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

Wai Keen Vong (waikeen.vong@adelaide.edu.au)
Amy Perfors (amy.perfors@adelaide.edu.au)
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Abstract

The study of semi-supervised category learning has shown mixed results on how people jointly use labeled and unlabeled information when learning categories. Here we investigate the possibility that people are sensitive to the value of both labeled and unlabeled items, and that this depends on the structure of the underlying categories. We use an unconstrained free-sorting categorization experiment with a mixture of both labeled and unlabeled stimuli. The results showed that when the distribution of stimuli involved distinct clusters, participants preferred to use the same strategies to sort the stimuli regardless of whether they were given any additional category label information. However, when the stimuli distribution was ambiguous, the sorting strategies people used were strongly influenced by the labeled information given. We capture performance in both cases with an extension to Anderson's Rational Model that does not know the exact number of category labels in advance.

Keywords: Semi-supervised learning, unsupervised learning, categorization, Bayesian modeling

Introduction

How do people acquire knowledge about concepts and categories? As Gibson, Rogers, and Zhu (2013) have recently argued, traditional supervised and unsupervised approaches provide insufficient and ecologically implausible explanations for real world category learning. In the real world, learners are not provided with labeled category information with every object they encounter (like in supervised category learning tasks), nor do they receive only unlabeled information (like in unsupervised category learning tasks). Rather, people learn through a mixture of both labeled and unlabeled information. This is known as *semi-supervised category learning*.

Previous research has suggested the rapid acquisition of concepts and categories from a few sparse labeled examples is the result of strong inductive biases (Xu & Tenenbaum, 2007). Semi-supervised learning presents another possibility: people are able to use the large amount of unlabeled information they receive to determine how to organize objects into categories, and in conjunction, combine this with sparse labeled information to map linguistic tokens onto these category representations (Bloom, 2002).

So far however, the empirical evidence regarding semi-supervised learning has been mixed, with only some empirical evidence showing that the presence of additional unlabeled examples affects categorization behaviour. A number of studies have found that in one-dimensional classification tasks, presenting labeled examples that differ from the distribution of unlabeled examples causes a shift in people's estimates of the category boundary (Lake & McClelland, 2011; Kalish, Rogers, Lang, & Zhu, 2011; Zhu, Rogers, Qian, & Kalish, 2007).

However, other studies involving two-dimensional stimuli found no evidence of semi-supervised learning (Vandist, De Schryver, & Rosseel, 2009; McDonnell, Jew, & Gureckis, 2012). In these experiments, participants who were given additional unlabeled information responded like participants in the supervised conditions, effectively ignoring any additional unlabeled information that was given. This raises the possibility that (for whatever reason) the role of labels depends on the dimensionality of the categorization task; yet even here, the evidence is mixed: Rogers, Kalish, Gibson, Harrison, and Zhu (2010) showed evidence of semi-supervised learning in another 2D categorization task, but only when participants were required to respond rapidly. This pattern of results suggests that the value of labeled and unlabeled information is not fixed as suggested by Lake and McClelland (2011), but is dependent on the nature of the task. However, the precise nature of this dependence is not well understood.

One issue with previous semi-supervised learning studies is that they have focused only on semi-supervised *classification* tasks in which learners are presented with both labeled and unlabeled examples, and then asked to classify novel examples into one of the labeled categories. This paradigm assumes a fixed number of categories, but in the real world people must infer how many categories there are, based on a limited number of observed category labels.

This is the problem of category discovery, in which the learner must decide when to form a new category based on what they have previously learned about other categories (Bruner, Goodnow, & Austin, 1956; Pothos et al., 2011). When only some of the items are labeled, the problem becomes more difficult. For example, suppose a child might see many spoons and forks, some labeled and some not. However, they also see a few unlabeled examples of chopsticks. Will the child realize that chopsticks are neither spoons nor forks, but rather belong to an entirely separate category, without having heard the label? If people are able to use the underlying category structure to shape their generalizations, they should recognize in this case the chopsticks belong to a new category.

