The helpfulness of category labels in semi-supervised learning depends on category structure

Wai Keen Vong
Daniel J. Navarro
Amy Perfors
School of Psychology
University of Adelaide

Abstract

The study of semi-supervised category learning has generally focused on how additional unlabeled information with given labeled information might benefit category learning. The literature is also somewhat contradictory, sometimes appearing to show a benefit to unlabeled information and sometimes not. In this paper, we frame the problem differently, focusing on when labels might be helpful to a learner who has access to lots of unlabeled information. Using an unconstrained free-sorting categorization experiment, we show that labels are useful to participants only when the category structure is ambiguous and that people's responses are driven by the specific set of labels they see. We present an extension of Anderson's Rational Model of Categorization that captures this effect.

Introduction

Imagine you are walking through an art gallery with an artist friend. As you proceed, your friend occasionally stops to point out a particular painting and tell you the name of the artist who painted it, thereby providing you with labeled data for that painting. All around you are dozens of other paintings in various styles. Although your friend has not commented on these paintings, this unlabeled data might be very informative to you: it can refine your understanding of particular painters or what art styles exist. You might even detect groups of similar paintings that you suspect were painted by the same person, even though your friend has not commented on them.

Learning from both labeled and unlabeled data, as in this example, is referred to as semi-supervised learning (Zhu, Rogers, Qian, & Kalish, 2007; Gibson, Rogers, & Zhu, 2013) and has not been studied to the same extent as other categorization problems. Most research investigates supervised learning, where each example is paired with the appropriate category label (Medin & Schaffer, 1978; Nosofsky, 1986). Other research explores unsupervised learning, where people must learn categories based without any category labels or feedback (Love, 2002; Pothos & Chater, 2002; Pothos et al., 2011). Despite the attention focused on supervised and unsupervised learning, in real life the majority of situations involve mostly

semi-supervised learning: a few labeled instances in conjunction with a large set of unlabeled experiences.

What do we know about human semi-supervised learning? Unfortunately, the literature is somewhat split about its effectiveness. One possibility is that receiving both unlabeled and labeled examples provides very little information over either source alone. Consistent with this, some studies have found that adding unlabeled data has no effect when labeled examples have already been provided (Vandist, De Schryver, & Rosseel, 2009; McDonnell, Jew, & Gureckis, 2012). Similarly, others have found that people are able to learn the structure of categories in an unsupervised manner, and only labels to map words onto existing category representations (Bloom, 2000). Both of these areas of research suggest that semi-supervised learning is not very different from either supervised or unsupervised learning. However, there is evidence for the other possibility too: some studies have found that adding unlabeled data can affect category learning in both humans (Zhu et al., 2007; Lake & McClelland, 2011; Kalish, Rogers, Lang, & Zhu, 2011; Gibson et al., 2013) and computers (Chapelle, Schölkopf, & Zien, 2006).

How can we reconcile these apparently contradictory findings? We begin by noting that the typical framing of semi-supervised learning tasks is somewhat puzzling. Although semi-supervised learning extends both supervised and unsupervised learning, papers on the topic almost invariably compare it to supervised learning. By adopting this perspective, researchers are led to ask whether unlabeled data provides any additional benefit to the learner over and above what can be learned from labeled data. This framing is implicit in the way our art gallery example was described: it was simply assumed that the labeled examples from the artist friend would be useful, and the question was whether the unlabeled paintings might be an additional source of information.

Yet this framing is easily reversed. Consider instead the following variant: As you walk around the art gallery, you see hundreds of examples of paintings. From this wealth of data you might form theories about art styles, pick out individual paintings, and so on. As you do so, your friend points to a few paintings and tells you that those are by Picasso and Monet. So far, the literature on semi-supervised learning has typically assumed that the labeled examples are distributed in a similar fashion as the unlabeled examples. However, this is not true in our art gallery example nor (often) in real life: there may have been many other painters such as Magritte, Pollock and Rembrandt which you only saw unlabeled examples of. More generally, the distribution from which the world generates raw (unlabeled) data need not be at all similar to the one from which a knowledgeable teacher chooses to select (labeled) examples, and a child learning language cannot assume that people are labeling all and only the relevant categories of objects. Indeed, what is relevant changes from context to context, and what is labeled is conditioned on many things (attention, conversational goal) other than providing the optimal category learning information. Thus by framing the problem of semi-supervised learning as one in which unlabeled data as the primary source of information, the focus now shifts to the evidentiary value of the additional labeled data.

