Psychon Bull Rev
DOI 10.3758/s13423-013-0556-3

BRIEF REPORT

Revealing human inductive biases for category learning by simulating cultural transmission

Kevin R. Canini · Thomas L. Griffiths · Wolf Vanpaemel · Michael L. Kalish

Abstract We explored people's inductive biases in category learning—that is, the factors that make learning category structures easy or hard—using iterated learning. This method uses the responses of one participant to train the next, simulating cultural transmission and converging on category structures that people find easy to learn. We applied this method to four different stimulus sets, varying in the identifiability of their underlying dimensions. The results of iterated learning provide an unusually clear picture of people's inductive biases. The category structures that emerge often correspond to a linear boundary on a single dimension, when such a dimension can be identified. However, other kinds of category structures also appear, depending on the nature of the stimuli. The results from this single experiment are consistent with previous empirical findings that were gleaned from decades of research into human category learning.

Keywords Category learning · Bayesian modeling · Mathematical models

Electronic supplementary material The online version of this article (doi:10.3758/s13423-013-0556-3) contains supplementary material, which is available to authorized users.

K. R. Canini
Computer Science Division, University of California, Berkeley, CA, USA

T. L. Griffiths (✉)
Department of Psychology, University of California, 3210 Tolman Hall #1650, Berkeley, CA 94720-1650, USA
e-mail: tom_griffiths@berkeley.edu

W. Vanpaemel
Faculty of Psychology and Educational Sciences, University of Leuven, Leuven, Belgium

M. L. Kalish
Institute of Cognitive Science, University of Louisiana, Lafayette, LA, USA

The ability to learn new categories from labeled examples is a basic component of human cognition, and one of the earliest to be studied by psychologists (Hull, 1920). As with many cognitive tasks, category learning can be characterized as a form of induction—an inference to underdetermined hypotheses from limited data (Bruner, Goodnow, & Austin, 1956). This directly implies a role for what machine learning researchers refer to as the *inductive biases* of a learner—those factors that make a learner more likely to entertain one hypothesis than another (Mitchell, 1997). In category learning, these inductive biases determine whether a particular category structure is easy or hard to learn.

Inductive biases for category learning have often been studied using "supervised" learning tasks, in which participants learn experimenter-designed categories in order to study the rates at which they learn and how they generalize (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Nosofsky, 1986, 1987; Nosofsky, Palmeri, & McKinley, 1994). However, this is an inefficient way to study human inductive biases. Establishing what is easy or hard to learn requires teaching people many different kinds of categories and comparing the results (e.g., Shepard, Hovland, & Jenkins, 1961). Over the last few decades, a clear picture has emerged from pursuing this method, but it has taken many experiments to produce this picture.

An alternative approach has been to use an "unsupervised" learning task, in which participants are asked to organize a set of stimuli into categories by themselves in order to examine which structures they naturally identify (e.g., Ahn & Medin, 1992; Anderson, 1991; Handel & Imai, 1972; Imai & Garner, 1965; Love, Medin, & Gureckis, 2004; Medin, Wattenmaker, & Hampson, 1987; Milton & Wills, 2004; Pothos & Chater,

Published online: 07 January 2014

2002; Pothos et al., 2011; Regehr & Brooks, 1995). Unsupervised learning clearly reflects people's inductive biases, but arguably taps different cognitive processes than does supervised learning (Milton & Wills, 2004; Regehr & Brooks, 1995).

For this article, we used a new approach to investigate human category learning, in which participants performed a standard supervised learning task, but we manipulated the categories that they learned so as to directly reveal their inductive biases using a single, compact experiment. The novel method that we used is known as *iterated learning* and has its roots in accounts of language evolution (Kirby, 2001). Participants are arranged into a chain in which the responses from the first participant are used as training data for the second participant, and so on. The result is a simple simulation of the cultural transmission of information that can be conducted in the laboratory. Mathematical analyses of this process have shown that as the chain gets longer, the responses that people produce come to reflect their inductive biases (Griffiths & Kalish, 2007). Intuitively, each time that people learn, they impose their inductive biases on the data, so iterating this process converges on responses that just express those biases. Experiments using this method have confirmed that the results are consistent with those of more traditional methods for estimating human inductive biases (Griffiths, Christian, & Kalish, 2008; Kalish, Griffiths, & Lewandowsky, 2007; Lewandowsky, Griffiths, & Kalish, 2009). Iterated learning can be seen as striking a middle ground between unsupervised and supervised learning, since the feedback that is provided comes from other participants.