In principle, categorization models that view semi-supervised category learning as being governed by the same underlying process as supervised and unsupervised learning should be able to capture this behavior by jointly using labeled and unlabeled information to discover the correct number of categories. However, while a number of categorization models have been successfully fitted to previous empirical studies of semi-supervised learning (Zhu et al., 2010; Gibson et al., 2013), they have only tackled the problem of semi-supervised classification with a fixed number of labeled categories. So far, there have not been any attempts to show

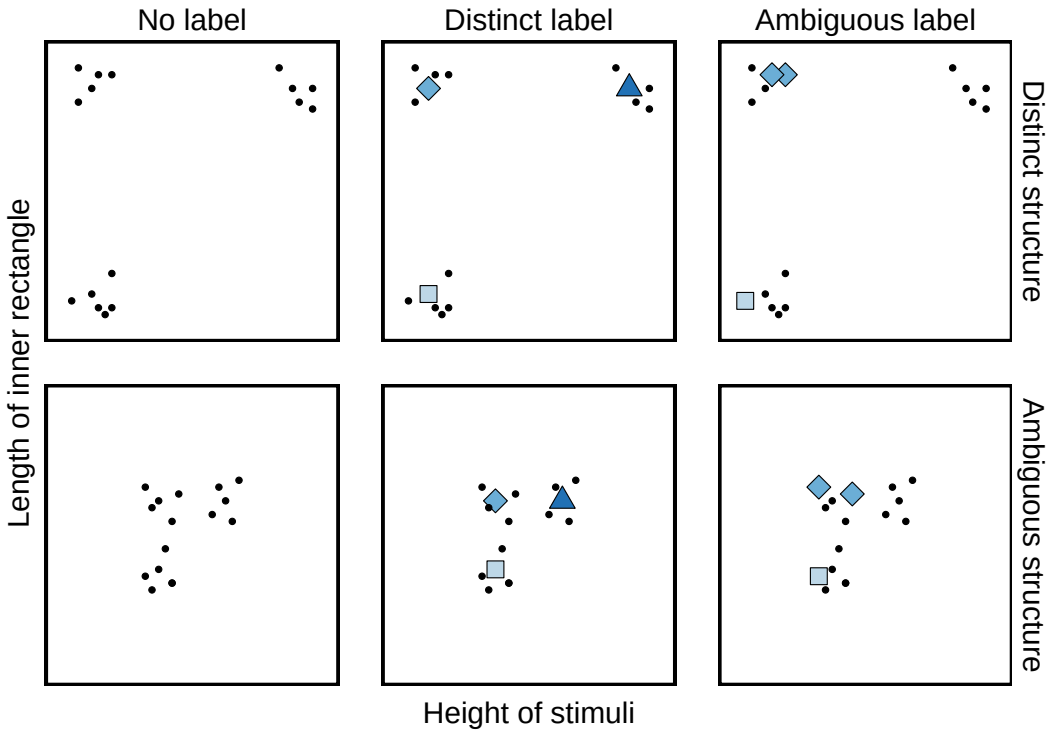

Figure 1: A visualization of the experimental design. The stimuli varied along two continuous dimensions (stimuli height and the length of the inner rectangle). The small black dots represent the unlabeled examples, while the larger stimuli represent the labeled examples, with each shape corresponding to a different category label (dax, wug or fep). Note that while participants in both labeled conditions saw three labeled examples, participants in the distinct and ambiguous label conditions saw either three or two different types of labels respectively.

that semi-supervised category learning models can jointly use labeled and unlabeled information to infer the number of categories and to discover new categories.

