Cognition 120 (2011) 106–118

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Can semi-supervised learning explain incorrect beliefs about categories?

Charles W. Kalish [a,*], Timothy T. Rogers [b], Jonathan Lang [c], Xiaojin Zhu [d]

[a] Department of Educational Science, University of Wisconsin-Madison, 1025 West Johnson St., Madison, WI 53705, United States
[b] Department of Psychology, University of Wisconsin-Madison, United States
[c] Department of Philosophy, University of Wisconsin-Madison, United States
[d] Department of Computer Science, University of Wisconsin-Madison, United States

ARTICLE INFO

Article history:
Received 27 May 2010
Revised 17 February 2011
Accepted 14 March 2011
Available online 6 April 2011

Keywords:
Categorization
Learning
Stereotyping

ABSTRACT

Three experiments with 88 college-aged participants explored how unlabeled experiences—learning episodes in which people encounter objects without information about their category membership—influence beliefs about category structure. Participants performed a simple one-dimensional categorization task in a brief supervised learning phase, then made a large number of unsupervised categorization decisions about new items. In all three experiments, the unsupervised experience altered participants' implicit and explicit mental category boundaries, their explicit beliefs about the most representative members of each category, and even their memory for the items encountered during the supervised learning phase. These changes were influenced by both the range and frequency distribution of the unlabeled stimuli: mental category boundaries shifted toward the middle of the range and toward the trough of the bimodal distribution of unlabeled items, whereas beliefs about the most representative category members shifted toward the modes of the unlabeled distribution. One consequence of this shift in representations is a false-consensus effect (Experiment 3) where participants, despite receiving very disparate training experiences, show strong agreement in judgments about representativeness and boundary location following unsupervised category judgments.

1. Introduction

If there is any consensus in current theorizing about human categorization and induction it is probably that, by and large, people are good at it. Most theories assume that human learning mechanisms accurately model the statistical structure of the world in order to make correct inferences about the category membership or hidden properties of novel items when they are encountered. Some theories explicitly propose that people behave optimally in this regard (Anderson, 1991; Kemp, Perfors, & Tenenbaum, 2007); others do not directly argue for optimality but nevertheless view categorization and inductive inference as arising from some statistical approximation of the true structure of the environment (Kruschke & Johansen, 1999; Rogers & McClelland, 2004). At the same time, it is apparent that people often form mistaken representations of categories. Social stereotypes constitute one obvious example—people believe that boys are better at math than girls (Herbert & Stipek, 2005) or that attractive people are more ethical than plain people (Smith, McIntosh, & Bazzini, 1999) despite the absence of real differences in the distributions of these attributes in the population. In the natural world, common misconceptions include the beliefs that whales are fish, or that bats lay eggs. A complete theory of human categorization must account for both the successes and the failures. The current paper explores semi-supervised learning as part of the explanation for how categorization can go wrong.

Semi-supervised learning (SSL) has been well studied in the field of machine learning (e.g., Chapelle, Zien, & Schol-

* Corresponding author. Address: Educational Psychology, Rm 880b EdSciences, 1025 W. Johnson St., Madison, WI 53706, United States. Tel.: +1 608 262 9920, fax: +1 608 262 0843.
E-mail address: cwkalish@wisc.edu (C.W. Kalish).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.03.002

kopf, 2006; Zhu & Goldberg, 2009), but only recently have the key ideas begun to inform theories about human categorization and induction (Love, Medin, & Gureckis, 2004; Vandist, De Schryver, & Rosseel, 2009; Zhu, Rogers, Qian, & Kalish, 2007). For this reason, we begin by introducing semi-supervised learning and reviewing recent studies of semi-supervised learning in people. We then consider how semi-supervised learning may lead people to draw incorrect conclusions about category structure.

Most approaches to object categorization assume either that category learning is fully supervised or fully unsupervised. In supervised approaches, the learner is always provided, in each learning episode, with some depiction or description of the stimulus (usually a vector of observed features x) as well as true information about the property to be learned (e.g. the category label or some other novel feature y). On the basis of experience with some set of n such training items, the learner acquires a mapping from the stimulus feature space to the learned property which can then be applied to novel items whose features are observed but whose true category label or target property is unknown. We refer to the feature to be predicted as the "label"; instances where this property is known are "labeled", those for which it is unknown are "unlabeled". There is no implication that labels must be words. The best known computational models of human categorization—for instance, ALCOVE (Kruschke, 2002), the generalized context model (Nosofsky & Palmeri, 1997), the Rational model (Anderson, 1991), and connectionist approaches (Rogers & McClelland, 2004)—have mainly been explored in the context of supervised learning.

Unsupervised approaches also suggest that the learner encounters, in each learning episode, a depiction or description of a stimulus, but is not provided with any information about the item's true category membership (see Pothos & Bailey, 2009; Zeithamova & Maddox, 2009). Instead, the learner must simply learn to group items into categories based on their observed properties, and to use the resulting "unsupervised" categories to make inferences about a new item's class membership or unobserved properties. Unsupervised learning is exemplified by the many available methods of cluster analysis (Du, 2010; Kojima, Perrier, Imoto & Miyano, 2010), competitive-learning algorithms (Si & Treves, 2009), and methods for computing topographic maps such as the Kohonen learning rule (Kwok & Smith, 2005).

The category learning problems normally faced by human beings are, however, neither fully supervised nor fully unsupervised. Children (and adults of course) very frequently encounter objects in the world without labels (that is, without an authority providing the true class membership or hidden properties of each item), but occasionally they also receive labels, as when mom or dad points out an object and names it, or when teachers explicitly teach new facts about familiar items in class. Semi-supervised learning is an approach to knowledge acquisition in which the learner has access to both labeled and unlabeled examples and must combine these to categorize and make inferences about new items.

In machine learning, where semi-supervised learning has been formally studied, a key insight has been that, for some kinds of problems, the learner can converge more rapidly on true beliefs about category structure if information from both labeled and unlabeled sources is combined. To illustrate the intuition, consider the example of a traveler camping in the wilderness of a foreign land (see Fig. 1). The traveler observes an animal in the shadows and can discern it is about 1.5 feet long. His companion tells him the animal is a dax. Sometime later he observes a larger animal, about 2.5 feet long, and his companion informs him it is a zog. From this labeled information, the traveler may infer that daxes are typically about 1.5 feet long; zogs are typically about 2.5 feet long; and the boundary between them is somewhere around 2 feet in length. But now suppose the knowledgeable companion leaves to collect firewood, and the traveler is left alone to observe animals scuffling in the shadows. Over time he observes animals of different lengths, and he notices that there are two clusters: a group a smaller animals about a foot in length, and a group of larger animals about 2 feet in length. The traveler might reasonably infer that the smaller group consists of daxes and the larger group of zogs, and might adjust his beliefs about these categories accordingly, despite not receiving any further instruction about true category labels. For instance, he may end up deciding that daxes tend to be about a foot long (rather than 1.5 feet); that zogs tend to be about 2 feet long (rather than 2.5 feet); and that the boundary between these is about 1.5 feet in length. Thus the probability density of the unlabeled distribution might reasonably be used to adjust conclusions about both the central tendencies and boundary between categories, compared to the conclusions drawn from labeled data alone.