Now the relevant question is: When and how might labeled examples be beneficial for category learning above and beyond having only unlabeled examples? One method to assess this would be to compare semi-supervised learning to performance in purely unsupervised learning. Traditionally, the problem of grouping objects into categories has been explored primarily from an unsupervised perspective (Medin, Wattenmaker, & Michalski,

1987; Pothos & Close, 2008; Pothos et al., 2011). However, one of the challenges of unsupervised categorization is the sheer combinatoric explosion of possible ways to sort a group of objects into categories. The number of ways to sort n items is given by the nth Bell number, which grows very rapidly as a function of n: even having only ten stimuli can result in over 100,000 possible different classifications (Medin & Ross, 1997). Despite this search challenge, one can easily imagine circumstances where labeled instances need not be necessary. For instance, people might not need any labels to determine that a Picasso painting was not created by the same artist who created a Monet – the styles are so different that it is obvious just from the unlabeled data that there were two separate categories of artists. In such a situation, semi-supervised learning might not be noticeably different from unsupervised learning.

On the other hand, distinguishing the work of Klee from that of Kandinsky represents a much harder problem for the novice. In fact, when the training items are similar, unsupervised categorization is hard and people have difficulty in determining how many categories to sort objects into and to do so in a consistent manner (Pothos et al., 2011). We hypothesize that it is in precisely these kinds of relatively ambiguous situations where some additional labeled examples may be beneficial, where even just a few labeled examples can substantially reduce the difficult of this huge search problem.

But how would *just a few* labeled examples help so dramatically? One possibility is that labeled examples might serve as a cue to people about what dimensions to attend to. For example, if the labels suggest that there are multiple relevant dimensions, the presence of the labeled data may prompt people to switch from a unidimensional classification strategy to a multidimensional one. While people tend to exhibit a strong unidimensional bias in unsupervised learning (Ashby, Queller, & Berretty, 1999; Medin et al., 1987), some recent work has shown that the presence of a sufficient number of labeled examples can cause people to shift towards multi-dimensional classification strategies (Vandist et al., 2009; McDonnell et al., 2012). However, this is not the only possibility as to how labeled examples might drive categorization. A different set of labeled examples might guide the learner into classifying using only a single dimension instead. Thus, we also hypothesize that labeled examples serve as cues as to which classification strategies to pursue, and that different sets of labeled examples should lead to different classifications.

This paper investigates how and when a small number of labeled examples improves category learning outcomes based on unsupervised data. We test these predictions through an experiment in which people sort unlabeled multidimensional rectangle stimuli into categories. In some conditions, the true category structure is distinct, while in others it is ambiguous. Conditions also differ by whether the labels people are given pick out distinct categories or not. Consistent with our hypotheses, we find (a) that people rely on labels when the underlying category structure is ambiguous, and (b) in that case, people's classification strategies are affected by the labeled examples they receive. In addition, we develop a modified version of the Rational Model of Categorization (Anderson, 1991) and show that it naturally captures people's behaviour in this novel semi-supervised paradigm.

Method

Our experiment took the form of an unconstrained free-sorting task in a semi-supervised setting. Participants were shown 16 two-dimensional stimuli, a maximum of

three of which were labeled (depending on the condition). They were asked to sort the objects into different categories any way they wished. Different conditions manipulated both the kinds of structures people saw as well as the labels associated with the stimuli. The goal of the experiment was to examine which (if any) of these settings promoted semi-supervised learning.

Participants

Data were analyzed from 504 participants (312 males) recruited from Amazon Mechanical Turk and paid either US$0.30 or US$0.50. An additional 34 did not complete the experiment and 52 were excluded for failing to properly respond to a check question (see below). The age of participants ranged from 18 to 69 (mean: 33.3) 56% of participants were from the USA, 39% were from India, and the remainder were from other countries.

Stimuli

The stimuli used in the experiment, shown in Figure 1, consisted of white rectangles with a black border. Inside each of the white rectangles was an inner gray rectangle along to the bottom-right corner. The stimuli varied along two continuous dimensions[1] corresponding to the height of the white rectangle (25 to 65 pixels high) and the length of the inner gray rectangle (10 to 50 pixels long). There were two different stimulus sets, one for each of the two stimulus structure conditions described below. Depending on the label condition, three of the stimuli might have been labeled with a nonsense word (*dax*, *fep* or *wug*), which appeared underneath the associated stimuli. A total of 16 different stimuli were used, all presented simultaneously on the screen.

Design

Participants were randomly assigned to conditions based on two between-subject manipulations. In the first, we varied the coherence of the underlying category structure. In the DISTINCT STRUCTURE condition, the stimuli consisted of three well-separated clusters that varied along both stimulus dimensions, as shown in the top row of Figure 2. The AMBIGUOUS STRUCTURE condition also consisted of three equally sized clusters, but they were much closer together in the stimulus space, as in the bottom row of Figure 2. This made it difficult to distinguish the cluster boundaries from feature information alone. In all conditions the participants were not told how many categories there were: they were instructed to sort the stimuli into as many categories as they felt was necessary.