Supervised and unsupervised learning have been used to explore several basic questions about human inductive biases for category learning: Does category learning differ when the dimensions used to define the stimuli are separable (i.e., clearly individuated; Garner & Felfoldy, 1970) or integral (Handel & Imai, 1972; Nosofksy, 1987)? Are boundaries defined along a single dimension easier to learn than those defined using multiple dimensions (Ahn & Medin, 1992; Ashby et al., 1998; Imai & Garner, 1965; Kruschke, 1993; Medin et al., 1987)? Do people form abstract prototypes, remember specific exemplars, or learn rules (Ashby & Gott, 1988; Erickson & Kruschke, 1998; Nosofsky, 1986; Nosofsky et al., 1994; Reed, 1972)? We used the iterated-learning approach to conduct a single experiment that generated a clear picture of human inductive biases for category learning, providing new insight into these basic questions.

Method

Participants

A total of 1,440 participants were included in the experiment, and a further 570 were excluded for reasons explained below. Of the included participants, 480 were students at the University of California, Berkeley, who received course credit, and 960 were recruited from Amazon Mechanical Turk (www.mturk.com, restricted to participants from the United States with a 95 % approval rating or greater) and received a payment of $0.50. The experiment had 16 conditions, resulting from the combination of four stimulus sets and four initial category structures. Each condition was replicated with six chains, two of which were made up solely of Berkeley participants, and four of which were made up solely of Mechanical Turk participants. Each replication of each condition consisted of an iterated learning chain of 15 generations. Each participant was randomly assigned to an incomplete chain in their pool (either Berkeley or Mechanical Turk), occupying the next available generation in the chain.

Stimuli

Four sets of stimuli were used, two with separable dimensions and two with integral dimensions (see Fig. 1). The stimuli with separable dimensions were "Shepard circles"—circles of varying diameters, each with a radius drawn at a varying angle (Shepard, 1964)—and rectangles that varied in height and width (Krantz & Tversky, 1975). The stimuli with integral dimensions were both sets of amorphous blobs, one from Cortese and Dyre (1996), which we call "Cortese blobs," and the other from Shepard and Cermak (1973), which we call "Shepard blobs." The two dimensions used to define these stimuli correspond to the parameters of periodic functions that are converted to closed loops and are not easily identified from the stimuli themselves.

For each stimulus set, we constructed an equal-spaced, 8-by-8 square grid of stimuli. The grid of 64 stimuli used for each type of stimulus set is depicted in the second and third rows of Fig. 1. We refer to the position of a stimulus in this 8-by-8 grid as the *canonical coordinates* of that stimulus. We also derived the multidimensional scaling (MDS; Kruskal, 1964) coordinates, shown in the fourth row of Fig. 1, which depict the ways that the stimuli are represented in psychological space. These coordinates were computed using similarity ratings provided by a different group of participants (details of this analysis appear in the supplementary materials).

Procedure

Each participant completed a training session and a test session. In the training session, the participant was trained to reproduce the category memberships of a random selection of 32 of the 64 stimuli. In each training trial, the participant classified a single stimulus from the training set, with feedback.