Such an adjustment will be beneficial to the learner given certain assumptions about the relationship between the unlabeled distribution and the category structure. Specifically, if it is the case that the labeled categories correspond to dense regions in the unlabeled distribution, and that category boundaries align with sparsely occupied re-

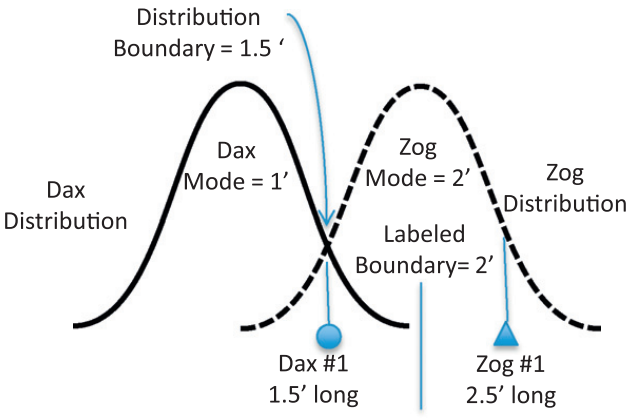

Fig. 1. Distributions of labeled and unlabeled items. Dax's are small, with a modal length of 1′. Zog's are larger, with a modal length of 2′. The learner is trained on a relatively large Dax (1.5′) and a large Zog (2.5′). This experience leads to an incorrect "Labeled" boundary (2′). Experience with unlabeled items from the distributions of Daxs and Zogs allows the learner to reline the boundary, and match the trough between the distributions (distribution boundary).

gions of the unlabeled distribution, then the unlabeled distribution can be fruitfully employed to glean better estimates of the categories' central tendencies and boundaries. Indeed, Gaussian mixture models capturing these intuitions have shown that, under the correct assumptions, a semi-supervised learner can converge on true beliefs about category structure with very few labeled items.

This use of labeled and unlabeled data can, however, lead the learner toward incorrect conclusions when the assumptions fail. To see this, consider that, in the animal kingdom, size is not always a good cue to category membership. House cats and mountain cats, for instance, are quite different in size despite being members of the same category, whereas mountain cats and coyotes are of comparable size despite being members of different categories. Suppose our traveler is camping in a forest where there are many feral cats about a foot long, some mountain cats that are about 3 feet long, and some coyotes that are about 4 feet long. The traveler first glimpses a 3-foot mountain cat in the shadows and is told that it is called a dax, and later views a 4-foot coyote and is instructed that it is a zog. From this labeled information, the traveler might correctly conclude that the size boundary between daxes and zogs is about 3.5 feet. Now, when the traveler is left alone he views many animals that are about a foot long (feral cats), and many that are between 3–4 feet in length. In this case, the true category boundary between cats and coyotes (around 3.5 feet) does not align well with the gap in the unlabeled distribution (around 2.5 feet). If the traveler alters his beliefs about category structure to conform better to the unlabeled distribution, he will arrive at the incorrect conclusion—that daxes are about 1 foot in length; that zogs are about 3.5 feet in length; and that the boundary between these is around 2.5 feet. He may even end up concluding that his guide was wrong about the 3-foot-long animal he previously labeled as a dax! In general, when the true category labels are distributed over items in ways that do not "line up with" the probability density of the unlabeled data, semi-supervised learning of this kind can mislead the learner.

In this paper we consider whether this is a potential mechanism by which people draw incorrect conclusions about category structure. This possibility requires, of course, that people do actually combine labeled and unlabeled observations when learning about category structure. There is remarkably little research on this topic in the literature. Zaki and Nosofsky (2004) approached the problem indirectly when they studied what they called "learning during transfer" (i.e., learning without feedback during the test phase of a categorization study), but they did not attempt to measure or model the magnitude of such learning. VanDist and colleagues (2009) explicitly studied learning from labeled and unlabeled examples in a simple categorization task and failed to find any influence of unlabeled examples on human performance: Participants did not learn a category boundary any faster when unlabeled trials (no feedback) were interleaved with labeled (corrective feedback). Their task, however, required learning of an "information-integration" boundary combining two non-integral dimensions, which is thought to depend on immediate corrective feedback (Ashby, Queller, & Berretty, 1999). The task also employed a fairly high proportion of labeled trials (50%) in the "semi-supervised" condition, possibly minimizing any impact of unlabeled data.

Zhu and colleagues (Zhu, Rogers, Qian, & Kalish, 2007) provide the best evidence for human semi-supervised learning in a one-dimensional two-category learning task. Participants viewed visually-complex shapes varying along a line in a multidimensional feature space, and had to learn to assign each shape to one of two classes. In a supervised phase, participants learned to classify a single exemplar from each category (each appearing 10 times), and acquired a category boundary midway between the two labeled items. They then made a large number of categorization judgments without feedback (unlabeled trials). Unlabeled items were selected from a bimodal distribution aligned so that the two original labeled items fell either on the rightmost or leftmost tails of the two peaks. Over the course of categorizing these unlabeled items, the participants' mental category boundaries tended to shift away from the original learned boundary between the labeled items and toward the trough in the bimodal distribution from which unlabeled items were selected—that is, toward the left when the labeled trials appeared on the rightmost tails, and toward the right when the labeled items appeared on the leftmost tails. Participant behavior was well fit by a mixture-of-Gaussians model of semi-supervised learning, suggesting that people are indeed apt to assume that category labels "pick out" clusters in the feature representation space.

In the following three experiments, we used the method employed by Zhu and colleagues to investigate how human categorization behavior changes with exposure to unlabeled data. The key difference is that, in the earlier work, the category structures suggested by labeled and unlabeled data were roughly consistent: each labeled item was closest to a different peak in the unlabeled distribution. Here we consider what happens when the distribution of unlabeled items grossly violates the category structure suggested by labeled data alone. Do learners then ignore the unlabeled distribution and base their conclusions solely upon the labeled data? Or do they change their beliefs about category structure based on the unlabeled distribution, making judgments that end up violating the information provided by the labeled experience?

Like Zhu et al. (2007), our studies employ stimuli that vary along a single dimension. This is a somewhat artificial scenario insofar as most natural categories involve items that vary along multiple stimulus dimensions simultaneously, but it brings at least two advantages. First, one-dimensional category learning is the only setting where semi-supervised learning in people has been conclusively demonstrated to date. Second, stimuli that vary in multiple dimensions raise additional questions about feature selection: how do people decide which dimensions are important for the categorization problem? Such questions, though certainly of considerable interest, are orthogonal to the issues explored in the current work. For these reasons, we have explored a 1D category learning task. In the general discussion, we will consider further the appli-

cability of our findings to real-world category learning problems, and how the approach might be extended to multidimensional stimulus domains.

2. Experiment 1

Experiment 1 introduces the basic paradigm used in the three experiments. Participants encountered schematic images of women varying along a single dimension, width. The women were presented as coming from one of two islands. In an initial supervised phase, two labeled examples—one from each category—were presented repeatedly. At each presentation participants guessed the island from which the woman came and received corrective feedback. A block of unlabeled examples followed in an unsupervised phase. In the baseline condition, a modest number of unlabeled items (28) fell at evenly-spaced points between the two labeled items, allowing us to estimate the learned boundary between categories following the supervised experience. In the experimental conditions, a large number of unlabeled examples (411) were sampled from a mixture of two Gaussian distributions positioned so that the peaks and trough of the bimodal distribution violated the category structure suggested by the two labeled items (see Fig. 2). In all three experiments, the central question was whether unsupervised classification of items sampled from the bimodal distribution of unlabeled examples would lead participants to draw conclusions about the structure of the two categories that contradicted the conclusions drawn from the supervised learning of labeled examples.

2.1. Method

2.1.1. Participants

The participants included 43 undergraduate students at a large mid-western university who volunteered for the study in order to receive extra credit in their Psychology classes. 24 participants were in the baseline condition and 29 were in the experimental condition.

2.1.2. Design and procedure

Participants were seated at individual computers in a computer classroom. The experiment began by asking participants to imagine that they were anthropologists studying the inhabitants of two remote islands, Pitolan and AinaKanu. The participants were informed that they would see various silhouettes of women from the two islands, and that their task was to learn which women were from Pitolan and which were from AinaKanu. After reading the basic instructions, participants moved to categorization trials. Each trial consisted of a single stimulus (silhouette) appearing on the computer screen with the prompt, "Which island is she from?" The stimulus appeared centered on the screen. Two buttons, each labeled with an island, appeared below offset to the right and left. Assignment of island name to thin and wide stimulus categories was randomized across participants. Participants indicated their categorization decision by clicking one of the two buttons. Selecting an island produced correct/incorrect feedback in the supervised phase but no feedback in the unsupervised phase. A second button click ended the trial and began the next. There was no time limit on responses.

