be explained by simple measures of the logical complexity of those structures. If we can tease apart the relative contributions of generalization and memory to human performance, we can identify the importance of quantities such as logical complexity in each. Based on the results of our experiments, we anticipate that iterated learning may provide a useful paradigm for further investigation of the effects of memory demands on human inductive biases.

10.2. Robustness to different response selection strategies

Our formal analysis of the consequences of iterated learning assumes that learners sample hypotheses from their posterior distribution. Without this "probability matching" assumption, general results for the consequences of iterated learning are hard to find (although see Griffiths & Kalish, 2007, and Kirby et al., 2007, for some exceptions). Our expectation that this assumption will hold for human learners is based on the fact that probability matching of various forms is often found in human learning (Myers, 1976; Vulkan, 2000), and is commonly assumed both in Bayesian models of cognition (e.g., Anderson, 1990; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) and in a variety of models of categorization and choice behavior in the form of Luce's (1959) choice rule (Ashby, 1992; Ashby & Alfonso-Reese, 1995; Ashby & Maddox, 1993; Kruschke, 1992; Nosofsky, 1986, 1987). However, one additional advantage of the generalization task that was used in our experiments is that it corresponds to a special case in which the predictions of our framework are robust to variation in the response selection strategy used by participants.

A common approach to modeling response strategies that are more systematic than probability matching is to use an exponentiated version of the Luce choice rule, taking the probability of selecting hypothesis h to be proportional to $P(h|d)^\gamma$, where γ is a free parameter (e.g., Ashby & Maddox, 1993; Kruschke, 1992). This reduces to probability matching when $\gamma = 1$, random responding when $\gamma = 0$, and deterministically selecting the hypothesis with highest posterior probability as $\gamma \to \infty$. Intermediate values of γ capture intermediate response strategies. In the case of our generalization task, use of an exponentiated Luce choice rule would mean selecting hypotheses with probability proportional to the exponentiated versions of the posterior distributions defined by Equations 5 and 7. However, because these posterior distributions both correspond to the prior, $P(h)$, normalized over the set of consistent hypotheses, use of the exponentiated choice rule with parameter γ is equivalent to probability matching with a prior that is proportional to $P(h)^\gamma$. Consequently, iterated learning will converge to the distribution over hypotheses proportional to $P(h)^\gamma$.

The analysis outlined in the previous paragraph implies that the prior on category structures estimated from the responses of participants in our experiments and the stationary distribution of iterated learning will be somewhat robust to variation in response selection strategies that can be captured using an exponentiated Luce choice rule. Specifically, although we cannot know the actual prior probability of a particular hypothesis without knowing γ, the ordering of the prior probabilities of the different hypotheses will be the same for any choice of $\gamma \in (0, \infty)$ because this ordering is preserved under exponentiation. Thus, our conclusions about the relative prior probabilities of category structures of different types should hold, even if participants are not selecting hypotheses by probability matching. A more detailed analysis of the class of models to which this kind of analysis applies and its implications for language evolution is given in Kirby et al. (2007).

10.3. When will iterated learning be useful?

Iterated learning is a new way to measure the inductive biases of human learners. However, previous work has established at least two ways to explore these biases: using computational

models and conducting experiments measuring the difficulty with which people learn or remember different hypotheses. Both computational models and experiments have been used to study inductive biases for the category structures we used in our experiments (Feldman, 2000; Nosofsky et al., 1994; Shepard et al., 1961), which was one of the reasons we choose to use these stimuli to demonstrate the consequences of iterated learning. For iterated learning to be useful more generally, there need to be circumstances under which it provides a more effective means of investigating inductive biases than either computational modeling or traditional experimental methods.

As mentioned in the introduction, the main advantage of iterated learning over estimating inductive biases using computational models is that it is requires only weak assumptions about the form of those biases. It is not necessary to define a model or specify a prior to use iterated learning to investigate inductive biases, and our formal analysis of iterated learning indicates that it will converge to the prior under very general conditions.[5] This means that iterated learning can be used to explore inductive biases in contexts where researchers have very little idea what form those biases might take, or with stimuli that are so complex that defining a computational model is difficult. Iterated learning is thus valuable as an exploratory tool, providing information in contexts where precise predictions are hard to formulate. It can also be used to test computational models: If a model assumes a particular inductive bias (or asserts that people use a particular prior), then that bias should be produced as a result of iterated learning.