The second experimental manipulation, shown in the columns of Figure 2, varied the informativeness of the labels that were included. As a baseline, the NO LABEL condition was fully unsupervised with no labels at all. In the DISTINCT LABEL condition, people saw a helpful and informative set of labels: one labeled example from each of the three clusters. By contrast, in the AMBIGUOUS LABEL condition people saw potentially misleading labels: one came from the first cluster, two came from the second cluster, and none came from the third cluster. Of interest is how people's categorizations were affected by the informativeness of the label in combination with the structural coherence in the unsupervised data.

[1] The extent of the variation along the two stimulus dimensions was calibrated by applying multidimensional scaling to a set of pairwise similarity ratings to ensure both dimensions were equally salient.

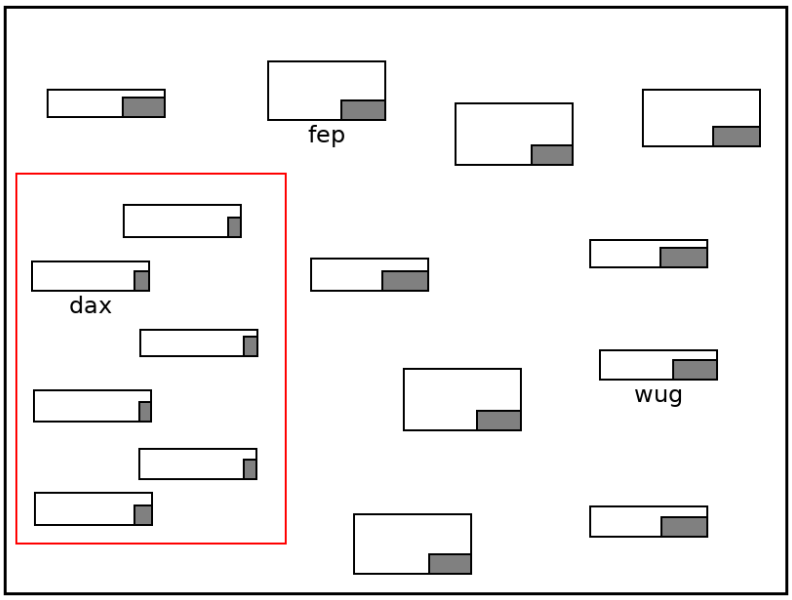

Figure 1. Screenshot from the task illustrating the stimuli and labels used in the experiment (this example is from the DISTINCT STRUCTURE and DISTINCT LABELS condition). In all conditions, people were asked to sort the rectangles into categories by dragging them around the screen into clusters. In this screenshot the participant has already drawn one box around one of the categories they identified.

Procedure

The experiment was run online through a custom website. The cover story informed participants that archaeologists had discovered a number of unknown objects on a recent expedition, and needed help to sort them into different categories. In the labeled conditions participants were told that the archaeologists had discovered some of the names of the objects, which could be used as a guide on how they sorted the stimuli. In the NO LABEL condition the instructions simply recommended using the appearance of the objects to guide the sorting. In all conditions, no indication was given of how many different categories were present in the data.

In order to make sure that people understood the sorting task, before beginning the main task the participants completed a demonstration trial. This trial contained three squares and three triangles of different sizes, where they were asked to sort the shapes into separate piles that they thought should naturally go together. The position of objects in both the demonstration trial and main task were arranged to be non-overlapping and randomly ordered for each participant. The user interface required participants to first click and drag on stimuli until they were sorted into piles they thought should belong together, and then to draw boxes around each pile. If people were unhappy with the boxes they could revert to the click and drag stage until they were satisfied. If anyone failed to group any stimuli inside a box or assigned any stimuli to multiple boxes, a warning would appear and they could not submit their response. The demonstration trial also served as an exclusion