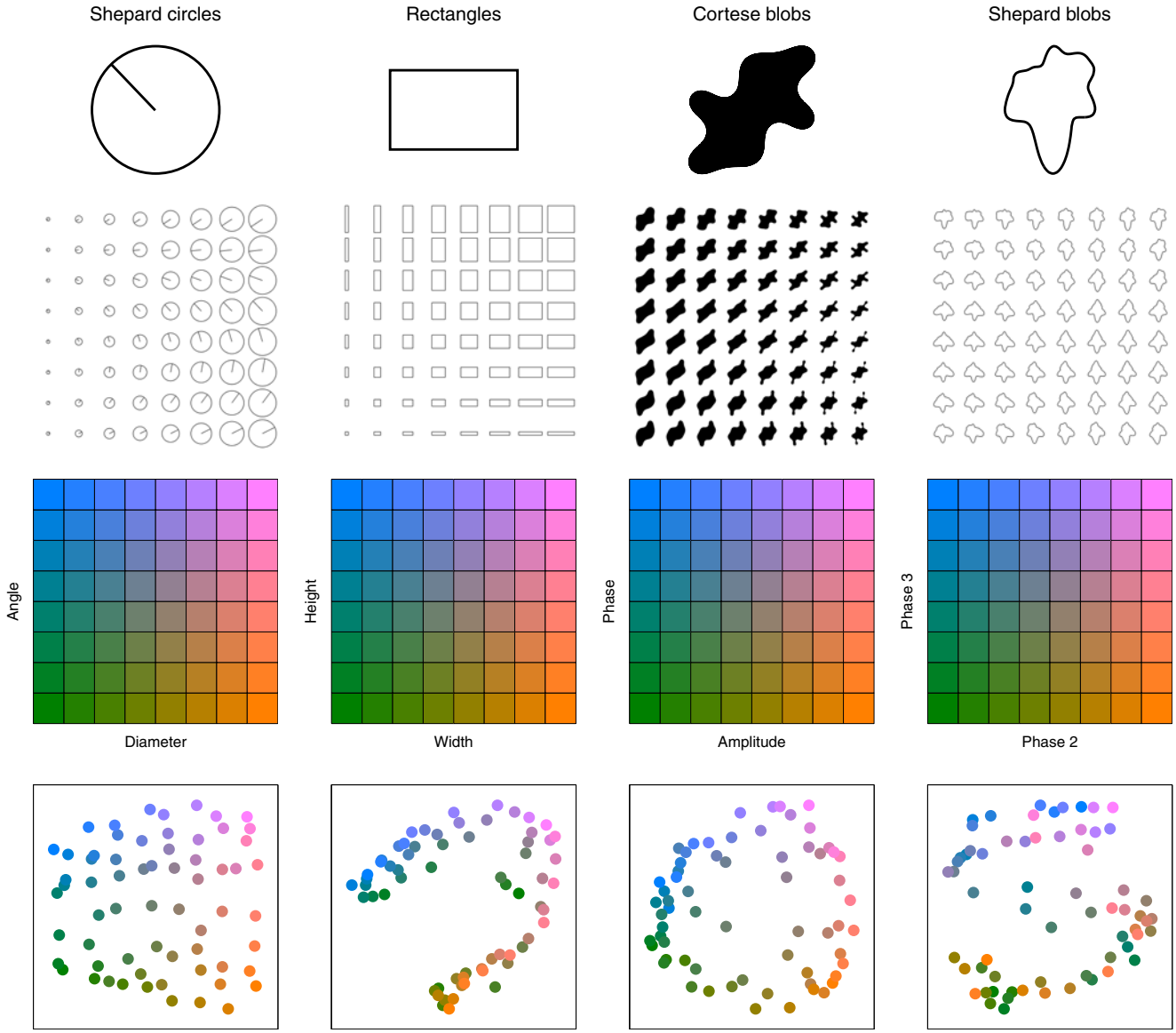

Fig. 1 The four stimulus sets. Each of the four columns corresponds to one stimulus set, as specified by the titles along the top of the figure. The first row shows representative items from the stimulus set. The second row presents all 64 stimuli from each stimulus set in the 8-by-8 grid of the canonical coordinate space. The third row represents each stimulus as a different colored square, arranged in the same 8-by-8 grid of canonical coordinates. The final row represents the stimulus sets in their multidimensional-scaling coordinates; in these images, each marker represents one stimulus, with the color corresponding to the color used to represent the same stimulus in the third row

For participants in the first generation of a chain, the feedback was based on one of the four initial category structures, shown in the first and second columns of Fig. 2 (the first column shows canonical coordinates, the second MDS coordinates, with different colors indicating category membership). The different initial category structures are included so as to allow diagnosis of the chains losing the influence of their initialization and converging on category structures that simply reflect inductive biases, and the initial structures were selected to represent a range of possible category structures. The first two are linear boundaries, one aligned with a single dimension and one using both dimensions. The other two are discretized versions of the category structures described by McKinley and Nosofsky (1995), corresponding to boundaries produced by taking each category to be a mixture of Gaussians.

For the participants in the remaining generations, feedback was provided according to the test session responses of the participant in the previous generation. Participants were not made aware that their test responses would be used in later generations, and they did not have any contact with other learners from different generations. The training session was organized into blocks containing 32 trials each, with the order of presentation of the stimuli being randomized within each block.