The difference between traditional uses of computational models in investigating inductive biases and iterated learning has an analogue in approaches to density estimation in statistics. Density estimation is the problem of determining the probability distribution that best characterizes observed data. Statisticians have explored a variety of solutions to this problem, but one of the basic distinctions between these approaches is between "parametric" and "nonparametric" methods (e.g., Silverman, 1986). Parametric methods assume that the distribution takes a particular form, such as a Gaussian, and then identify the distribution by estimating its parameters, such as its mean and variance. Nonparametric methods make very weak assumptions about the form of the distribution, defining procedures that work for estimating probability distributions of any form, given enough data. Parametric density estimation works well with small quantities of data, but can lead to a poor estimate if the true distribution takes a form other than that assumed (e.g., a Gaussian will poorly characterize a bimodal distribution). Nonparametric density estimation requires more data, but will ultimately yield a good estimate of any distribution.

Determining the priors of human learners is a form of density estimation: The goal is to estimate a probability distribution. Defining a computational model and fitting the parameters of that model to the data defines a class of possible priors and then finds the distribution within that class that best fits the data, and is thus a parametric approach. This will work well in cases where priors are simple or modelers know a lot about the form of those priors, as for the category structures explored in this article. In these cases, it is possible to specify a class of distributions that will come close to the true distribution. By contrast, iterated learning is a nonparametric method: It will yield samples from the prior regardless of the form of that prior, and those samples can be used to construct estimates of the corresponding distribution that make only weak assumptions about the nature of human inductive biases. Thus, in cases

where priors are complex or little is known, iterated learning will be a valuable alternative to traditional methods, allowing people's inductive biases to reveal themselves directly.

Other experimental methods exist for investigating inductive biases. These methods typically aim to establish the difficulty of learning or remembering particular hypotheses. The resulting experiments require a set of hypotheses to be investigated, and the hypotheses are either exhaustively enumerated from some restricted class (as is the case for the category structures used in our experiments) or chosen to test a theoretical principle. Iterated learning provides information about inductive biases in a way that does not require gathering information about every hypothesis under consideration. Once iterated learning converges, the hypotheses produced are sampled in proportion to their prior probability. Thus, the frequency with which the hypotheses appear are informative about their prior probability, and hypotheses with low prior probability might not appear at all. As a consequence, iterated learning can be used in contexts where the set of hypotheses is difficult to enumerate or there are no theories to guide selection. For example, it can be used to explore inductive biases in very large sets of hypotheses (such as all 2^{2k} categories defined on k binary dimensions), and will be efficient in doing so as long as only a small number of those hypotheses have high prior probability.

Again, a comparison to statistical methods can provide insight into the value of iterated learning as an experimental method for investigating inductive biases. In this case, the comparison is to the Monte Carlo method for evaluating expectations of functions of a random variable. Given a probability distribution $p(x)$ and a function $f(x)$, the expectation of $f(x)$ with respect to $p(x)$ is

$$E_{p(x)}[f(x)] = \sum_{x \in \mathcal{X}} f(x)p(x) \tag{8}$$

where $\mathcal{X}$ is the set of all values of x. Normally, evaluating this expectation would require evaluating $f(x)$ and $p(x)$ at all values of x, which could take a long time if $\mathcal{X}$ is large. The Monte Carlo method approximates the expectation as

$$E_{p(x)}[f(x)] \approx \frac{1}{m} \sum_{i=1}^{m} f(x^{(i)}) \tag{9}$$

where $x^{(i)}$ is the ith sample from the probability distribution $p(x)$. Because sampling produces values of x for which $p(x)$ is high with high probability, accurate approximations can be obtained with only a few samples. In cases where $\mathcal{X}$ is large and $p(x)$ is high for only a few values of x, sampling can be more efficient than exhaustively enumerating all values of x. Likewise, iterated learning provides samples of hypotheses that have high prior probability, providing information about which hypotheses people favor without requiring enumeration of the entire set of possibilities, and will be more efficient than traditional experimental methods when exploring strong inductive biases expressed in large sets of hypotheses.

10.4. Conclusion

Our formal analyses suggested that iterated learning might be a possible method for identifying human inductive biases in the laboratory. By using a task based on a small and well-studied

set of category structures, we have been able to show that iterated learning converges toward a distribution on category structures that reflects people's inductive biases, and to make precise predictions about the dynamics of iterated learning that provide a good characterization of the human data. Our use of this task was motivated by the existence of a literature characterizing people's inductive biases for different category structures, allowing us to evaluate the correspondence between the outcome of iterated learning and the results presented in this literature. However, our results open the door to future research investigating people's inductive biases in contexts where they remain unknown. By reproducing iterated learning in the laboratory, we can begin to map out the implicit biases that are at the heart of the remarkable human ability to solve inductive problems.