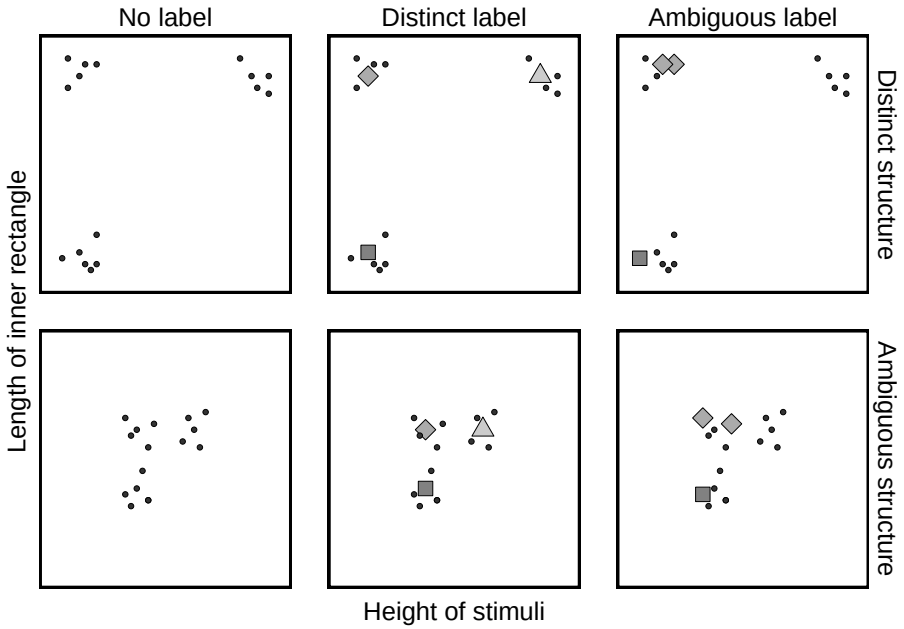

Figure 2. A visualization of the experimental design. The stimuli varied along two continuous dimensions (stimuli height and the length of the inner rectangle). The small black dots represent the unlabeled examples, while the larger stimuli represent the labeled examples, with each shape corresponding to a different category label (*dax*, *wug* or *fep*).

criterion: 52 people failed to sort those stimuli in a sensible way (i.e., not by size or shape) and their data from the main experiment were therefore excluded from further analysis.

Results

Participants produced 216 unique sorts out of 504 solutions analyzed. This level of variability is commensurate with similar tasks in unsupervised categorization (e.g. Pothos et al., 2011). However, the extent of the variability was very different across conditions. To quantify this variability we use the adjusted Rand index (*adjR*), which measures the similarity between two classifications (Hubert & Arabie, 1985). It has a maximum of one when both classifications are the same, and drops to zero if the agreement between them is no more than would be expected by chance. The average *adjR* score among all pairs of participants in each condition is shown in Figure 3, and reveals two key findings.

The first finding was that people did indeed appear to find the ambiguous structure more ambiguous: responses in the DISTINCT STRUCTURE condition were more similar to one another than those in the AMBIGUOUS STRUCTURE condition. Consistent with this, a two-way ANOVA on structure $\times$ label revealed a significant main effect of structure ($F(1, 498) = 293.5$, $p < 0.001$).

The second finding, of more importance, is that the effect of labels was different in different contexts: while there was a significant main effect of label ($F(2, 498) = 14.2$, $p < 0.001$), there was also a significant interaction between the structure condition and

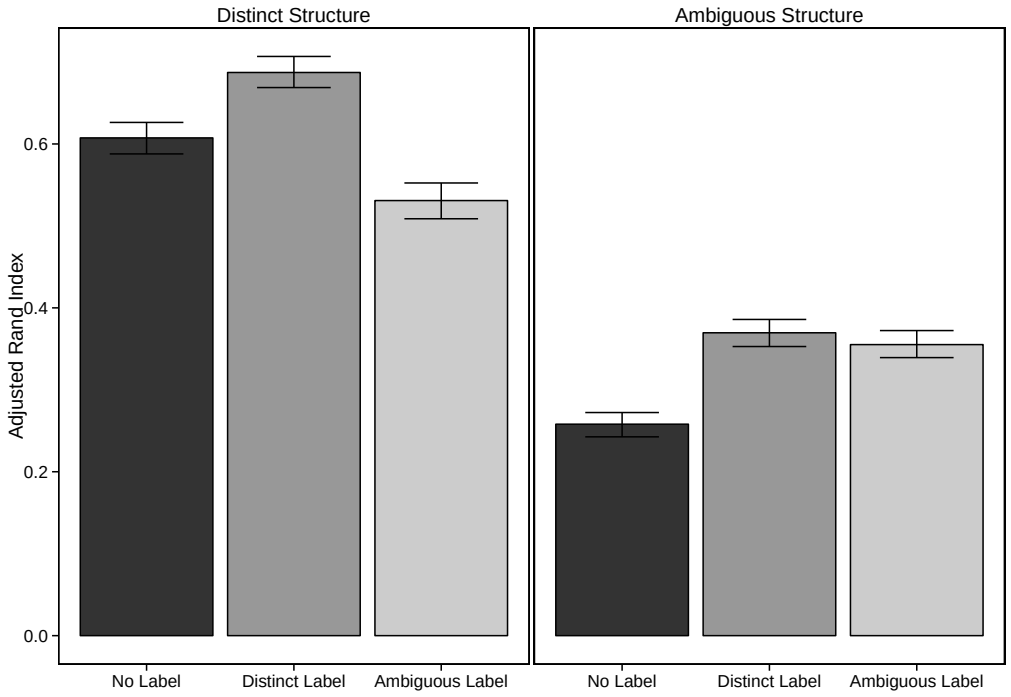

Figure 3. Agreement between participants within condition. Each bar plots the average similarity between solutions (i.e., adjusted Rand index) taken across all subjects in the same condition. Error bars are bootstrapped 95% confidence intervals.

label condition ($F(2, 498) = 10.9$, $p < 0.001$). In the AMBIGUOUS STRUCTURE condition, adding labels increased the degree of agreement among participants regardless of which label set was provided. However, in the DISTINCT STRUCTURE condition, the effect was more subtle. When the DISTINCT LABELS were provided, the labeled data were consistent with the structure of the unlabeled data, and the agreement among participants increased relative to the NO LABEL condition. But when the AMBIGUOUS LABELS were provided, the structure among the labeled examples did not precisely match the structure of the unlabeled data. As a result, the agreement among participants dropped relative to the NO LABEL condition.