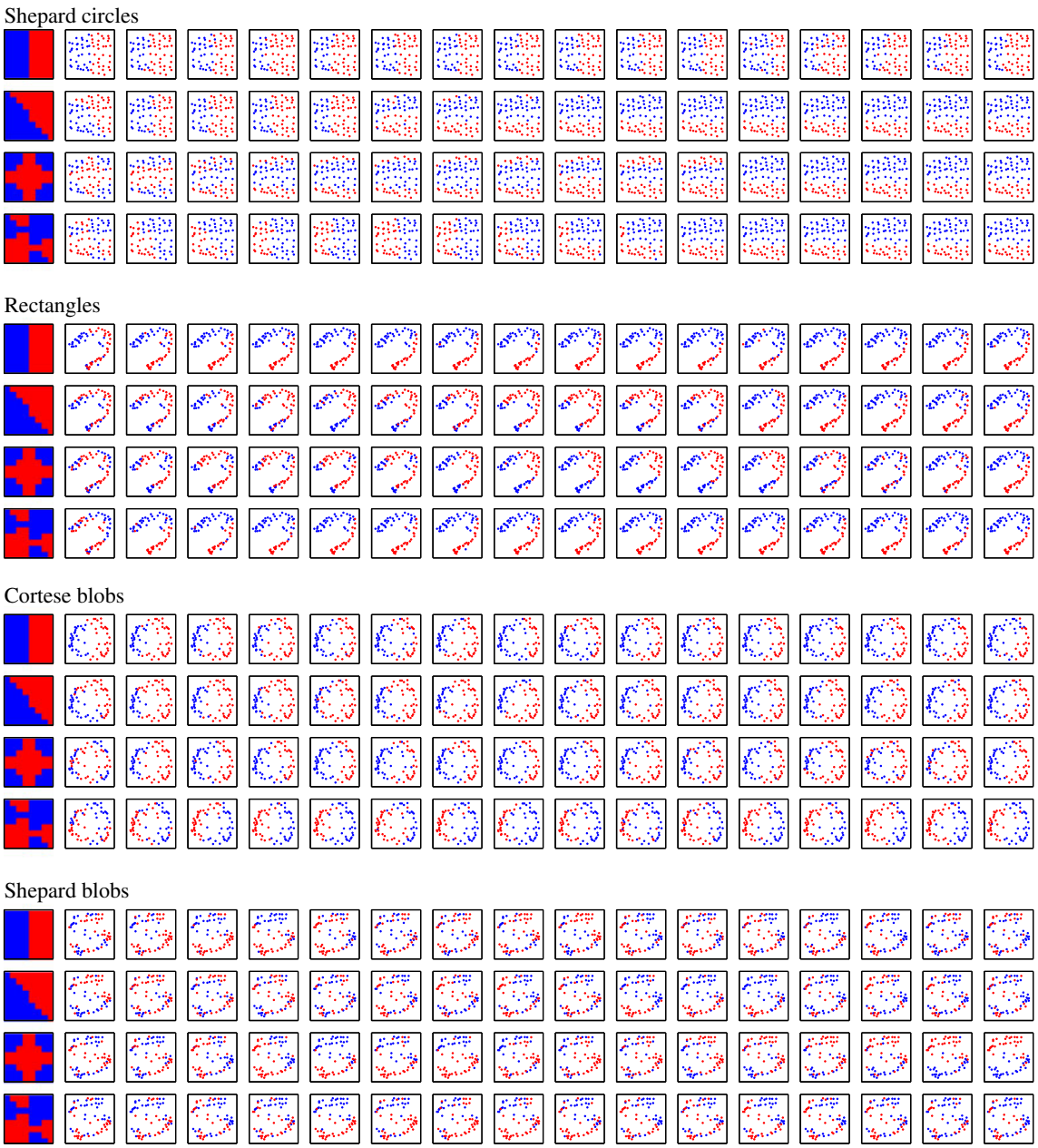

Fig. 2 Samples of responses from the iterated-learning experiment. Each row is an iterated-learning chain; one chain is shown for each combination of stimulus set and initial category structure. The colors indicate category membership, and each image shows the generalization responses of a single learner. Each learner learned from examples of the category structure shown to the immediate left. In each chain, the first and second images show the initial category structure in canonical and multidimensional-scaling (MDS) coordinates, respectively. All of the subsequent images are category structures produced by human learners, presented in MDS coordinates

In the test session, the participant classified all 64 items in a random order without feedback. Participants continued to the test session only if they had correctly answered at least 22 of the 32 training trials in any training block.[1] Otherwise, the participant completed another block of the training session, with exactly the same set of stimuli. If a participant had not reached the learning threshold after 20 blocks or 30 min, the experiment was ended and the participant was excluded. Of the participants, 38 reached the maximum number of blocks, and 25 reached the time limit without achieving the learning criterion. A further 31 participants were excluded because some of their data were not sent to the database server due to network connection issues. All of the excluded participants were identified in the course of the experiment and replaced

[1] This threshold was selected because 22 correct responses out of 32 trials indicates with $p < .05$ that the responses are not purely random, according to an exact binomial test. This also corresponds to the threshold for "positive evidence" (a Bayes factor of 3 or more; Kass & Raftery, 1995) in favor of the probability of a correct response being greater than .5.

by others to fill in their positions in the chains before the experiment continued. The data of 31 additional participants were excluded because they were assigned to duplicate positions in a chain that had already been filled by another participant; this occurred when many participants attempted to take the experiment simultaneously and our system was unable to properly allocate them all to unique chains.

