Sparse category labels obstruct generalization of category membership

John V. McDonnell (john.mcdonnell@nyu.edu)
Carol A. Jew (carol.jew@nyu.edu)
Todd M. Gureckis (todd.gureckis@nyu.edu)
New York University, Department of Psychology
6 Washington Place, New York, NY 10003 USA

Abstract

Studies of human category learning typically focus on situations where explicit category labels accompany each example (supervised learning) or on situations were people must infer category structure entirely from the distribution of unlabeled examples (unsupervised learning). However, real-world category learning likely involves a mixture of both types of learning (semi-supervised learning). Surprisingly, a number of recent findings suggest that people have difficulty learning in semi-supervised tasks. To further explore this issue, we devised a category learning task in which the distribution of labeled and unlabeled items suggested alternative organizations of a category. This design allowed us to determine whether learners combined information from both types of episodes via their patterns of generalization at test. In contrast with the prediction of many models, we find little evidence that unlabeled items influenced categorization behavior when labeled items were also present. **Keywords:** Semi-supervised category learning; rule induction; unsupervised learning

Introduction

Category learning is a critical cognitive ability which is central to many aspects of cognition. As a result, considerable research over the last 50–60 years has explored the psychology of category learning using laboratory tasks. The majority of this work can be divided into two groups. Most research has focused on *supervised learning* tasks where corrective feedback or category labels are presented following or alongside each observation of a stimulus (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). More recently, there has been an interest in *unsupervised learning*, wherein participants must organize examples in the absence of explicit instruction using the distributional properties of the stimuli (e.g., Clapper & Bower, 1994; Love, 2002; Pothos et al., 2011). However, neither of these situations adequately reflect the problem of real world category learning, in which feedback is not altogether absent nor always present, but is typically sparse and intermittent. Such tasks require learners to combine information from both labeled and unlabeled episodes. In machine learning, this problem is frequently studied under the name *semi-supervised learning* (for review, see Zhu, 2005).

Aside from offering a more ecologically relevant approach to the study of category learning, the study of semi-supervised learning has important implications for theories of human concept learning. Consider the problem of learning a concrete noun such as *horse*. One proposal is that word learning essentially links sound tokens (words) to already-acquired hypotheses or representations (Bloom, 2000; Gentner, 1982). Under this view, the label information from a teacher or parent about a single example horse must be integrated with the child's pre-linguistic grouping of objects in their environment into classes.

A similar position is advocated by a number of influential theories of category learning which hold that supervised and unsupervised learning are subserved by a single underlying learning process (e.g., the rational model of categorization, Anderson, 1991; or the Supervised and Unsupervised STratified Adaptive Incremental Network, abbreviated SUSTAIN, Love, Medin, & Gureckis, 2004). Such models naturally predict that semi-supervised learning should not only be possible, but may be the primary way in which people learn categories and their respective names.

Can people acquire categories via semi-supervised learning?

Despite these arguments, recent empirical attempts to demonstrate semi-supervised category learning in the lab have met with mixed success. For example, Vandist, De Schryver, and Rosseel (2009) found that adding unlabeled training examples to a mostly supervised task offered no additional benefit beyond learning from only the supervised trials. However, the category structures they tested (known as *Information-Integration* categories) are typically difficult for people to learn even in fully unsupervised settings (Ashby, Queller, & Berretty, 1999), which may explain the limited impact that the unlabeled examples had.

On the other hand, Kalish, Rogers, Lang, and Zhu (2011) showed that after learning a simple category distinction on a single dimension from a small set of labeled examples, participants' estimate of the category boundary could be shifted by the presentation of a large number of unlabeled examples whose distribution was shifted compared to the labeled set (see also Lake & McClelland, 2011). While this study provides some evidence of semi-supervised learning, there remain alternative explanations of the effect. For example, since the central tendency of both categories are shifted in these studies it is unclear whether people are separately updating each category representation or responding to the global shift in the stimulus space.