On close inspection it turns out that most answers were variants[2] of one of the three classification schemes shown in Figure 4, which we refer to as the three "canonical classifications" for the task. Participants almost always approximately (a) sorted into three categories using both stimulus dimensions, (b) sorted into two categories based on height, or (c) sorted into two categories based on length. We assigned people to one of the three classifications by calculating the *adjR* value between each person's sort and each of the three canonical classifications, and then selected the one that was highest as their classification

[2]It was not unusual for a participant to classify *most* of the stimuli according to one of these schemes, with some of the boundary cases being different; situations like this meant that 187 of the 216 distinct unique sorts were only produced by a single participant.

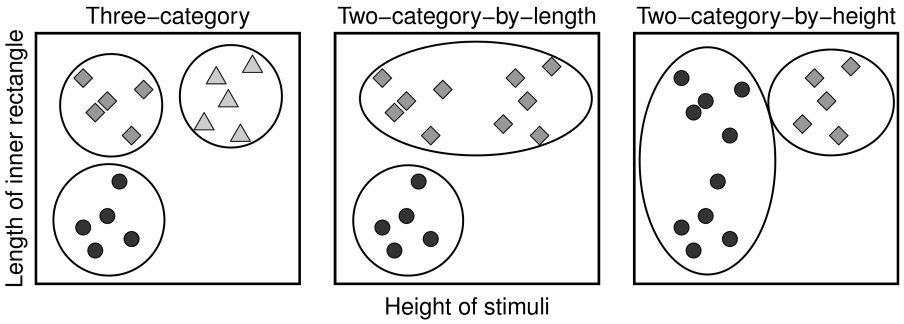

Figure 4. The three canonical classifications used to classify people's responses in the task. While this figure only depicts the canonical classifications for the AMBIGUOUS STRUCTURE condition, the strategies are analogous for the DISTINCT STRUCTURE condition. The three-category strategy required attending to both stimulus dimensions when sorting. On the other hand, the two-category-by-length and two-category-by-height strategies only required attending to a single stimulus dimension corresponding to either the length of the inner rectangle or the height of the stimuli respectively.

strategy.[3] The breakdown of classification type by condition is shown in the top row of Figure 5.

In the DISTINCT STRUCTURE condition the results were straightforward. The choice of labeling scheme had no effect on the classification strategy ($\chi^2(4) = 1.90, p = 0.75$) and participants tended to use the three category solution regardless of the nature of the labels. Even when one cluster of stimuli was given no labels at all, as in the AMBIGUOUS LABEL condition, people detected the unlabeled cluster and did not attempt to group those items with items in the labeled clusters. This suggests that if the category structure is coherent and obvious enough, labels make very little difference to people's categorizations.

For the AMBIGUOUS STRUCTURE condition the story is more complex, and there is a significant difference in classifications depending on the nature of the labels shown ($\chi^2(4) = 26.48, p < 0.001$).[4] In the NO LABEL condition, people were evenly split between the three classification schemes. This reflects the fact that the raw stimulus information was not sufficient for people to infer how to categorize the items. When labels were provided, participants relied on them heavily. In the DISTINCT LABEL condition people preferred the three category solution, since the labeling scheme explicitly picked out the three clusters. In the AMBIGUOUS LABEL condition, the labels ruled out the two-category-by-height strategy, but did not distinguish between the other two strategies. This is reflected in the data, with people split evenly between the three-category and two-category-by-length strategies.

Although the overall pattern of results is a complicated interaction between stim-

[3] We also ran analyses in which we grouped participants into an "other" strategy if their solutions were insufficiently similar to any of the canonical classifications (e.g. having *adjR* values below a certain threshold like 0.2 for all three canonical classifications). The qualitative pattern of results remains unchanged across different threshold values ranging from 0.1 to 0.5.

[4] Significant differences were also observed between each pair of label conditions (NO LABEL and DISTINCT LABEL: $\chi^2(2) = 10.47, p < .01$, NO LABEL and AMBIGUOUS LABELS: $\chi^2(2) = 15.61, p < .01$ and DISTINCT LABELS and AMBIGUOUS LABELS: $\chi^2(2) = 12.69, p < .01$).