Notes

1. More formally, the total variation distance between $P(h_n)$ and the prior $P(h)$, defined as $\frac{1}{2}\sum_i |P(h_n = i) - P(h = i)|$, is upper bounded by a constant multiplied by λ_2^n (Rosenthal, 1995). This upper bound thus decreases geometrically in n, with the rate being determined by λ_2.
2. This assumption is not necessary for either positive or mixed examples, as any likelihood function that is constant for all hypotheses consistent with the set of observed examples will yield the same result. For example, both the "strong sampling" and "weak sampling" schemes for determining the likelihood outlined in Tenenbaum and Griffiths (2001) yield the same posterior distribution in this case.
3. These are the prior probabilities of individual structures of the six types, rather than the probability assigned to each type in the prior; thus, they do not sum to 1. The latter probability can be obtained by multiplying these probabilities by the number of structures of each of the six types, being 6, 6, 24, 8, 24, and 2, respectively.
4. Consistent with this interpretation, the rate of random responding was far lower among the participants from Brown University, with only 11 participants from Brown being eliminated across all experiments reported in this article. All participants from Brown had specifically volunteered for the experiments and were reimbursed at an hourly rate, thus being significantly more motivated to spend time on the task than the other participants, who were fulfilling a course requirement and received a fixed amount of credit for completing the experiment.
5. Convergence to the prior requires that the underlying Markov chain be *ergodic*. In this case, the main requirement is that the prior and likelihood are such that there is a nonzero probability of starting at one hypothesis and ending at another after some finite number of iterations for every pair of hypotheses.

Acknowledgments

This work was supported by Grants 0704034 and 0544705 from the National Science Foundation (to Thomas L. Griffiths and Michael L. Kalish, respectively), and a grant from the Louisiana Board of Regents to Michael L. Kalish. A shorter version of this article was presented at the 28th Annual Conference of the Cognitive Science Society. The data appearing

in this article were collected while Thomas L. Griffiths and Brian R. Christian were at Brown University. We thank Anu Asnaani, Joe Austerweil, Charles Barousse, Rebecca Cremona, Alana Firl, Vikash Mansinghka, Laurie Robinette, and Diana Tamir for discussions about this project and assistance in running experiments.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

Ashby, F. G. (1992). *Multidimensional models of perception and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39*, 216–233.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372–400.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science, 24*, 259–260.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Decision Processes, 11*, 1–27.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34*, 71–105.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input–output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review, 103*, 566–581.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10*, 335–344.

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective*. Cambridge, MA: MIT Press.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 673–680). Sydney, Australia: Association for Computational Linguistics.

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

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 827–832). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Griffiths, T. L., & Kalish, M. L. (2007). A Bayesian view of language evolution by iterated learning. *Cognitive Science, 31*, 441–480.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (in press). Categorization as nonparametric Bayesian density estimation. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind*. Oxford, England: Oxford University Press.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford, England: Oxford University Press.

Hurwicz, L. (1973). The design of mechanisms for resource allocation. *American Economic Review, 63*, 1–30.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review, 14*, 288–294.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review, 111*, 1072–1099.

Kearns, M., & Vazirani, U. (1994). *An introduction to computational learning theory.* Cambridge, MA: MIT Press.

Kemp, C., & Tenenbaum, J. B. (2003). Theory-based induction. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 658–663). Mahwah, NJ: Lawrence Erlbaum Associates.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation, 5*, 102–110.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences, 104*, 5241–5245.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, 605–621.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*, 232–257.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.

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

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Norris, J. R. (1997). *Markov chains.* Cambridge, England: Cambridge University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 87–108.

Nosofsky, R. M., Gluck, M., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*, 352–369.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition.* Oxford, England: Oxford University Press.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 393–407.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*, 665–681.

Rosenthal, J. S. (1995). Convergence rates of Markov chains. *SIAM Review, 37*, 387–405.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press.

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

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13, Whole No. 517).