Finally, Rogers, Kalish, Gibson, Harrison, and Zhu (2010) compared learning in a semi-supervised learning condition with a fully supervised condition. In this study, adding unlabeled items to a supervised category learning task caused faster learning only when trials were speeded. However, the question of whether people can integrate labeled and unlabeled training examples is logically separate from claims about

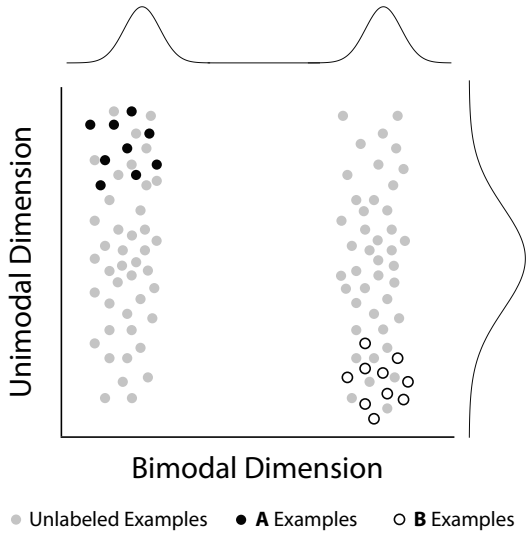

Figure 1: A schematic depiction of the design used in the experiment. Category stimuli varied along two continuous dimensions. The plot edges represent the marginal distribution of examples. Unlabeled examples fall in two columnar clusters, while two clusters of examples with labels *A* and *B* appear in the corners of the space. Taken alone, the distribution of labeled examples is ambiguous concerning how to generalize, since a rule on either dimension alone could explain the labels.

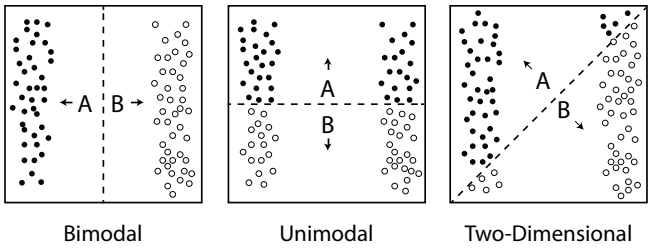

Figure 2: A range of possible category strategies consistent with the category in Figure 1. In the Bimodal strategy, the learner classifies all the items that fall in each clustered column with the label given to the labeled items within the column. In the Unimodal strategy, the learner divides the example along the unimodal dimension. This strategy is acceptable if the unlabeled examples are ignored. The 2D strategy is more complex (in the sense that it depends on attention to both stimulus features) but is also consistent with the labeled examples and inconsistent with the unlabeled distribution.

learning rates between tasks. For example, a participant's learning rate might vary based on features of the overall task context rather than the information conveyed by any subset of examples.

Collectively, these results tell a surprisingly unclear story. Despite decades of research on supervised and unsupervised learning with artificial stimuli, studies which have attempted to combine these two forms of learning fail to show robust and consistent effects. Some find limited evidence of semi-supervised learning while others fail to find any evidence at all. The goal of the present study is to attempt to revisit this issue with a novel design which may be more diagnostic of semi-supervised learning. As will be revealed shortly, our results add modest light to an already murky picture.

Evaluating semi-supervised learning through patterns of generalization at test.

Our study (summarized abstractly in Figure 1) departs from the studies described above in a number of ways. In some previous work, the distributional properties of both labeled and unlabeled examples were identical (e.g., Vandist et al., 2009). In contrast, we manipulated the distribution of examples so that the distribution of unlabeled examples and the distribution of labeled examples suggested alternative organizations of the category. In particular, the labeled items alone were ambiguous about the basis for the category difference. However, the distribution of unlabeled examples suggested a clear organization of the categories along a single dimension. Our prediction was that if people combine information from both the labeled and unlabeled examples, they will generalize the label information according to the distribution implied by the unlabeled examples. This should be clearly captured in their patterns of generalization in a test phase (see Figure 2).