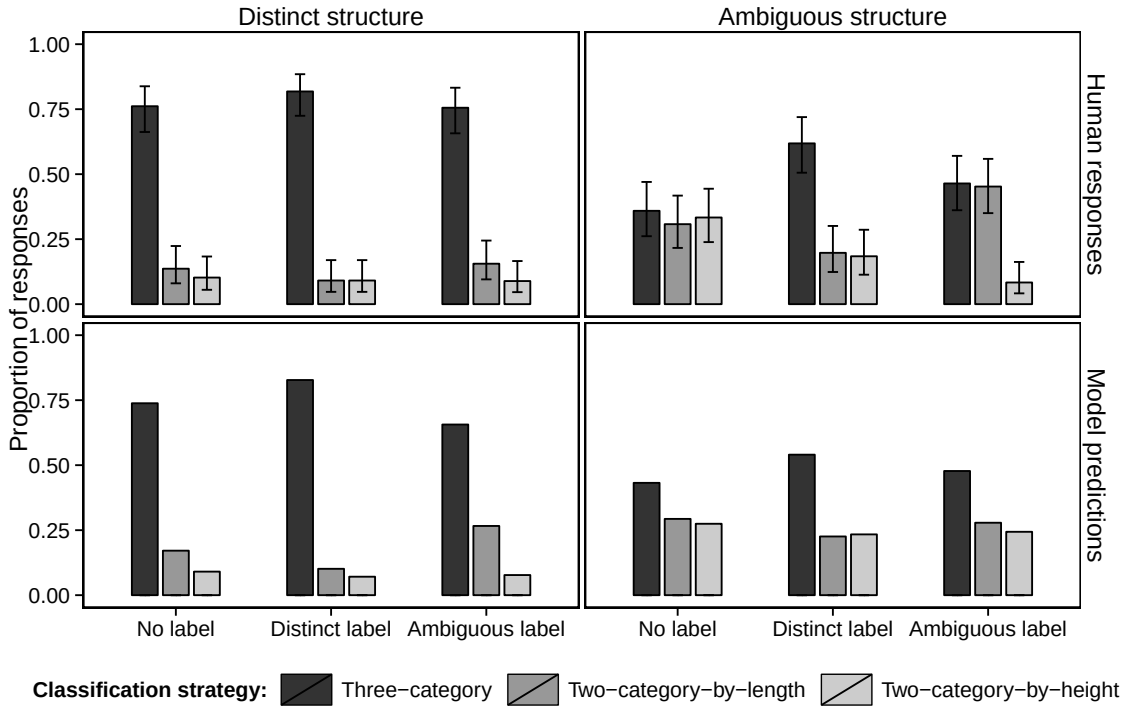

Figure 5. Comparison between the proportion of strategies used by humans and predicted by the Rational model across each of the experimental conditions. Error bars plot 95% confidence intervals for the human responses. People in the DISTINCT STRUCTURE mostly relied on unlabeled information, with labeled examples having little effect in their choice of classification strategy. In contrast, there was a strong effect in how labels were used by people in the AMBIGUOUS STRUCTURE conditions. The rational model of categorization captures people's responses reasonably well in both conditions.

ulus structure and labeling scheme, the interpretation of this interaction effect is simple. When the stimulus structure was unambiguous, providing additional labeled data had no influence on how people learned. In such cases semi-supervised learning looks the same as unsupervised learning. In contrast, when the stimulus structure was ambiguous, even a very small number of labeled examples had a big impact on how people learned, pushing people towards one solution or another depending on the information provided by the labels.

Model fitting

It appears that people produced sensible behavior in this task, but one question remains: can we account for this performance based on standard psychological theories of categorization, or is it necessary to postulate entirely different mechanisms or abilities? To address this question, we applied a modified version of Anderson's (1991) Rational Model of Categorization (RMC) to the task. The RMC is a Bayesian category learning model that has previously been applied to a variety of tasks in supervised learning (Anderson, 1991),

unsupervised learning (Clapper & Bower, 2002; Pothos et al., 2011) and semi-supervised learning (Zhu et al., 2010). We chose to focus on the RMC because it lends itself well to the situation our participants were in: it assumes that stimuli belong to one of several categories, but does not know how many categories exist and so attempts to infer this from the data. However, there is no inherent reason why other successful category learning models such as SUSTAIN (Love, Medin, & Gureckis, 2004) could not also be similarly adapted. The RMC learns the number of categories by relying on a simple sequential assignment method known as the Chinese restaurant process, which specifies the prior probability of a particular category (proportional to the number of items in that category) and the prior probability of a new category (a constant). For a detailed discussion of the RMC in the form we implemented it, see Sanborn, Griffiths, and Navarro (2010).