2.1.2.1. Supervised phase. In both baseline and experimental conditions participants began with a supervised learning phase in which they viewed five instances each of two labeled stimuli presented in random order. There were two different sets of labeled stimuli utilized for this experiment, with half of the participants in each condition receiving one set of stimuli and the other half receiving the other set of stimuli. The "left" labeled set consisted of two relatively thin stimuli, one with a pixel-width of 80 and the other with a pixel-width of 115. The "right" set consisted of two somewhat wider stimuli of pixel-width 135 and pixel-width 165. Critically the boundary between each set of labeled stimuli (the mean of the two) was offset from the trough in the distribution of unlabeled stimuli and split either the leftmost or rightmost mode—in this sense, the boundary suggested by the labeled examples was inconsistent with the boundary suggested by the unla-

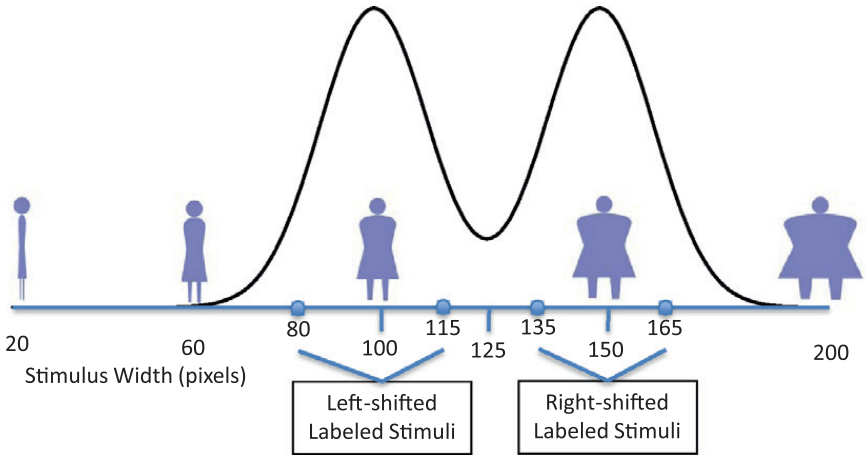

Fig. 2. Examples of stimuli and frequency distribution (Experiment 1). Bi-modal distribution of stimuli had peaks at 100 and 150 with mean of 125 and range from 20 to 200. Left-shifted labeled stimuli had values of 80 and 115. Right-shifted labeled stimuli had values of 135 and 165.

beled distribution (the trough). In this phase participants received corrective feedback: The corrective feedback constituted the labeling for stimuli in this phase. When the stimulus was assigned to the correct island, the word "yay!" flashed upon the screen. When the participant made an incorrect assignment, the word "boo!" flashed upon the screen.

2.1.2.2. Unsupervised phase: Baseline condition. Since our primary research question was whether participants' beliefs about category membership would change after exposure to unlabeled examples drawn from a shifted distribution, we first needed to assess the beliefs they formed from the supervised phase alone. The baseline condition was designed to estimate where participants implicitly place the category boundary, and subsequently to measure their explicit beliefs about the most representative members of each category, following the supervised phase. Thus immediately following the supervised phase, participants in the baseline condition classified seven unique items four times each (28 trials total), including the two labeled items and five additional items appearing at equally-spaced intervals between these. The 28 items appeared in random orders generated separately for each participant. There was no feedback in this unsupervised phase, rendering the stimuli "unlabeled".

2.1.2.3. Unsupervised phase: Experimental condition. Immediately following the supervised training, participants in the experimental condition viewed and categorized 411 unlabeled items selected as follows. First they viewed 37 stimuli in random order sampled in steps of 5 pixels along the full range from the smallest (20) to the largest (200) possible values. As in the baseline condition, this "grid sampling" allowed us to estimate where participants implicitly place the category boundary following the supervised phase. In contrast to the baseline condition, however, the grid spanned the full allowable stimulus range rather than just the range between labeled items. Following this, participants viewed 300 items sampled from a mixture of two Gaussian distributions (the frequency distribution) plus an additional repetition of the 37 grid items randomly intermixed to ensure full coverage of the range. The two Gaussians had means at 100 and 150 with standard deviations of 13; the trough between the two modes was at 125. Finally, the 37 grid stimuli were again presented at the end to provide a final estimate of the implicit category boundary after exposure to the bimodal distribution. Thus participants viewed and categorized 411 items in total. From the experimenters' perspective these 411 decisions were divided into: initial grid (1–37), Gaussian distribution plus intermixed grid (38–374), and final grid (375–411). As in the baseline condition, participants received no feedback about their decisions, so the items were unlabeled.

2.1.2.4. Assessment of explicit beliefs about categories. Following both the baseline and experimental unlabeled conditions we elicited explicit judgments about the most typical members of each category and about the boundary between categories following the unsupervised experience. The participants read the following instructions: "Now we would like you to show us your idea of the typical or average islander. Use the slider to select the body size most representative of a person on the island named." Participants set the position of a slider that controlled the width of a schematic woman's silhouette on the screen. Once participants had finished this task, they read the following instructions: "Now please show us the woman at the boundary between the two islands. Use the slider to indicate the point at which women shift from one island to the other." Participants again used the slider to indicate their explicit, final category boundaries.

2.1.2.5. Materials. The stimuli consisted of simple shapes suggestive of female silhouettes (see Fig. 1). The shapes differed from each other only in terms of the width of the torso, which ranged from 20 (for the thinnest silhouette, which measured 20 pixels across at its widest point) to 200 (for the widest silhouette, which measured 200 pixels across at its widest point). The size of the head was fixed.

2.1.2.6. Measures. The experiment is designed to measure participants' implicit and explicit beliefs about category structure (size of the most representative examples of each category and location of the boundary between them) and to compare these against theoretically relevant landmarks in the labeled and unlabeled distributions. From the labeled distribution, the relevant landmarks are the "labeled boundary," that is, the point midway between the two labeled examples in the supervised phase, as well as the labeled points themselves. If participants employ just the labeled information in guiding their category decisions (the null hypothesis), they should place the boundary between categories near the labeled boundary, and should judge items near the labeled points to be most representative of each category.

From the unlabeled distribution, the relevant landmarks are the "distribution boundary," that is, the trough in the unlabeled distribution, and the modes of the bimodal distribution. If participants' beliefs about category structure are altered by exposure to the unlabeled distribution (SSL hypothesis), they should place their category decision boundary closer to the trough than to the labeled boundary, and their judgments of the most typical items in each category should be shifted away from the labeled points and toward the modes of the unlabeled distribution.

To test these different hypotheses, we must measure the participants' beliefs about the location of the category boundary and the most representative items. With regard to the boundary, we considered three measures derived from the participants' responses: initial implicit, final implicit, and explicit. To estimate implicit beliefs about the location of the boundary between categories, each participant's categorization judgments were fit to logistic functions. Assignment of a stimulus to the wide island was coded as 1, assignment to the thin island as 0. The point at which the estimated logistic function crossed 50% was taken as the implicit category boundary. Each participant in the experimental condition responded to two sets of grid stimuli, thus generating two implicit boundary judg-

ments: initial and final. As participants in the baseline condition did not experience the two Gaussians, only a single initial implicit boundary was calculated in this condition. Finally, participants in the baseline and experimental conditions also made explicit judgments about the location of the boundary. To assess participants' beliefs about the most representative members of each category, we simply took their explicit judgments expressed using the slider. These measures of beliefs about category structure were then compared to the boundary and representativeness landmarks from the labeled and unlabeled distributions.