Supervised category-learning experiments have almost exclusively used relatively balanced categories, with roughly equal numbers of members of each category. To focus our investigation of human inductive biases on these cases, we excluded 445 participants who assigned more than twice as many stimuli to one category than to the other in the test session. These participants were identified as the experiment was running and replaced immediately. This had an effect that was equivalent to conditioning the resulting distribution over category structures on the categories being relatively balanced. Although excluding these participants may seem wasteful, it prevented the iterated learning chains from exploring the very large space of unbalanced category structures, which would ultimately produce a large amount of uninformative data.

Results

No qualitative differences were found between the two participant pools in the analyses presented below, so their data were combined. Figure 2 shows one representative chain of 15 generations for each of the 16 conditions, with different colors being used to indicate category membership.[2]

Convergence analysis

The chains shown in Fig. 2 appear to have converged to similar category structures for each stimulus set, regardless of initialization. To quantitatively verify that all chains converged, we examined the distribution of responses across generations to determine whether this distribution was influenced by the initial category structures. First, we performed a cluster analysis of the test session responses, with the variation of information (VI) metric (Meila, 2003) being used as the distance function. The VI metric is an information-theoretic measure of the distance between partitions that depends only on how stimuli are classified, and not on the locations of those stimuli. The VI metric is invariant to relabelings of the categories, so two structures that are identical but switch the category labels would have a VI value of zero.[3]

For each of the 15 generations, the 96 category structures from all six replications of the four stimulus types and four initial conditions were automatically allocated to ten clusters using the k-means algorithm. We then examined how the frequencies with which category structures were assigned to these clusters were influenced by the initial category structures, for each stimulus type separately. We aggregated the frequencies of the clusters over the last ten generations and then used a Bayesian test for independence (Nandram & Choi, 2007, p. 222) to compare each pair of initial category structures (i.e., six pairs for each stimulus type). Of the 24 resulting comparisons, four produced a Bayes factor greater than 1 (i.e., providing evidence for an effect of initial category structure). One comparison corresponded to the two linear boundaries for rectangles, which resulted in a difference in the boundaries being produced by participants in those chains (although both boundaries appeared with some frequency in the chains resulting from the other initial category structures). The other three comparisons showed a difference between the one-dimensional linear boundary and the other initial category structures for the Shepard circles. The dimension that this rule picked out was not the dimension that the other chains producing a one-dimensional linear boundary favored. This difference endured through most of the experiment, being reduced only in the last couple of generations. We thus omitted the chains for the Shepard circles initialized with a one-dimensional linear boundary from our subsequent analysis.[4]

Visualizing inductive biases

In order to characterize both the overall pattern and the variability in the category structures produced by iterated learning in the chains for a stimulus set, we performed an analysis that made it possible to visualize all of the category structures for each stimulus set simultaneously. We used the VI distances between category structures as the input to a metric MDS analysis (Torgerson, 1958) of the last

[2] A set of figures showing all of the results are included in the supplementary materials. To promote further exploration of the results by other researchers, the full set of results is available online at http://cocosci.berkeley.edu/iteratedCatData/.

[3] In order to compare across the different stimulus sets, we identified stimuli by their canonical coordinates. Although the MDS analysis showed that people formed representations that differed from the canonical coordinates for some stimuli, the analysis of convergence required only that we find clusters that could be used to summarize the distribution of responses for each generation. The canonical coordinates were thus sufficient for this analysis.

[4] The persistence of the other linear boundary suggests that it might occupy another mode of the distribution over category structures reflecting people's inductive biases. However, since this mode was not visited by any of the other chains, we made the conservative decision to omit these results.

ten generations of responses.[5] The results are shown in Fig. 3, with each dot in a panel corresponding to one of the category structures generated by the participants. For all stimuli other than the Shepard circles, there are 240 (six replications for four initial category structures over ten generations) data points. The omission of one initial category structure for the Shepard circles resulted in 180 data points.