It is difficult to present people with an unknown number of categories within the classification paradigm that has previously been used to study semi-supervised learning. The current study therefore deviates from previous studies and uses an unconstrained free sorting task instead. Free sorting tasks present all of the stimuli together, with the goal of sorting the stimuli into coherent categories. Such tasks have been previously used for studying unsupervised categorization (Pothos & Close, 2008; Pothos et al., 2011), but none have been used to explore semi-supervised categorization where sorting behaviour may be influenced by additional labeled examples.

In this study we investigate how people jointly use labeled and unlabeled information in a sorting-based semi-supervised task, with the aim of reconciling some of the existing confusion about the role of labeled and unlabeled information in semi-supervised learning. Our main question is how the use of both labeled and unlabeled information depends on the underlying category structure. Are there situations where people tend to use one kind of information over the other? If so, why? Our results show that people use both types of information, but to different extents depending on how distinct the categories are and the labels they are given. We also demonstrate that people's sorting behavior can be accounted for by an extension of the Rational Model of Categorization (Anderson, 1991) in which the number of category labels is not preset.

Method

Participants

590 participants were recruited from Amazon Mechanical Turk and paid $0.30 or $0.50 for their participation. Participants were randomly assigned to each of the experimental conditions, and the task took roughly five minutes to complete. 86 people were excluded for either not finishing the task (34) or providing a response in the sample trial that revealed they did not understand or were not engaged with the task, as described below (52), leaving 504 participants in the experiment.

Materials

The stimuli consisted of white rectangles with an inner gray rectangle in the bottom-right corner. These stimuli were mapped along two continuous dimensions corresponding to the height of the white rectangle (which ranged between 25 to 65 pixels) and the length of the inner gray rectangle (which ranged between 10 and 50 pixels). In the experiment, people were presented with 16 different stimuli on the screen together, with the stimulus values dependent on two different experimental manipulations. Some stimuli were labeled with a nonsense word (*dax*, *fep*, or *wug*) appearing underneath.

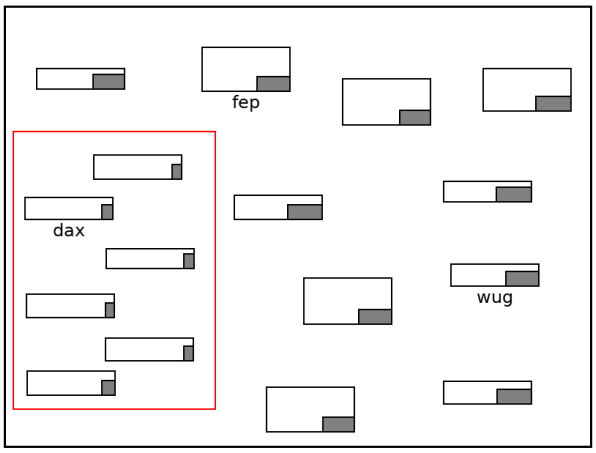

Figure 2: An example of the experimental task demonstrating the stimuli and labels used in the task. People were free to drag the stimuli around and sort them into categories. This figure illustrates the DISCRETE STRUCTURE with DISCRETE LABELS condition, with some of the stimuli grouped into one category.

We manipulated the labeled and unlabeled information given to participants in two different ways to examine their effects on categorization behaviour. The first manipulation involved different **category structures**, which are shown in Figure 1. In the DISTINCT STRUCTURE condition, there were three well-separated equally sized clusters that varied along both stimulus dimensions. The AMBIGUOUS STRUCTURE condition also consisted of three equally sized clusters, but were much closer together in the stimulus space, making it hard to distinguish the cluster boundaries. None of the participants were told that there were three underlying categories: they were simply instructed to sort the stimuli into as many categories as they wanted.

The second manipulation, also shown in Figure 1, involved varying the **category labels** that were presented with the stimuli. The NO LABEL condition corresponded to an unsupervised version of the task in which people saw no labels. In the DISTINCT LABEL condition, there was one label located near the center of each of the three clusters. Conversely, in the AMBIGUOUS LABEL condition, one label was from the first cluster, two were from the second, and none were from the third.