2.2. Results and discussion

Participants' implicit category boundaries in the baseline condition did not differ significantly from the labeled boundary, $t(23) = .90$, *ns* (see Fig. 3). Participants' explicit estimates of the boundary in this condition also did not differ from the labeled or the implicit boundary, $t(23) = 1.8$, $t(23) = 2.0$ respectively, both *ns*. Finally, participants' estimates of the most typical Pitolan and AinaKanu stimuli in the baseline condition were located near to, and did not differ significantly from, the two labeled items with which they were presented, both $t(23) < 1$, *ns*. In general, participants in the baseline condition drew reasonable conclusions following the brief supervised training: the two labeled items were representative of their classes and the boundary between classes lay about midway between them.

In contrast, the implicit and explicit category boundaries observed in the experimental condition did change after experience with unlabeled examples (see Fig. 2). Following experience with the unlabeled bimodal distribution, both the implicit and explicit category boundaries were significantly different from the labeled boundary in the direction of the distribution boundary (for the final implicit boundary estimates, mean difference = 18.1 pixels, $t(28) = 3.8$, $p < .001$; for the explicit boundary estimates, mean difference = 22.3 pixels, $t(28) = 7.0$, $p < .0001$).

The shift in category boundaries occurred very early in the unsupervised phase. Recall that the first 37 examples spanned the full range at 5-pixels intervals (presented in random order), allowing us to estimate participants' initial category boundaries. The most dramatic boundary shift happened during this initial grid exposure (mean difference = 13.8 pixels, 76% of the difference observed in the final implicit category boundary). One implication of this rapid shift is that participants' implicit beliefs about the category boundary must be shaped partly by information about the range of the stimuli in the domain. During the initial grid exposure, participants in the experimental group received no information about the frequency distribution of unlabeled examples, yet category boundaries still shifted closer to the middle of the stimulus range. In fact this early shift was large enough that, although the implicit boundaries continued to move following subsequent exposure to the frequency distribution, this added effect was not significant (mean difference between initial and final implicit boundaries = 4.2 pixels, $t(28) = 1.1$). Only the explicitly estimated boundaries were significantly shifted relative to the initial implicit boundary, $t(28) = 3.0$, $p < .01$. From this experiment, then, it is unclear whether the observed effects arise solely from the range of unlabeled items, or whether the frequency distribution also matters. Indeed, because we did not anticipate this strong effect of range information, Experiment 1 was not designed to adjudicate this question—in these stimuli the distribution boundary (125) was very near the midpoint of the

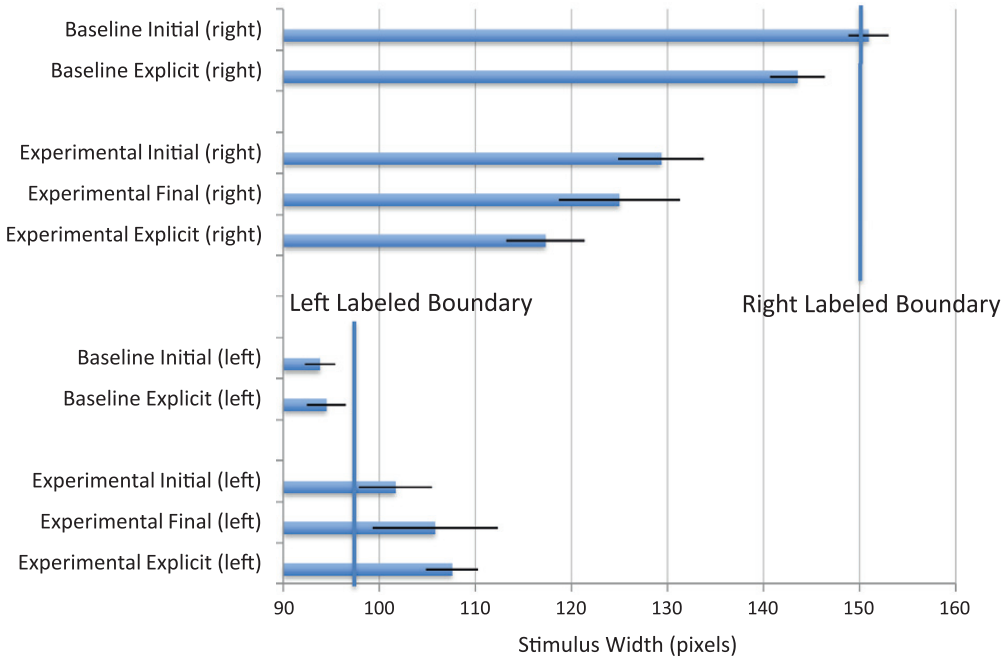

Fig. 3. Results from Experiment 1. Vertical lines indicate labeled category boundaries, from supervised phases. Bars indicate mean boundaries estimated from participants' responses to Initial and Final grid items, and mean explicit boundaries from participants' ratings. Right and Left labeled conditions are presented separately. Horizontal lines indicate one standard error +/− in boundary estimates or ratings.

stimulus range (110). Experiment 2 thus uses a different distribution in order to distinguish these effects.

Judgments of the most typical example of each category were also affected by the unlabeled examples. Recall that the labeled items appeared on either side of one mode in the unlabeled distribution. Of these two items, one always fell between the two peaks, and hence was closer to the other mode. We will refer to the category denoted by this item as the inconsistent category. The other, consistent category had labeled items relatively near one of the modes of the distribution. For the consistent category, judgments of the most typical item aligned well with the location of the labeled item. Though the mean estimate across participants was shifted significantly away from the labeled item and toward the nearest mode in the unlabeled distribution $t(23) = 2.3$, $p < .05$, this effect was driven by two outlying participants whose estimates were more than two standard deviations from the mean, $t(21) = 1.6$, ns with outliers excluded. Even including the outliers, the estimate of the most typical member of the consistent category did not differ significantly from the corresponding judgment from the baseline condition, $t(52) = .02$, ns. For the inconsistent category, however, judgments of the most typical member were shifted very strongly in the direction of the far mode (the mean difference from the labeled item was 41.7 pixels, $t(28) = 8.3$, $p < .001$). Thus judgments of the most typical category members were strongly influenced by exposure to the unlabeled data.

The general conclusion from Experiment 1 is that unlabeled examples do strongly influence both the implicit and explicit conclusions people draw about category structure. This result is consistent with earlier demonstrations of semi-supervised learning (Zhu et al., 2007) but extends this work by looking at a situation in which the labeled and unlabeled examples suggest very different conclusions about category structure.

One illustration of the magnitude of this effect is how the labeled items were categorized before and after exposure to unlabeled items. More than half of the participants in the experimental condition (16 of 29) changed their decisions about the category membership of one of the training items during the unsupervised phase. For example, the stimulus that initially defined the large island was eventually assigned to the small island after exposure to unlabeled data. Almost all of those who changed (14 of 16) did so after exposure to just the 37 range items. Effectively, many participants rapidly unlearned the original boundary and ended up mis-classifying the labeled items from the supervised phase.

Given these dramatic changes in categorization decisions, one might wonder whether the supervised experience has any lasting effect on participants' beliefs about category structure. Perhaps the initial supervised training merely allows the learner to associate one label with one direction of the dimension of interest and the other label with the other direction, and all subsequent learning is driven by structure in the unlabeled distributions. Put differently, perhaps the participants were doing fully unsupervised learning with the unlabeled data, and were only using the supervised training to figure out which label goes with which mode. Perhaps exposure to the unlabeled data completely erases any trace of the initial supervised learning.