Discussion

The iterated-learning paradigm offers a new perspective on the problem of characterizing how people learn categories. A chain of participants each take part in a traditional supervised category-learning task, with generalization judgments being made on the basis of a training set. Each chain provides a window into the dynamics of how people are learning from other people's judgments. Taken together, the structures that emerge from this simulated process of cultural transmission create a picture of human inductive biases for a set of stimuli. In the remainder of this article, we will consider how well this picture coheres with previous findings based on the empirical comparison of computational models of category learning, focusing on the three questions raised in the introduction.

Use of boundaries based on multiple dimensions

Previous work on category learning has provided strong evidence that people find it easier to learn categories that have boundaries that align with a single stimulus dimension than to learn those that depend on multiple dimensions (e.g., Ashby & Maddox, 2005; Ashby, Queller, & Berretty, 1999; Goudbeek, Swingley, & Smits, 2009; Kruschke, 1993; Shepard, Hovland, & Jenkins, 1961), and that they tend to produce such structures in unsupervised learning tasks (Ahn & Medin, 1992; Imai & Garner, 1965; Medin et al., 1987; Milton & Wills, 2004). The inductive biases revealed in our experiment, shown in Fig. 3, are largely consistent with these results: Where clear category boundaries were observed, those boundaries tended to be linear and aligned with a single (psychological) dimension. The most interesting deviation from this tendency was observed with the Shepard circles, where one of the more prevalent category structures involved a discontinuous category. However, even in this case the resulting category boundaries were both linear and both aligned with a single dimension.

[5] Metric MDS was used because the sheer number of responses would have made nonmetric MDS prohibitively computationally costly.

Separable and integral stimuli

Consistent with existing findings on the relation between the separability of dimensions and the formation of linear boundaries (e.g., Ashby & Maddox, 2005; Handel & Imai, 1972; McKinley & Nosofsky, 1996), the bias toward one-dimensional, linear boundaries varies with the stimulus type. The distinction between separable and integral representations is based partly on whether people can identify the dimensions along which stimuli vary (Garner & Felfoldy, 1970). This has direct implications for category learning, since people can only form rules or allocate attention over dimensions that they can identify. The results shown in Fig. 3 indicate that for the separable stimuli—the Shepard circles and rectangles—the vast majority of the category structures produced by the participants were linear boundaries aligned with a single dimension, whereas for the integral stimuli—the Cortese and Shepard blobs—these category structures were far less dominant. Consequently, our results suggest that people have a strong preference for simple category structures, using a one-dimensional linear boundary when the stimulus type allows it.

Perhaps the main surprise yielded by these results is that people still produced linear boundaries aligned with the dimensions that were used to define the integral stimuli. This may be a result of these stimuli (and, in particular, the Cortese blobs) not being truly integral, with participants being able to identify an underlying dimension that correlates with the complex dimensions used to define the stimuli. For example, the "fatness" of the Cortese blobs correlates with the amplitude dimension, and this may be why people produce category structures that favor boundaries orthogonal to this dimension. The use of a square array of stimuli might also provide a distributional cue as to the underlying dimensional structure. In future work, it would be interesting to explore inductive biases for other clearly integral stimuli (such as color chips varying in hue and saturation) with a circular stimulus array, to investigate whether a similar strong bias for simple structures could be observed.

Prototype and exemplar representations

Different models of category learning can be seen as positing that people favor different kinds of category structures, and thus have different inductive biases. The data produced by our experiment thus offer the opportunity to compare the inductive biases of computational models of category learning to those of people.[6] One enduring debate in the literature on category learning concerns whether people make

[6] We focused on a qualitative comparison here, providing a more quantitative comparison in Canini, Griffiths, Vanpaemel, and Kalish (2011). Given the unusual size and richness of our data set, we hope that it will be used as a benchmark for models of category learning in the future.