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.
Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life, 9*, 371–386.
Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453–489.
Strang, G. (1988). *Linear algebra and its applications* (3rd ed.). Philadelphia: Saunders.
Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning*. Unpublished doctoral dissertation, Massachussets Institute of Technology, Cambridge, MA.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–641.
Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 2277–2282). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.
Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance, 16*, 8–27.
Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*, 101–118.

Appendix: The Expectation–Maximization (EM) Algorithm for Detecting Chance Responding

Let θ denote the parameters of the prior, being the probability of category structures of Types I through VI, respectively. Assume that some proportion of our participants respond in a way consistent with θ, whereas the remainder respond at random, and let π denote the probability that a participant is "real" as opposed to "random." Random responding can be modeled in our Bayesian framework by using a prior that is uniform over all category structures (i.e., the probability of a single category structure of any type is 1/70). Our goal is to estimate θ and π, and to identify the "real" participants.

Given knowledge of θ and π, deciding whether a participant is real is straightforward. Letting $z_i = 1$ indicate that the ith participant is real and $z_i = 0$ indicate random responding, we can use Bayes's rule to compute

$$P(z_i = 1|x_i, \theta, \pi) = \frac{P(x_i|z_i = 1, \theta)P(z_i = 1|\pi)}{P(x_i|z_i = 1, \theta)P(z_i = 1|\pi) + P(x_i|z_i = 0)P(z_i = 0|\pi)} \tag{10}$$

$$= \frac{P(x_i|z_i = 1, \theta)\pi}{P(x_i|z_i = 1, \theta)\pi + P(x_i|z_i = 0)(1 - \pi)} \tag{11}$$

where x_i are the responses of the ith participant, and we have simplified $P(x_i|z_i = 1, \theta)$ and $P(z_i = 1|\pi)$ to reflect conditional independence from π and θ, respectively. The probabilities $P(x_i|z_i = 1)$ and $P(x_i|z_i = 0)$ can be computed by assuming that participants sample from their posterior distribution for each decision, computing the posterior via Equations 5 or 7 using either the prior defined by θ or a uniform prior. Given knowledge of z_i for all participants, estimating θ and π _is also straightforward. The maximum-likelihood estimate of π is just

$$\hat{\pi} = \frac{\sum_{i=1}^{n} I(z_i = 1)}{n} \tag{12}$$

where n is the total number of participants and $I(\cdot)$ takes the value 1 when its argument is true and 0 otherwise. The maximum-likelihood estimate of θ is

$$\hat{\theta} = \arg\max_{\theta} \sum_{i|z_i=1} \log P(x_i|z_i = 1, \theta) \tag{13}$$

being the value that maximizes the log-likelihood of the data produced by the participants for whom $z_i = 1$.

Unfortunately, we lack knowledge of both θ and π and the values of z_i. The Expectation–Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is an algorithm for maximum-likelihood parameter estimation in such circumstances, addressing the difficult problem of estimating θ, π, and z_i simultaneously by alternating between solving the two simpler problems outlined in the previous paragraphs. First, we initialize θ and π to some arbitrary values. Then, in the Expectation (E) step, we compute $P(z_i = 1|x_i, \theta, \pi)$ based on these values of θ and π. In the Maximization (M) step, we select new values of θ and π that maximize the expectation of the log-likelihood with respect to this distribution on the z_i. In our case, this means using the estimates

$$\hat{\pi} = \frac{\sum_{i=1}^{n} P(z_i = 1|x_i, \theta, \pi)}{n} \tag{14}$$

and

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} P(z_i = 1|x_i, \theta, \pi) \log P(x_i|z_i = 1, \theta) \tag{15}$$

where the q and p appearing in $P(z_i = 1|x_i, \theta, \pi)$ in both equations refer to the old values of those parameters. We then complete an E step with these new values of θ and π, apply an M step, and iterate until the log-likelihood of the data, , converges to a fixed value.

The basic idea of assuming that participants fall into groups associated with different probability distributions over responses and estimating the parameters of those groups using an iterative procedure such as the EM algorithm has previously been used for detecting different strategies used by a set of participants (e.g., Lee & Webb, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). In our case, it proved an extremely effective procedure for separating unmotivated participants from those engaged by the task. The efficacy of this method can be gauged by inspecting the responses of the participants classified as responding at random (i.e., for whom $P(z_i = 0|x_i, \theta, \pi) > 0.5$). Fig. A1 shows the proportion of trials on which these participants selected category structures of the six different types as a function of the number of iterations, together with Bayesian models using uniform priors (the constraints that the data impose on the hypotheses that learners can select mean that some structure can persist in their responses, so the Bayesian models are required to evaluate what random responding would look like). In all cases, the probability of selecting a category structure converges to a distribution in which there is very little effect of its type, particularly in comparison with the responses of the corresponding "real" participants shown in Fig. 7.