Procedure

The experiment began with instructions describing how archaeologists had discovered a number of unknown objects and needed help in sorting them into different categories. Those in the labeled conditions were additionally told that the archaeologists had discovered the names of some of the objects to help them out.

Before performing the main task, participants were given a sample trial to familiarize themselves with the free sorting interface. In the sample trial, they were shown three squares and three triangles of various sizes and asked to sort them into different categories. They were then asked to draw selection boxes around each pile to identify the separate categories. Once they were happy with their sort, they submitted their response. 52 participants failed to categorize the sample stimuli in a sensible way (either by shape or size). On the assumption that this reflected misunderstanding or non-engagement with the task, their data were excluded from analyses. The procedure during the main task was identical to the sample task, but the stimuli were the rectangles described above, whose feature values and category labels depended on which condition the participant was in. Figure 2 illustrates the task in the DISTINCT STRUCTURE and DISCRETE LABEL conditions.

Model

To model performance in the task we extended Anderson's (1991) Rational Model of Categorization (RMC), a highly successful model of categorization that has captured many findings from both supervised and unsupervised learning (Anderson, 1991; Pothos et al., 2011). Gibson et al. (2013) have recently adapted the RMC to handle semi-supervised learning, to make use of additional information from unlabeled examples. Additionally, while the RMC can grow flexibly with the number of clusters in the data using the Chinese Restaurant Process prior, it treats category labels as a discrete feature that can only handle a fixed number of labels. However, in our task the number of category labels is unknown. To account for this, we modified how the likelihood of labeled features were calculated to account for uncertainty over the kinds of possible labels.

In order to calculate the likelihood of category labels, we assume that each cluster k has a Chinese Restaurant Process prior that keeps track of the labeled examples in each cluster (see Sanborn, Griffiths, and Navarro (2010)). In addition, we specify a label concentration parameter l which measures how likely we expect to see new labels in a given cluster (which we set as 1 in each cluster for simplicity purposes). Thus, the likelihood of observing an existing label j in cluster k is given by:

$$p(j_{\text{existing}}|k) = \frac{n_{j,k}}{(N_k + l)}$$

where $n_{j,k}$ represents the number of times the label j has been observed in cluster k, and N_k represents the total number of labels observed in cluster k. Similarly, we can calculate the likelihood of observing a new label for a given cluster k by:

$$p(j_{\text{new}}|k) = \frac{l}{(N_k + l)}$$

What about the likelihood of unlabeled examples? Rather than ignoring the missing category label (and assigning it a likelihood of 1), we can attempt to impute the missing label using the observed label frequencies for a given cluster. That is, the likelihood of an unlabeled stimulus is given by the probability of its true label might be (using the observed frequencies), multiplied by the likelihood of observing each label and summing across all possible labels (including un-

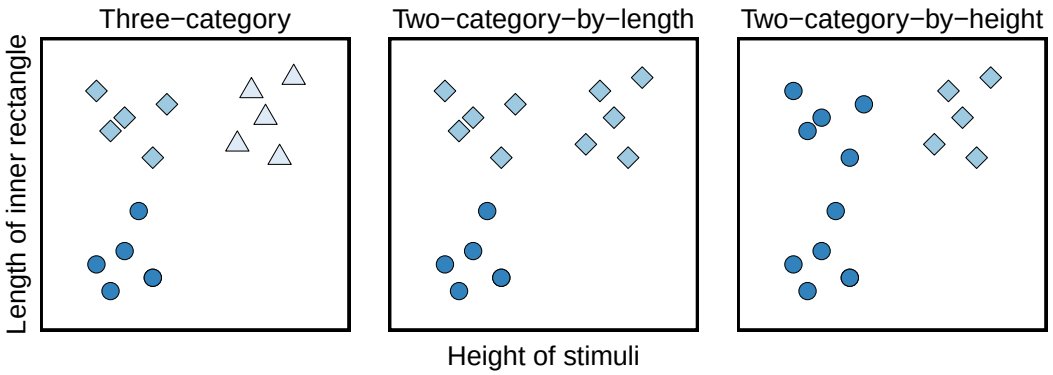

Figure 3: The three canonical classifications used to classify people's responses in the task. While this figure only depicts the canonical classifications for the AMBIGUOUS STRUCTURE condition, the strategies are analogous for the DISTINCT STRUCTURE condition. The three-category strategy required attending 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.