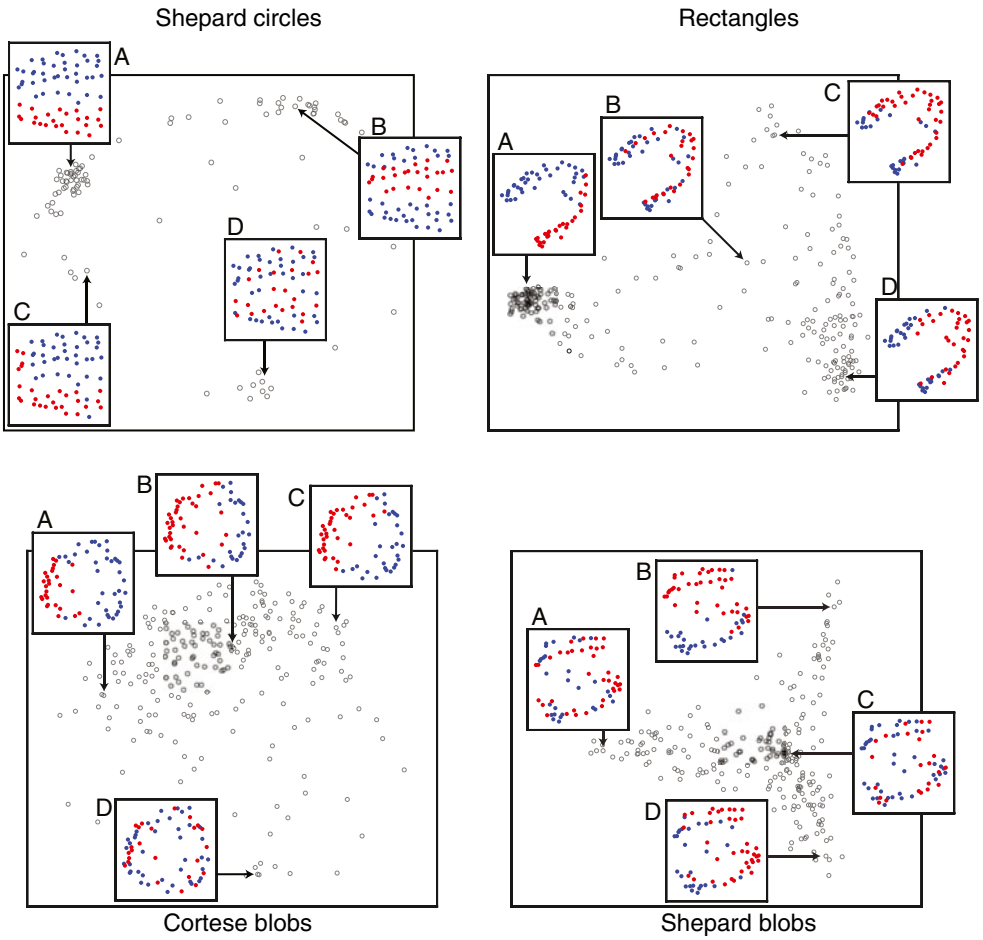

Fig. 3 Two-dimensional representation of the category structures produced by the last ten generations of human learners. Each medium-gray dot in the large squares represents the category structure produced by one participant. For each stimulus set, four individual category structures are shown in detail to demonstrate the variation among people's responses and identify the contents of the clusters

categorization decisions by comparison to a prototype abstracted from the training stimuli (e.g., Reed, 1972) or by recalling the training stimuli themselves (e.g., Nosofsky, 1986). One of the arguments in favor of storage of exemplars is that people can (eventually) learn complex category structures that cannot easily be represented in terms of prototypes (McKinley & Nosofsky, 1995). The inductive biases revealed in our experiment provide an interesting perspective on this debate. Complex category structures were produced by iterated learning for all stimulus sets. However, the proportion of category structures falling into this class depended on the stimulus set, being least pronounced for the separable stimuli and most pronounced for the (arguably most integral) Shepard blobs. The majority of category structures produced for most of the stimulus sets took a form that was more consistent with a prototype representation (or a simple rule), with a single coherent region for each category. Although exemplar models can mimic the predictions of prototype models (Medin & Schaffer, 1978), they do not do so generically. That is, if people were always using an exemplar strategy, we should not see this degree of coherence in the outcomes of iterated learning. Consequently, our results suggest that people can adopt different modes of learning, using prototypes or rules when the data allow it, and relying on exemplars when this is not possible. This conclusion is consistent with several previous models that have explored how these different approaches to category learning can be integrated (Ashby et al., 1998; Erickson & Kruschke, 1998; Nosofsky et al., 1994; Vanpaemel & Storms, 2008).