If unlabeled data come to dominate category representations, then participants who received labeled items centered on the left mode would end up with the same beliefs about category structure as those who received labeled items centered on the right mode. Despite starting with quite different labeled examples, the two groups received exactly the same unlabeled examples. An ANOVA with label condition (left-shifted, right-shifted) as a between-subjects variable and boundary estimate (initial, final) as a within-subjects variable revealed a main effect of label condition, $F(1, 27) = 11.8$, $\eta_p^2 = .30$, $p < .005$, with no main effect of boundary estimate and no interaction. Participants in the right-shifted and left-shifted label conditions had different implicit category boundaries for both the initial and the final grid items, $F(1, 43) = 12.4$, $p < .005$, and $F(1, 43) = 6.0$, $p < .05$, respectively, suggesting a persistent effect of the brief supervised experience even after 411 unlabeled trials. On the explicit measures of the most typical instances and of the category boundary, however, the left and right labeled conditions did not differ (all $t(27) < 1$).

In summary, people do change their category representations in response to unlabeled examples. The results of Experiment 1 suggest that these changes happen fairly rapidly and may be quite dramatic. Although unlabeled, these examples do carry information about the range of the stimuli and about the frequency distribution of stimulus values. Participants' category boundaries were influenced by unlabeled examples very quickly in the task, after exposure only to range information. Category boundaries were shifted toward the midpoint of the stimulus range relative to the training examples. It was less clear from Experiment 1 whether participants' behavior was influenced by the frequency distribution of unlabeled examples. One limitation of Experiment 1 was that the midpoint of the stimulus range was very close to the trough in between the two Gaussian distributions. That is, the "natural" boundary in the distribution was very close to the middle of the range. To test whether category representations are affected by the distribution, as well as or instead of the range, we need to use a different distribution.

3. Experiment 2

Experiment 2 had two goals. The first was to determine whether participants are sensitive to both the distribution and the range of the unlabeled distribution. Thus in this experiment the unlabeled examples were drawn from a mixture of two Gaussian distributions situated so that the trough between peaks was fairly distant from the midpoint of the range. As before, participants first completed a supervised learning phase with two labeled items; however in this case the two items were chosen so that the midpoint between them (the labeled boundary) lay between the midpoint of the range and the trough in the unlabeled distribution. If the range of the unlabeled examples has more influence than the density of the unlabeled

distribution, then the category boundary should shift away from the trough and toward the midpoint. If the boundary is more influenced by the density of the distribution than the range, it should shift away from the midpoint and toward the trough.

The second goal of Experiment 2 was to assess whether participants retain an accurate memory of the two labeled items after the unlabeled experience. In Experiment 1 we saw that participants judged the most typical member of the inconsistent category to be quite different from the one labeled member of the category they had previously viewed. This tendency may arise because the participant's memory of the labeled item itself has changed following exposure to the unlabeled distributions. Alternatively, it may be that participants remember the labeled item quite well but have come to realize that this item is not very representative of its category. To adjudicate these hypotheses, we added two additional questions to the explicit testing phase at the end of the experiment. In addition to asking participants to indicate the typical Pitolan and AinaKanu women and the boundary between them, we also asked them to reproduce as accurately as possible the two labeled items viewed at the beginning of the experiment.

3.1. Method

3.1.1. Participants

Eighteen undergraduate students at a large mid-western university participated in the study in order to receive extra credit in their Psychology classes.

3.1.2. Design and procedure

This experiment exactly replicated the experimental condition of Experiment 1 with two exceptions. First, the two labeled items and the range of the unlabeled distribution were altered. Specifically, unlabeled stimuli ranged from 20 to 250 pixels in width, reflecting a 50 pixels range increase from Experiment 1. This change resulted in 47 grid stimuli (instead of 37), and shifted the midpoint of the range to 135 pixels instead of the 110 pixels midpoint in Experiment 1 (see Fig. 4 for illustration). For Experiment 2, the two Gaussians had modes of 85 and 135, with a trough at 110. The two labeled stimuli presented during the supervised experience were at 96 and 150 pixels. Thus the labeled boundary was at 123 (midway between 96 and 150), the midpoint of the full range was larger than this value (135), and the trough between modes was smaller (110). The second difference from Experiment 1 was that, after estimating the typical Pitolan, AinaKanu, and the boundary, participants adjusted a slider to show the original Pitolan and AinaKanu women observed at the beginning of the experiment.

3.2. Results and discussion

In contrast to Experiment 1, participants' initial category boundaries, as estimated from the first grid items, did not differ from the labeled boundary, Mean = 119, $t(17) = 1.2$ (see Fig. 4). Relative to Experiment 1, the labeled boundary was closer to the range midpoint to begin with. For example, in the Right-shift condition of Experiment 1 the distance from labeled boundary to midpoint was 25 pixels. In Experiment 2 this distance was 13 pixels. Boundaries may not have changed so quickly in Experiment 2 because the midpoints of the labeled and unlabeled ranges were more similar. Final implicit boundaries were significantly different than initial implicit boundaries, $t(17) = 3.5$, $p < .01$, indicating that exposure to the frequency distribution had an effect over and above exposure to the range.

The final implicit category boundaries (estimated following experience with the bimodal distribution) did differ from the labeled boundary, $t(17) = 3.1$, $p < .005$. Critically, the difference was in the direction of the trough between the distributions and away from the midpoint of the range. The label boundary was smaller than the middle of the range to begin with, and the implicit boundary became even smaller after exposure to the unlabeled distribution. The mean final implicit boundary was significantly less than (thinner than) the midpoint, $t(17) = 5.3$, $p < .001$ but not significantly different from the trough between the distributions, $t(28) = -.8$. Participants' explicit boundary

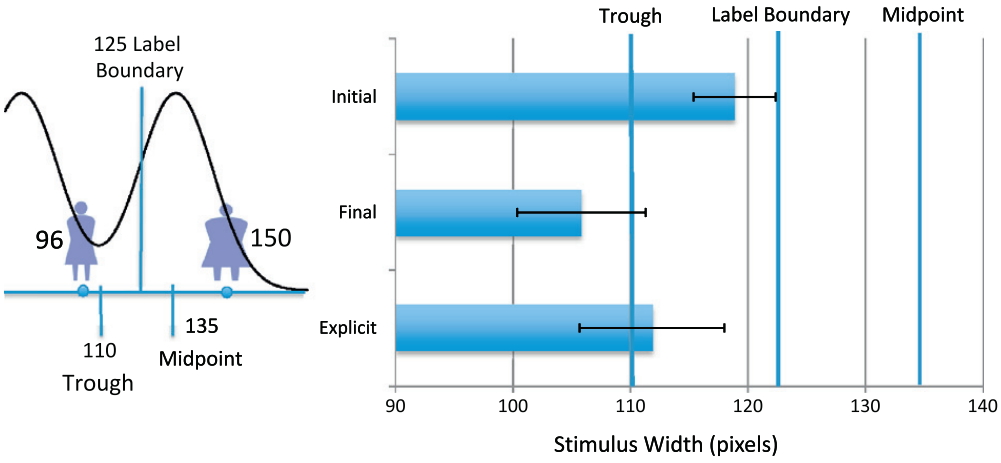

Fig. 4. Distributions and Results from Experiment 2. Vertical lines indicate objectively delined (potential) category boundaries: the trough between the two distributions, the label boundary, and the midpoint of the stimulus range. Bars indicate mean boundaries estimated from initial and Iinal grid items, and mean explicit boundary ratings. Horizontal lines indicate one standard error +/− in boundary estimates or ratings.

judgments showed the same pattern. The mean explicit boundary was significantly smaller than the range midpoint, $t(17) = 3.8$, $p<.05$, but not from the trough of the distribution $t(17) = .3$.