observed labels).[1] This simple imputation method means that the likelihood of an unlabeled item in cluster k is given by:

$$p(j_{\text{unlabeled}}|k) = \frac{l^2 + \sum_j n_{j,k}^2}{(N_k + l)^2}$$

For new observations, the likelihood of labeled examples where the label has been previously observed, as well as unlabeled examples should both be favored over new labels. By modifying how the model treats existing labeled examples, new labeled examples and unlabeled examples, our extension to the RMC influences how likely different clusterings of the given stimuli are.

Results

One difficulty with unconstrained categorization tasks is identifying a useful measure of performance. Previous work have used the frequency of the preferred classification (Pothos et al., 2011), and although this measure is useful in examining the consistency between responses, it does not deal well when the distribution of sorting strategies is multi-modal. An initial analysis of the responses showed that most participants followed one of three different strategies, schematically shown in Figure 3: a three category classification along both stimulus dimensions, or a two-category classification along one of the dimensions (either length or height).

A large number of participants employed one of these three strategies, but many opted for classifications that were close to but not exactly any of them. We therefore measured the extent to which each person's classification matched to each of the three canonical strategies using the Adjusted Rand Index (*adjR*), which calculates the similarity between two different classifications (Hubert & Arabie, 1985). When two classifications are identical the *adjR* value is 1, and when they match at chance levels it is 0. For each person we calculated the three *adjR* values comparing their classification to each of the three canonical ones, and took their strategy to be the canonical one that produced the highest *adjR* value.[2]

The top row of Figure 4 shows the histogram of strategies used among the different conditions. We found that participants in the DISTINCT STRUCTURE strongly preferred a three category classification regardless of the presence or nature of the category labels. Indeed, no significant differences were observed between the different **category label** conditions within the DISTINCT STRUCTURE condition ($\chi^2(4) = 1.90, p = 0.75$). Even when one cluster of stimuli was given no labels at all, as in the AMBIGUOUS LABEL condition, people still strongly preferred to sort that cluster into its own separate category. These results suggest that people were primarily using the underlying structure to discover the nature of the categories, and did not need each category to be labeled separately.

In contrast, when the underlying category structure was ambiguous people varied substantially in their classification strategies depending on the presence and nature of the labels. As Figure 4 shows, when participants were given no labels they were equally likely employ any one of the three canonical strategies. However, and in contrast to the DISTINCT STRUCTURE condition, the presence of additional label information guided people to different strategies depending on the nature of the labels. In the DISTINCT LABEL condition, most participants preferred a three category classification strategy, ruling out either of the two category strategies as being inconsistent with the labels presented (by having two different labels in the same category). Similarly, people in the AMBIGUOUS LABEL condition ruled out the inconsistent two-category-by-height strategy, and their responses were equally split among the remaining two canonical strategies. Overall, we found a significant difference between the

[1] More precisely, we impute missing values by replacing them with their expected value. In this case the thing that is treated as missing is the probability of the object label. We also considered applying stochastic methods to impute the missing label itself, but it seems unlikely that this more complex approach would make much difference to the RMC predictions.

[2] While a number of participants used other sorting strategies, accounting for this with an "other" strategy variable produced no qualitative differences for the analyses in this study.