Limitations and extensions

Our experiment is intended as a first demonstration of the power of the iterated-learning method in studying category learning. Its findings could be extended in several ways. First, we focused on the case in which participants learned only two categories; some of the more complex category structures might disappear if it was possible to learn multiple categories (for some preliminary work in this direction, see Xu, Dowman, & Griffiths, 2013). Second, the stimulus sets that we used differed in their representations in psychological space—forming a rough square for the Shepard circles, a

horseshoe for the rectangles, and variations on circles for the Cortese and Shepard blobs. It would be interesting to investigate whether people's inductive biases change when the distribution of stimuli over psychological space changes, and whether this could account for some of the differences between separable and integral stimuli that were observed in our experiment. Research on unsupervised learning has suggested that surprisingly subtle factors can influence the category structures that people tend to produce (e.g., Milton & Wills, 2004); if iterated learning were to reveal a similar sensitivity in inductive biases for a task based on supervised learning, this might be a factor that needs to be taken into account by models of category learning.

Conclusions

Iterated learning provides a way to explore human inductive biases for category learning. By running a single experiment, we obtained results that are consistent with insights that were gleaned from decades of research into human category learning. These results complement previous research and provide an unusually clear picture of the factors that contribute to human category learning. We view these results as a first step toward gaining a more comprehensive understanding of how people learn categories. By running iterated-learning experiments with other stimuli and other tasks, we can begin to build toward a catalogue of human inductive biases that can be used to guide the development of both psychological theory and machine-learning systems that come closer to human performance.

Author note Preliminary results of this work were presented at the Annual Conference of the Cognitive Science Society (Canini et al., 2011). K.R.C. and T.L.G. were supported by Grant Nos. IIS-0845410 and BCS-0704034 from the National Science Foundation and by Grant Nos. FA-9550-07-1-0351, FA-9550-10-1-0232, and FA-9550-13-1-0170 from the Air Force Office of Scientific Research. W.V. acknowledges the support of research grants from the KU Leuven Research Council (OT/11/032 and CREA/11/005).

References

Ahn, W.-K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science, 16,* 81–121.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429. doi:10.1037/0033-295X.98.3.409

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442–481. doi:10.1037/0033-295X.105.3.442

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53. doi:10.1037/0278-7393.14.1.33

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., Queller, S., & Berretty, P. T. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61,* 1178–1199.

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

Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2011). Discovering inductive biases in categorization through iterated learning. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Expanding the space of cognitive science: Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1667–1672). Austin, TX: Cognitive Science Society.

Cortese, J. M., & Dyre, B. P. (1996). Perceptual similarity of shapes generated from Fourier descriptors. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 133–143. doi:10.1037/0096-1523.22.1.133

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107–140. doi:10.1037/0096-3445.127.2.107

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1,* 225–241.

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1913–1933.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science, 32,* 68–107. doi:10.1080/03640210701801974

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science, 31,* 441–480. doi:10.1080/15326900701326576

Handel, S., & Imai, S. (1972). The free classification of analyzable and unanalyzable stimuli. *Perception & Psychophysics, 12,* 108–116.

Hull, C. L. (1920). Quantitative aspects of evolution of concepts: An experimental study. *Psychological Monographs, 28*(1), 1–86. doi:10.1037/h0093130

Imai, S., & Garner, W. R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology, 69,* 596–608.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. doi:10.1080/01621459.1995.10476572

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.

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology, 12,* 4–34.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science, 5,* 3–36.

Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* 1–27.

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science, 33,* 969–998.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332. doi:10.1037/0033-295X.111.2.309

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 128–148. doi:10.1037/0096-1523.21.1.128

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 294–317. doi:10.1037/0096-1523.22.2.294

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238. doi:10.1037/0033-295X.85.3.207

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19,* 242–279.

Meila, M. (2003). Comparing clusterings by the variation of information. In B. Schölkopf & M. K. Warmuth (Eds.), *Learning theory and kernel machines* (Vol. 2777, pp. 173–187). Berlin, Germany: Springer.

Milton, F., & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 407–415. doi:10.1037/0278-7393.30.2.407

Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw Hill.

Nandram, B., & Choi, J. W. (2007). Alternative tests of independence in two-way categorical tables. *Journal of Data Science, 5,* 217–237.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. doi:10.1037/0096-3445.115.1.39

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. doi:10.1037/0278-7393.13.1.87

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79. doi:10.1037/0033-295X.101.1.53

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26,* 303–343.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition, 121,* 83–100. doi:10.1016/j.cognition.2011.06.002

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

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 347–363. doi:10.1037/0278-7393.21.2.347

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1,* 54–87.

Shepard, R. N., & Cermak, G. W. (1973). Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cognitive Psychology, 4,* 351–377.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75,* 1–42.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: Wiley.

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review, 15,* 732–749. doi:10.3758/PBR.15.4.732

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B, 280*(20123073), 1–8. doi:10.1098/rspb.2012.3073