Judgments of typicality again showed the effect of experience with unlabeled examples. The most typical thin-islander (67) was significantly thinner than the labeled item for that category (96), $t(17) = 7.0$, $p < .001$. The typical wide-islander (163) was not reliably wider than the labeled item (which was already on the extreme edge of the corresponding mode, 150), $t(17) = 1.6$, $p = .06$. The estimated typicalities did not simply reflect the modes of the unlabeled distribution. Rather there was a kind of caricature effect in which the most typical items were judged to be more extreme than the modal unlabeled items. The mean judgment for the most typical thin category stimulus was significantly smaller than the mode of the thin distribution, $t(17) = 4.4$, and the mean judgment for the typical wide category stimulus (163) was wider than the mode of the wide distribution (135), $t(17) = 3.4$, both $p < .005$. Note that typicality was only measured after exposure to the distribution (in the explicit judgments): There was no measure of typicality after the initial grid alone. Thus, unlike category boundaries, we cannot determine whether typicality judgments were sensitive to the range or the distribution, or both.

Interestingly, participants' memory for the original labeled items showed a remarkably similar pattern. The example of the thin category was remembered as being significantly thinner (73) than it actually was (85), $t(17) = 4.6$, $p < .005$, and the wider labeled example (150) was remembered as somewhat though not significantly wider, mean = 157, $t(17) = .8$, ns. Strikingly, memory for the two labeled items did not differ significantly from estimates of the most typical items in each category, both $t(17) < 1.5$, ns—suggesting that memory for the labeled items is altered by the unlabeled experience. Of course, it is possible that memory would have changed the same way without the unlabeled experience: There was no control comparable to the baseline condition in Experiment 1. However, it seems unlikely that memory drift would so closely match typicality judgments.

In Experiment 2, the two labeled examples fell on different sides of all relevant landmarks from the labeled and unlabeled distributions: they received different labels, were on either side of the trough between the Gaussian distributions, and were on either side of the midpoint of the stimulus range. Thus unlike Experiment 1, the category structure suggested by the supervised experience was roughly consistent with that suggested by the unlabeled distribution. Nonetheless, almost half of the participants ($N = 8$) ended up mis-classifying one of the original labeled examples. Only one participant ended up with a boundary that was too high (classifying both labeled examples as thin), while seven ended up with a boundary that was too low (classifying both labeled examples as wide). These categorization errors are consistent with the view that the mental category boundary moves toward the trough with some degree of noise so that the membership of the labeled item nearest the new boundary becomes uncertain despite the previous supervised experience.

The results of Experiment 2 confirm and extend the findings from Experiment 1. In general, both experiments demonstrate that unlabeled examples affect people's category representations. Whereas Experiment 1 suggests that people are sensitive to the range of unlabeled items, Experiment 2 demonstrates that the density of the distribution also has a strong influence on category judgments: Boundaries tend to shift toward the trough of the distribution even when this differs from the location of the range midpoint. Experiment 2 also confirmed that explicit beliefs about category structure are influenced by unlabeled experience, and further showed that memory for labeled items can be distorted through exposure to unlabeled items.

4. Experiment 3

Experiments 1 and 2 showed that people's beliefs about category structure can be strongly influenced by the distribution of unlabeled items to which they are exposed, with the consequence that learners can sometimes draw conclusions about category membership that directly contradict their supervised experience. In Experiment 3, we consider whether these tendencies can lead groups of individuals to collectively form shared incorrect beliefs about category structure. The key idea is that individuals living in a common environment are likely to be exposed to similar distributions of unlabeled experiences. The preceding experiments suggest that, by virtue of this shared experience, individuals who receive very different supervised learning experiences should have their beliefs about category structure "reshaped" by unlabeled experience in similar ways—so that, despite extreme differences in the explicit feedback they receive, they may, following experience with unlabeled items, come to agree on certain aspects of category structure.

In Experiments 1 and 2, the learner only rarely received definitive information about a woman's island of origin—there were only two labeled examples—and each participant received the same two labeled examples. Imagine, however, a different case in which the observed information about each item—the width of the silhouette—is completely irrelevant to determining the correct category. Suppose, for instance, that each learner views one Aina-Kanu and one Pitolan woman, but that each example is sampled completely at random from a uniform distribution on the range, with each learner viewing a different random sample. From these supervised learning experiences, different individuals will form widely variable beliefs about the characteristic AinaKanu and Pitolan women and the boundary between them. With exposure to the unlabeled items, however, each individual should "shift" her mental category boundary away from the initial labeled points and toward the trough/midpoint of the distribution. Similarly, representations of most typical islanders should shift from the labeled items toward the two modes of the distribution. Since the distribution landmarks are all the same for each learner, one might initially observe considerable disagreement among learners following the supervised experience, but increasing agreement across learners the boundary and most characteristic

examples with increasing exposure to the unlabeled items. These are the predictions tested in Experiment 3.

4.1. Methods

4.1.1. Participants

Twenty-four undergraduate students at a large midwestern university volunteered for the study in order to receive extra credit in their Psychology classes.

4.1.2. Design and procedure

Experiment 3 was identical to the experimental condition of Experiment 1 with the exception that in the supervised phase each participant saw a randomly selected pair of examples sampled from a uniform distribution over the range. Pairs were chosen with the constraint that the two stimuli be at least 20 pixels apart and that neither member of the pair be less than 30 pixels from the maximum or minimum of the range. The implicit boundary between categories was assessed with an initial set of "grid" items presented immediately following the supervised learning phase, and again with a final grid following exposure to the bimodal distribution of unlabeled items. Explicit beliefs about the boundary and about the most representative Ainakanu and Pitolan were assessed at the end of the experiment using the method from Experiments 1 and 2.

4.2. Results and discussion

The primary question in Experiment 3 is whether, following very different supervised learning experiences, individuals will come to agree about the location of the category boundary and characteristic examples after encounters with a common set of unlabeled instances. In contrast to Experiments 1 and 2, the absolute magnitude of each individual's category boundary is of less interest than is the variance across individuals. Fig. 5 presents two measures of dispersion. The horizontal bars indicate mean differences from the trough in the unlabeled distribution. Error bars represent the standard deviations of these boundaries. From the figure it is clear that, despite

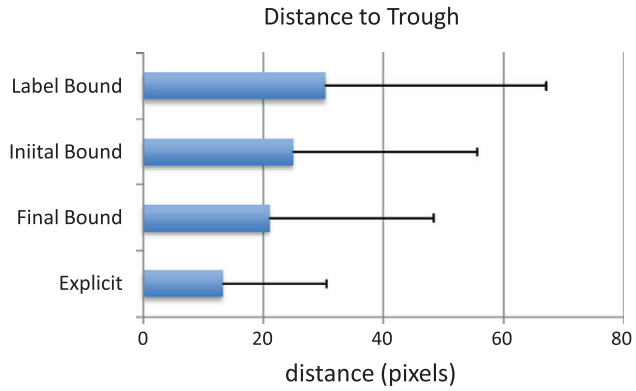

Fig. 5. Mean absolute distances of boundaries to the trough in the distribution. Label bound is the mean between the two labeled examples. The bar represents the mean value across all participants. Initial and final bounds are mean estimates from grid stimuli. The explicit boundary is the mean rating of the category boundary. All error bars indicate one standard deviation.

wide variation in the initial labeled boundaries, individuals increasingly come to agree on the location of the boundary following exposure to the unlabeled distribution.

To formally test this observation, we compared the variance in category boundaries both before and after unlabeled experience. The Morgan-Pitman test is based on $VarY - VarX = Covar(X - Y,\ X + Y)$. Thus the test asks whether there is a significant correlation between the sums of the scores and the differences of the scores. For example, to assess the change in variance from initial to final estimates we sum the two scores for each participant, and take the difference of the two scores for each participant. We then compute the covariance between these two sets of scores (sums and differences). The variance in the training boundaries was 1011. The variance in both initial and final boundary estimates was reduced significantly, 442; $r = .54$ and 408; $r = .51$ respectively, both $p < .01$. The variance in the explicit category boundary ratings was much lower than the variance in the training boundaries, 199; $r = .83$, $p < .001$. Indeed, pairwise comparisons reveal that variance was reduced in two steps: Training > Initial, Final > Explicit.