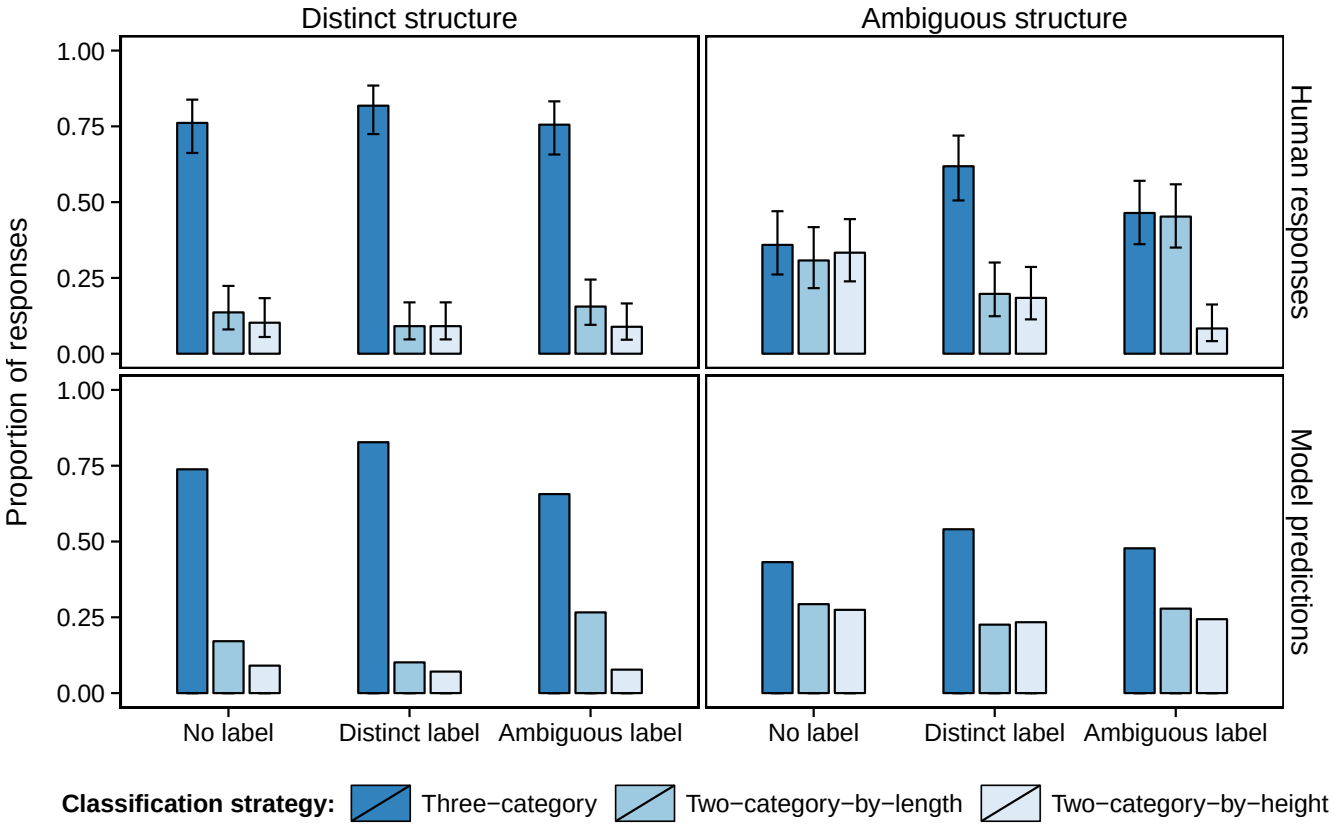

Figure 4: Comparison between the proportion of strategies used by humans and predicted by the Rational model across each of the experimental conditions. Error bars plot 95% confidence intervals for the human responses. People in the DISTINCT STRUCTURE mostly relied on unlabeled information, with labeled 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 to guide their classification. The Rational model was able to capture most of the results seen in the human data.