It was necessary to modify the RMC slightly in order to apply to this task. A critical feature of the RMC is that category labels are viewed as an additional feature possessed by stimuli. From this perspective our task involves two continuous features (height and length) and one discrete one (label). A category is associated with a probability distribution over all three features. In Anderson's (1991) formulation, the number of possible values that a discrete feature can take is assumed to be known in advance. In our task this assumption is inappropriate, since the number of possible labels is not known to the learner. Fortunately this is easy to rectify: we assume that the distribution over labels is itself sampled from a Chinese restaurant process, consistent with the prior distribution over category assignments. Thus, labels of the same type would tend to belong to the same clusters, while items with unseen labels would be more likely to be assigned to new clusters.

Each run of the RMC outputs a set of category assignments for the observed stimuli (directly analogous to the responses we collected from participants). This output was compared to human responses by applying the same procedure that we applied to the human data: assigning each classification to one of the three canonical strategies based on the *adjR* index. Results for each condition reflect 5000 independent runs, with the order that the stimuli were presented to the model randomized between runs.

The output of the RMC, plotted in the bottom row of Figure 5, is qualitatively consistent with the pattern of responses produced by human subjects. For example, in the DISTINCT STRUCTURE conditions, the model predicted that the three category classification would be preferred regardless of the nature of the labels. It also predicted, similarly to people, greater variation in the strategies in the AMBIGUOUS STRUCTURE conditions. There were a few cases where the model predictions did not exactly match the responses given by people, most notably in the AMBIGUOUS STRUCTURE, AMBIGUOUS LABEL condition, where it did not rule out the two category by height classification like people did.[5]

Overall, the correlation between the predictions of the modified RMC and the data from participants in the proportion of responses for each strategy was 0.92. This suggests that despite its imperfections, the RMC is able to roughly reproduce human performance for a novel semi-supervised task. Given that this is the first study that we are aware of that tries to compare semi-supervised learning to unsupervised learning (rather than to supervised learning) and where the number of labels is not known, it is reassuring to see that existing theory generalizes well to this situation.

[5]The model's responses in this condition suggest that this result is driven primarily by runs where it did not observe the labeled instances necessary for correct classification until near the end of the run.

Discussion

Most of the literature on semi-supervised learning takes supervised learning as its starting point, and examines the extent to which additional unlabeled data shifts people's learned category representations relative to people only presented only with labeled data. The results in this area have been mixed, with studies finding that in some situations unlabeled data has an effect in semi-supervised learning (Zhu et al., 2007; Lake & McClelland, 2011; Kalish et al., 2011) and in others where it does not (McDonnell et al., 2012). Our work adopts a very different framing of the semi-supervised learning problem: instead of asking how semi-supervised learning differs from supervised learning, we ask how it differs from unsupervised learning. Instead of asking when unlabeled data have an influence on learning, we investigate when labeled data are helpful.

Our core results bear a superficial similarity to previous work, insofar as our key finding is that labeled data is sometimes helpful, and sometimes it has no effect on learning. However, our experimental manipulations make it clear when and why it happens. When the unlabeled data is informative enough that the category structure is unambiguous, people do not need labeled data to guide learning. As Bloom (2000) suggests, semi-supervised learning appears indistinguishable from unsupervised learning in this scenario. In contrast, when the unlabeled data is ambiguous, labels become more powerful and have a large effect on the categories that people infer – in this case, the specific set of labels shown helps people determine which dimensions are relevant for classification. This includes whether to stick with a simpler unidimensional strategy or to switch to a more complex multidimensional classification strategy. Of course, ambiguous situations may not be the only kind of instance where labeled examples are useful. The results from Vandist et al. (2009) suggest that labeled examples can also help in learning complex Information-Integration categories – in that case, the categories are well-separated and not ambiguous but still require integrating information from multiple dimensions.

The historical prevalence of supervised learning as a topic of interest in cognitive science and machine learning has implicitly taken supervised learning to be the natural reference point against which semi-supervised learning should be assessed. In our view, this assumption also reflects an incomplete view of human semi-supervised learning. The category learning problems people – especially children – face in real life do not usually involve a few unlabeled examples in addition to many labeled ones. Rather, the world naturally presents people with a rich distribution of unlabeled data, which helpful teachers (such as parents) supplement by labeling.

Comparing semi-supervised learning to unsupervised learning sheds light on the critical role that labeled data plays in human learning. In particular, much of the difficulty in how humans learn categories is in the unsupervised aspects of determining how things should be grouped together. Here we argue that labels play a fundamental part in making sense of it, especially when the categories are ambiguous without them. It is an open question to what extent categories in the natural world are ambiguous in this way. Future work should investigate cases where labeled examples are informative in other ways, such as when objects belong to multiple cross-cutting categories (Shafto, Kemp, Mansinghka, & Tenenbaum, 2011) or when items organized into taxonomies have multiple labels (Canini & Griffiths, 2011).