In addition to becoming less variable, the location of implicitly and explicitly assessed category boundaries also changed. For each individual, we computed the distance of their initial, final, and explicit boundary estimates to the trough in the unlabeled distribution. (Note that with this distribution of unlabeled examples the trough and midpoint of the range are very similar.) With increasing exposure to the unlabeled items, boundary estimates moved increasingly away from the labeled boundary and toward the trough between the peaks of the unlabeled distribution. Initial boundary estimates were significantly closer to the trough than were the labeled boundaries, $t(23) = 2.5$, $p < .05$, demonstrating the effect of exposure to the full range of unlabeled items (as in Experiment 1). Initial and final estimates did not differ in distance, $t(23) = .3$. Finally, explicit boundary estimates were significantly closer to the trough than were initial implicit boundary estimates, $t(23) = 3.3$, $p < .005$. In other words, each individual's boundary estimates moved reliably away from his or her idiosyncratic label boundary and toward the trough in the unlabeled distribution. Because this distribution was common across individuals, these shifts produced increasing agreement as to the location of the boundary.

As in Experiment 1, boundaries often shifted so that the original labeled items changed class after the unlabeled experience. Twenty-one of the participants received training boundaries that were significantly off-center (i.e., at least 10 pixels away from the range midpoint). Of these 21 participants, 12 made category boundary ratings that placed the two labeled items in the same category.

Finally, estimates of the most representative member of each class also changed in systematic ways. Though the mean typicality judgments did not differ significantly from the mean of the corresponding labeled points, there was considerably greater agreement across individuals. The stimuli identified as the most typical or representative instances of their respective categories (after exposure to unlabeled examples) were significantly less variable than

the training items participants saw, variances of 1202 vs. 461 for the "wider" category, and 1463 vs. 365 for the "thinner", both $t(22) > 2.3$, $p < .05$, Morgan–Pittman tests. Thus participants showed substantially greater agreement about the characteristic width of the thinner and wider islanders than warranted by their direct supervised experience.

Note that the consensus achieved by participants in Experiment 3 had a very interesting quality. Participants generally came to agree on the boundary between the island with the thinner women and the island with the wider women. They also came to agree on the typical sizes for thin-island and wide-island women. They disagreed, however, about which island was which. Because labels were randomly assigned to training items, half the participants thought that the AinaKanu women were wide and the Pitolan women were thin, and half the participants thought just the opposite! The participants actually ended up with two distinct, even opposing, stereotypes. Based on the unlabeled experience they all agreed that there was a wide and a thin island, and agreed on the size of the difference. However, based on the labeled experience, they fell into one of two camps about which label indicates which size. Although this may seem to be a very incongruous (and even unrealistic) state of affairs, it may actually be just the kind of in-group/out-group belief observed in the social psychological literature. We all agree that there is a group of nice/smart/beautiful/moral people and a group of mean/ignorant/ugly/evil people. We just disagree about which group is mine and which group is yours.

5. General discussion

In three experiments we have shown that the beliefs people form about category structure are strongly influenced by the unlabeled instances they encounter. The baseline condition of Experiment 1 showed that, immediately following supervised learning with a single instance from each of two categories, participants (i) acquire mental category boundaries near the midpoint between the labeled items and (ii) judge each labeled item to be representative of its category. After categorizing many new items without feedback, however, these beliefs change: Mental category boundaries shift toward the midpoint of the range of unlabeled items (Experiment 1) and toward low-density regions between modes of the unlabeled distribution (Experiment 2). This shift occurs quite rapidly, and is sufficiently robust that many participants end up making categorization decisions that actually violate their supervised learning experience. Finally we have shown that individuals who have quite different supervised learning experiences but are exposed to a common set of unsupervised experiences (Experiment 3) can come to agree on aspects of category structure that are inconsistent with their supervised experience. In Experiment 3, there was no true category boundary between Ainakanu and Pitolan women: examples of these categories were sampled from the same random distribution. Nevertheless, after exposure to the unlabeled distribution, participants largely agreed that one island mainly had wide women and one mainly had narrow women; they agreed on the typical width of women from each island; and they agreed as to the location of the boundary between the categories. The simple example illustrates one mechanism by which groups of people, despite quite variable supervised feedback, can converge on incorrect beliefs about categories.

5.1. Relationship to other work

The results of the three experiments support and extend prior research on semi-supervised learning. Zhu and colleagues (2007) showed that category boundaries shift in response to unlabeled data. In this case, however, the labeled and unlabeled distributions were roughly consistent with one another. Experiments 1 and 3 suggest that when the discrepancy between labeled and unlabeled data is high, people's boundaries shift very quickly and dramatically, leading to re-categorization of even the previously-labeled items. Experiment 2 closely replicates the Zhu et al. results.

In contrast to the results of the current study, Vandist and colleagues (2008) have argued that unlabeled examples do not influence category boundary learning. One possible reason for this discrepancy is the relative frequencies of labeled and unlabeled examples: Vandist et al. provided labels on 50% of their learning trials, and so potentially limited the impact of unlabeled distributions. There are, however, other differences between the studies that might also explain the seeming contradiction. Vandist's paradigm required participants to learn an "information integration" boundary, that is, an oblique boundary in 2d stimulus space with psychologically separable dimensions. Such tasks are often thought to require extensive supervised experience (Ashby, Queller, & Berretty, 1999). We further note that, in Vandist's experiment, the unlabeled examples did not carry any additional information about the range or distribution of stimulus values beyond that provided by the labeled items. The unlabeled items were drawn from exactly the same distribution as were the labeled items, and the midpoint of the stimulus values and the trough in the distribution of examples coincided exactly with the labeled category boundary. Categorization of the unlabeled examples would therefore have reinforced the labeled boundary, rather than shifting it as in the current experiments. Vandist et al.'s work thus allowed these authors to look only at the rate at which the boundary was learned, and not at effects on the category structure acquired.

One of the implications of the current results is methodological. Zaki and Nosofsky (2004) note the possibility of what they termed "learning during transfer." In a typical categorization experiment, the subjects are taught a distinction (using labeled examples) and then asked to categorize a set of "transfer" stimuli (without label feedback) in order to measure what was learned from the training. The measures intended to diagnose category representations may in fact change such representations (see Zhu et al., 2010, for further evidence). The current study confirms that learning can happen during transfer. Although it is difficult to quantify, category representations seem to change rapidly and significantly. The current study used

a very brief category learning phase (only 10 trials) and the criteria for categorization decisions were very clear (stimuli varied along a single dimension). These features of the design may have contributed to the influence of unlabeled examples. We might expect more extensive supervised experience to reduce the influence of unlabeled items. Also, in our study the unlabeled examples varied only on the dimension relevant to the category boundary. With more complex stimuli the variation in the unlabeled examples may be less obvious, and thus have a smaller effect on category representations (see Rogers, Kalish, Gibson, Harrison, & Zhu, 2010). Finally, the current study used a blocked design in which all training occurred before all unlabeled experience. It is interesting to speculate whether a small number of labeled examples interspersed with the unlabeled examples would be sufficient to maintain the trained category representations. These considerations provide directions for future research on semi-supervised learning, as well as hypotheses about minimizing the effects of learning during transfer.

5.2. Strengths and weaknesses of semi-supervised learning

We began the introduction with a suggestion that semi-supervised learning may be one mechanism by which people form incorrect beliefs about categories, and the current results support this hypothesis. It should be clear, however, that semi-supervised learning can also greatly benefit category learning in many cases. Whether semi-supervised learning leads to correct or incorrect beliefs about categories depends on the relation between the distribution of labeled and of unlabeled observations. If category labels tend to apply across contiguous high-density regions of the unlabeled feature space, and category boundaries tend to follow low-density regions in this space, then semi-supervised learning will lead to correct beliefs even when labels are very sparse. Only when category structure conflicts with unlabeled structure and labeled experience is rare will semi-supervised learning fail.