category label conditions in the AMBIGUOUS STRUCTURE ($\chi^2(4) = 26.48, p < .001$).[3]

We also applied our extended RMC to the current task. Each condition was run for 5000 iterations, randomizing the stimuli order presented to the model, and then classifying the model's response to one of the three canonical strategies using the same method as we did for people.[4] The results, as shown in the bottom row of Figure 4, demonstrate that the model qualitatively matched many of the response patterns found in people. In particular, the model was able to recognize that the structure of examples in the DISTINCT STRUCTURE was more useful in determining how to classify responses. Likewise, for the AMBIGUOUS STRUCTURE, the model relied less on the distributional information and was able to use labeled information when sorting the stimuli, with the exception of the AMBIGUOUS STRUCTURE, AMBIGUOUS LABEL condition, where the model's predictions deviated slightly from how people responded. Overall, the average correlation between the model's predictions and the human responses was 0.92, suggesting that the modified RMC was able to perform semi-supervised learning, even when the number of category labels was not fixed.

Discussion

This study investigated the effect of the underlying category structure and the nature of the labels on people's semi-supervised learning with an unconstrained free-sorting paradigm. Our results suggest that the information provided by the category structure and the category labels are both useful, and their utility depends on how distinct or ambiguous each is. In addition, we presented an extension of the RMC that can account for uncertainty over feature labels which was able to capture the kinds of categorization strategies people used, suggesting that an existing category learning model can handle a novel categorization task in semi-supervised learning.

Our main result helps to reconcile previous work in semi-

[3] Significant differences were also observed between each pair of **category 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: $\chi^2(2) = 12.69, p < .01$).

[4] In contrast to Gibson et al. (2013), we used the grouping over categories as the model response, rather than predicting the category label along the label dimension, as this was better suited to our task.

supervised learning which found mixed effects of the role of labels and unlabeled examples. We find that the effectiveness of labeled and unlabeled information is mediated by the underlying category structure. When the clusters were sufficiently distinct, labels did not add much: people were easily able to organize them into the obvious clustering regardless. On the other hand, when the underlying distribution of objects was less clear, people used the labeled information as a cue. These results suggest that people are sensitive to the value provided by both labeled and unlabeled information, and can adjust their behaviour in categorization tasks accordingly.

Many of the cases where people used additional unlabeled information in categorization in earlier studies were in tasks that used one-dimensional stimuli, suggesting that the simpler category structure may have been easier for participants to attend to the underlying distribution of stimuli (Zhu et al., 2007; Lake & McClelland, 2011; Kalish et al., 2011). On the other hand, for tasks with two-dimensional stimuli where the category structure was less obvious, found little benefit to unlabeled information (Vandist et al., 2009; McDonnell et al., 2012). Our results suggest that people in these tasks may have had a preference to attend to labeled information, rather than attending to the category structure along both stimulus dimensions in unlabeled information. Why were people able to use both dimensions to sort the categories in our task then? One possible explanation is that free sorting tasks require less working memory to remember all of the stimuli, and also allows for comparisons between the stimuli. This contrasts with the strong preference for classification rules that lie along a single dimension in classification designs (Ashby, Queller, & Berretty, 1999). Alternatively, and in line with the results of Pothos and Close (2008), sorting objects into categories using both dimensions may have produced more intuitive categories than sorting along a single dimension.

Semi-supervised category learning is a fundamental part of human experience and yet much understudied relative to supervised and unsupervised category learning. While our design differed from previous studies in human semi-supervised learning, it presents results that help to reconcile mixed results in the semi-supervised literature, as well as a computational model that can capture human performance in a sorting task in which the number of labels is not pre-defined. Much remains to be done before this important aspect of category learning is understood, but this work is one step along that road.

Acknowledgments

This research was supported by ARC grant DP0773794. DJN received salary support from ARC grant FT110100431, and AP from ARC grant DE120102378.

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. (2002). *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.

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).

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).

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).

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.

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).