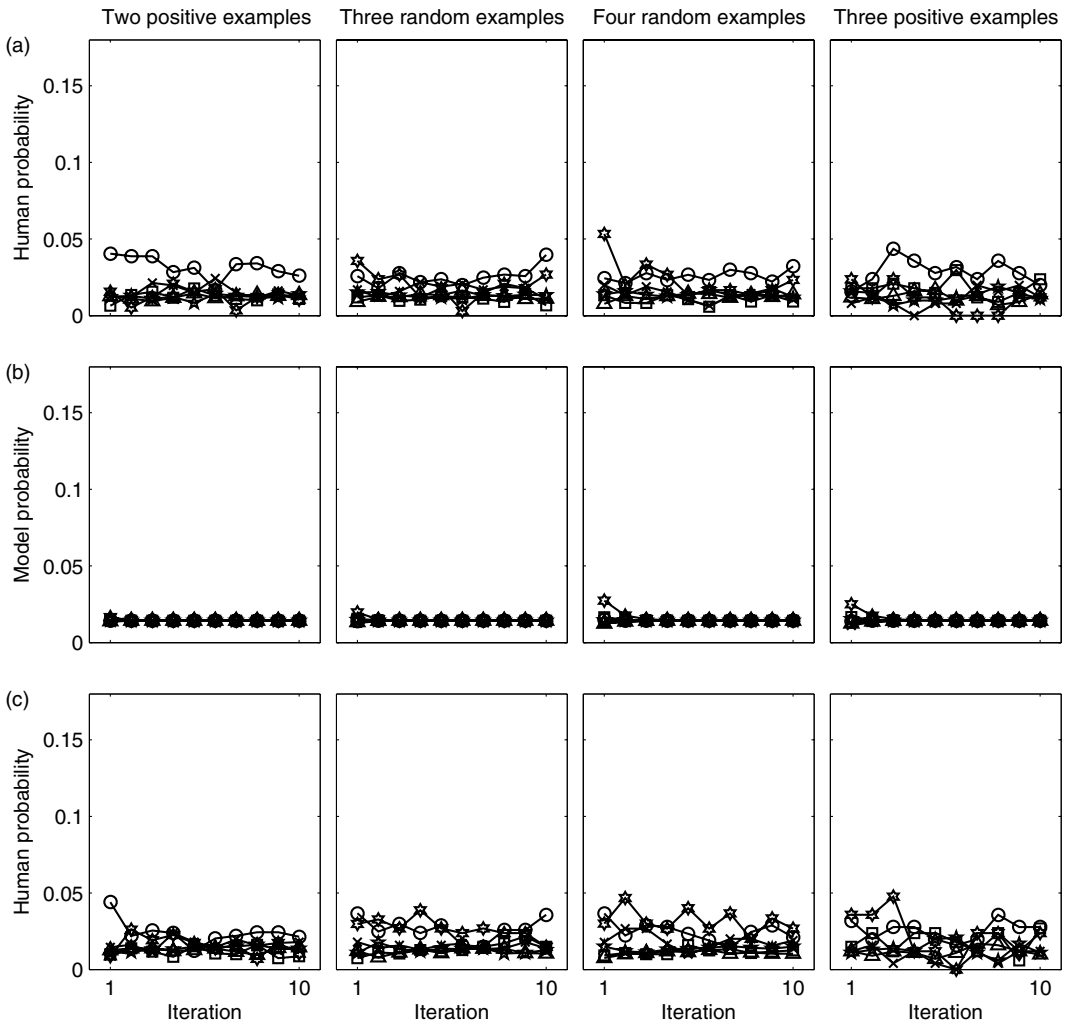

Fig. A1. Responses of participants classified as responding at random in Experiments 1A and 2, for (a) dependent chains, (b) the Bayesian model, and (c) independent chains. Each plot shows the probability that a participant selects a particular category structure of a given type as a function of the number of iterations of learning. Plot markers are those used in Fig. 5. All plots collapse across all chains. The Bayesian model predicts response probabilities across all iterations of the independent chains that match the first iteration of the dependent chains. In all cases, the probability of selecting a particular category structure at the end of the experiment depends only weakly on the type of that structure, consistent with random responding.