In many natural domains, it is probably true that category labels and boundaries map fairly well onto the structure of unlabeled experience. Rosch et al. (1976) famously argued, for instance, that mental category representations carve the world at its joints. To the extent that this is true, semi-supervised learning will benefit concept acquisition. Furthermore, common misconceptions about natural categories seem to occur in cases where the key assumption fails. For instance, the mistaken belief that bats lay eggs probably occurs because bats share many salient properties with birds and few with familiar mammals, and because there are many properties held in common amongst birds and amongst mammals, but relatively few shared between these groups. In other words, in our unlabeled experience we encounter many bird-like things and many mammal-like things, but few things in between—there is a "trough" in the observed feature space between birds and mammals. Moreover, most people get little direct experience showing them which items lay eggs and which do not (i.e., supervised experience). Thus, although the true boundary between egg-laying and live-bearing animals would fall between the bats and the birds in a multidimensional feature space, the semi-supervised learner will place the boundary at the trough between birds and mammals, and form incorrect beliefs about bats.

Of course, standard models of similarity-based inference make similar predictions. If bats are similar to known egg-laying animals and dissimilar from known live-bearing animals, then bats will be assumed to be egg-layers. Things that cluster together in known ways will be assumed to cluster together in unknown ways. Semi-supervised learning does not involve unique mechanisms. Rather the point is to illustrate some implications of the standard processes of similarity-based learning. Specifically, the intuition that unknown properties will be distributed like known ones (things similar in known ways will be similar in unknown ways) is not always correct. The perception of similarity-based structure can override direct experience that the true distribution of the unknown feature is orthogonal to the known features. The reliance on the predictive value of known/observable clusters can introduce error into learned classifications.

Errors from semi-supervised learning may be reduced in a variety of ways. Extensive supervised experience, or very strong weighting of labeled items, will overcome the tendency to conform to the distribution of unlabeled items. Also, learning about new properties of objects can effectively shift the distribution of unlabeled items in ways that would allow semi-supervised learning to benefit category acquisition. For instance, learning about the ways that bats are similar to mammals—having fur, similar skeletal structure, expressing milk, and so on—will shift the bat away from the bird and toward the mammals in the unlabeled feature space, making it less likely that semi-supervised learning will lead to an incorrect conclusion.

Though these examples focus on misconceptions about natural categories, we believe that other conceptual domains may be more susceptible to these kinds of effects—in particular, domains where the unlabeled distribution is more likely to conflict with the categories we need to learn, and where direct experience of the true category label is rare. Social concepts might constitute one such domain. There is considerable structure in our everyday experience with other individuals—on the basis of properties we can readily observe (unlabeled experience), people fall naturally into clusters that roughly reflect their age, race, sex, socioeconomic status, attractiveness, etc. Yet these clumps in the unlabeled distribution may not be especially useful for drawing inferences about their unobserved properties—which people are smart, which are violent, which are good drivers, which are lazy, which are ethical—and we may get relatively little direct experience of these traits. Social stereotypes often seem to involve using clusters apparent in the unlabeled distribution to govern generalization about these infrequently-observed traits (e.g. women are bad at math, Latinos are lazy). To the extent that such stereotypes reflect incorrect beliefs, semi-supervised learning may provide one mechanism for understanding how they come to be formed in the first place.

One limitation of the current work is that it employed a very simple category-learning setting in which items varied on just one dimension and were assigned to just two categories. In real-world category learning, of course, items

vary in many different feature dimensions and may be assigned to many different categories. The current work was useful for illustrating how and why semi-supervised learning might lead to incorrect beliefs, but to assess the ecological validity of these ideas future work will need to employ multidimensional stimuli assigned to a variety of different categories.

6. Conclusion

The possibility that humans learn from semi-supervised experience has significant implications both for theories of human inductive inference and for the methods we employ to assess these abilities. Though we have investigated semi-supervised learning as a mechanism for understanding the occasional failures of human categorization, in most natural domains it seems likely that semi-supervised learning will more often lead to positive outcomes. So long as the category structure to be learned aligns well with the distribution of unlabeled examples, semi-supervised learning of the kind we have studied here will improve performance. It is likely that this assumption does hold for many of the inductive inference problems we are called upon to solve (see Rosch et al., 1976). The difficulty arises in those cases where natural discontinuities in encountered examples do not actually reflect the category boundaries we are called upon to learn, and where supervised experience is largely unavailable. The current study suggests that, under these conditions, category representations can be strongly distorted by the unlabeled examples. This mechanism may be especially useful for understanding incorrect beliefs in particular conceptual domains, such as social stereotyping.

Acknowledgements

This work was supported by grants from the NSF (DRM/DLS) 0745423 to C.W.K., NSF IIS-0953219, and NSF IIS-0916038 to X.Z. and AFOSR FA9550-09-1-0313 to T.T.R and X.Z. Thanks to Molly Garner for help with design and data collection.

References

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

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

Chapelle, O., Zien, A., & Scholkopf, B. (2006). *Semi-supervised learning*. Cambrdge, MA: MIT Press.

Du, K. L. (2010). Clustering: A neural network approach. *Neural Networks, 23*, 89–107.

Herbert, J., & Stipek, D. (2005). The emergence of gender differences in children's perceptions of their academic competence. *Journal of Applied Developmental Psychology, 26*, 276–295.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*, 307–321.

Kojima, K., Perrier, E., Imoto, S., & Miyano, S. (2010). Optimal search on clustered structural constraint for learning Bayesian network structure. *Journal of Machine Learning Research*, *11*.

Kruschke, J. R. (2002). ALCOVE: An exemplar-based connectionist model of category learning. In T. A. Polk & C. M. Seifert (Eds.), *Cognitive modeling* (pp. 537–574). Cambridge, MA, US: MIT Press.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1083–1119.

Kwok, T., & Smith, K. A. (2005). Optimization via Intermittency with a self-organizing neural network. *Neural Computation, 17*, 2454–2481.

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

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300.

Pothos, E. M., & Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the generalized context model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1062–1080.

Rogers, T. T., Kalish, C. W., Gibson, B. R., Harrison, J., & Zhu, X. (2010). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. In *Proceedings of the 32nd annual meeting of the cognitive science society, Portland, OR*.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rosch, E. et al. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.

Si, B., & Treves, A. (2009). The role of competitive learning in the generation of DG fields from EC inputs. *Cognitive Neurodynamics, 3*, 177–187.

Smith, S. M., McIntosh, W. D., & Bazzini, D. G. (1999). Are the beautiful good in Hollywood? An investigation of the beauty-and-goodness stereotype on film. *Basic and Applied Social Psychology, 21*, 69–80.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: The impact of feedback in learning the information-integration task. *Attention, Perception, & Psychophysics, 71*, 328–341.

Zaki, S. R., & Nosofsky, R. M. (2004). False prototype enhancement effects in dot pattern categorization. *Memory and Cognition, 32*, 390–398.

Zeithamova, D., & Maddox, W. T. (2009). Learning mode and exemplar sequencing in unsupervised category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 731–741.

Zhu, X., Rogers, T. T., Qian, R., & Kalish, C. W. (2007). Humans perform semi-supervised learning too. Paper presented at the twenty-second AAAI conference on artificial intelligence (AAAI-07).

Zhu, X., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. San Rafael, CA: Morgan & Claypool.

Zhu, X., Gibson, B. R., Jun, K., Rogers, T. T., Harrison, J., Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In *Proceedings of the 27th international conference on machine learning (ICML)*, Haifa, Israel.