Acknowledgments

Correspondence concerning this article should be addressed to Wai Keen Vong, School of Psychology, University of Adelaide SA 5005, Australia (waikeen.vong@adelaide.edu.au). Daniel J. Navarro received salary support from Australian Research Council grant FT110100431 and Amy Perfors from Australian Research Council grant DE120102378. Research costs were funded through Australian Research Council grant DP110104949.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199.

Bloom, P. (2000). *How children learn the meaning of words.* Cambridge, MA: MIT Press.

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

Canini, K. R., & Griffiths, T. L. (2011). A nonparametric bayesian model of multi-level category learning. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence.*

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning* (Vol. 2). MIT press Cambridge.

Clapper, J. P., & Bower, G. H. (2002). Adaptive categorization in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 908.

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science*, *5*(1), 132–172.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, *120*(1), 106–118.

Lake, B., & McClelland, J. (2011). Estimating the strength of unlabeled information during semi-supervised learning. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1400–1405).

Lewis, M., & Frank, M. C. (2013). An integrated model of concept learning and word-concept mapping. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society.*

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, *111*(2), 309.

McDonnell, J. V., Jew, C. A., & Gureckis, T. M. (2012). Sparse category labels obstruct generalization of category membership. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 749–754).

Medin, D. L., & Ross, B. H. (1997). *Cognitive psychology.* (Second ed.). Harcourt Brace Jovanovich.

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

Medin, D. L., Wattenmaker, W. D., & Michalski, R. S. (1987). Constraints and preferences in inductive learning: An experimental study of human and machine performance. *Cognitive Science*, *11*(3), 299–339.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122.

Nosofsky, R. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*(3), 303–343.

Pothos, E. M., & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition*, *107*(2), 581–602.

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*(1), 83–100.

Rogers, Kalish, C., 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 Conference of the Cognitive Science Society* (pp. 2320–2325).

Rogers, T., Kalish, C., Harrison, J., Zhu, X., & Gibson, B. R. (2010). Humans learn using manifolds, reluctantly. In *Advances in Neural Information Processing Systems* (pp. 730–738).

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1–25.

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*(2), 328–341.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Zhu, X., Gibson, B. R., Jun, K.-S., 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* (pp. 1247–1254).

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 864–870).

Appendix

As discussed in the main text, our version of the RMC does not assume the learner knows the number of possible labels in advance, and – much like the number of categories

itself – uses a Chinese restaurant process to capture this uncertainty (see Navarro, Griffiths, Steyvers, & Lee, 2006, for an overview). This implies that the probability of observing the jth label for a stimulus belonging to the kth category is

$$P(\text{label } j \mid \text{category } k) = \frac{n_{jk}}{n_{\cdot k} + l}$$

where l is a parameter that governs the learner's willingness to tolerate differently labeled items within the same category (fixed at $l = 1$ in all simulations). In this expression n_{jk} denotes the number of times the label j has been observed in cluster k, and $n_{\cdot k}$ is the total number of labeled examples assigned to category k. Relatedly, the probability of observing a new label for an item in category k is

$$P(\text{new label} \mid \text{category } k) = \frac{l}{n_{\cdot k} + l}$$

For unlabeled data, a complete solution would be to have the model treat the label as missing data, and to try to infer those labels via Bayesian inference by sampling from the posterior in the same way that the model infers the category assignment. In our applications we adopt a simplification in which the model simply computes the expected prior probability of the label that should have been assigned to that observation, integrating over all possible values for that label. For an item assigned to category k, this is given by:

$$\begin{aligned} P(\text{unlabeled} \mid \text{category } k) &= E_{P(\text{label} \mid \text{category } k)}[P(\text{label} \mid \text{category } k)] \\ &= P(\text{new label} \mid \text{category } k)^2 + \textstyle\sum_j P(\text{label } j \mid \text{category } k)^2 \\ &= \left(\frac{l}{n_{\cdot k}+l}\right)^2 + \textstyle\sum_j \left(\frac{n_{jk}}{n_{\cdot k}+l}\right)^2 \\ &= \frac{l^2+\sum_j {n_{jk}}^2}{(n_{\cdot k}+l)^2} \end{aligned}$$

This approach is an approximation (sufficient for our purposes) that uses the prior to integrate out the learner's uncertainty about the identity of the missing labels, and is considerably simpler than the full Bayesian solution that would use the full joint posterior distribution over all unobserved quantities to achieve the same end.

In all other respects, including parameter values, the model we used is identical to the version of the RMC described by Sanborn et al. (2010), in which we used Markov chain Monte Carlo (MCMC) methods to approximate Bayesian inference within the model. For each condition, the adapted RMC was run 5000 times, randomizing the order of the stimuli presented to the model each time.