In addition to comparing semi-supervised learning to fully supervised learning we also include a second control condition assessing fully unsupervised learning. In fact, this condition of our study represents a conceptual replication of a previous study on unsupervised category learning by Zeithamova and Maddox (2009, henceforth referred to as Z&M). This served two purposes. First, it allows us to establish a baseline measure of behavior for both extremes of supervision. Second, this ensures that participants *can* learn the category from the distribution of unlabeled examples alone.

Finally, rather than test a single semi-supervised learning condition, we systematically explore the effect of the *number* of labeled examples on semi-supervised learning. Our design thus interpolates between fully unsupervised learning to fully supervised learning by changing the relative amount of labeled versus unlabeled information.

The Experiment

In our experiment, we developed a cover story which provided a plausible explanation for why some category examples were unlabeled (but still came from the same category). The cover story asked participants to imagine that they were working in a TV repair shop in a town where people tuned special loop antennas to pick up one of only two possible channels (similar to Markant & Gureckis, 2010). Similar antennas tended to pick up similar channels. Although all the TVs were tuned to one of the two channels, many had broken tubes making it impossible to turn on and verify the channel. The participants' job over the course of the experiment was to determine how different settings of the antennas determine which channel the TV is tuned to pick up. They were reminded that learning about the antennas was possible even if the TV tube was broken.

The experiment was organized into two phases. The first was a training phase in which participants were shown various category members with and without labels (depending on condition). The second was a test phase in which participants were asked to classify novel examples. Decision bound models (Ashby, 1992) were fit to subjects' responses during the test phase in an attempt to infer the strategy they applied. We then analyzed the frequency by which different strategies were adopted as a function of condition.

Methods

Participants 124 New York University undergraduates participated for course credit. Participants were randomly assigned to one of four possible conditions: Unlabeled (N = 33), 10-Labeled (N = 31), 40-Labeled (N = 30), or 40-All-Labeled (N = 30). Four participants, three in the Unlabeled Condition and one in the 10-Labeled Condition, were classified as responding randomly (see the results section) and were dropped from the analysis, leaving 30 participants in each condition.

Materials The objects to be categorized were line stimuli varying in their length and orientation. The stimulus properties (lengths and angles) of the antennas were chosen to be similar to those used by Z&M (2009). The range of possible angles was different for each subject, but it covered 60° and was constrained not to cross the vertical or horizontal axes. The range of lengths was always between 100 and 560 pixels. The line stimuli were attached to pictures of TV via a stem. Category label information was given by changing what was showing on the TV screen. For unlabeled examples, the screen took on the appearance of broken glass. Participants were told that these TVs were broken, but still tuned correctly to one of the two channels. When category label information was given, the letters CH1 or CH2 appeared on the screen, indicating that the TV was set to pick up one of the channels (see the top row of Table 1 for examples).

Design During the training phase, the TVs were drawn from two elongated distributions which were naturally separable along one of the stimulus dimensions (the *bimodal* dimension). For reasons of control, stimuli were sampled using a discrete binning method described in Figure 3. This differs slightly from Z&M (2009) who used bivariate normal distributions but was necessary to ensure tight control over the distributional properties of the stimuli, and in particular to ensure that the distributions of labeled items were unbiased with respect to the particular categorization strategies.

Our primary experimental manipulation was to alter the training that participants received in the task. Four training conditions ranging from completely unsupervised to completely supervised (with no unlabeled training items) were included (see Table 1 for a summary).

Unlabeled Condition. In this condition, all TVs were broken (i.e., unlabeled). This condition is a traditional unsupervised category learning task and a conceptual replication of the *intermixed* condition from Z&M (2009), Exp. 1A.

10-Labeled Condition. This condition was identical to the Unlabeled Condition except that ten of the items in the corners were presented along with category labels (i.e., the appropriate channel).

40-Labeled Condition. This condition was similar to the Unlabeled Condition and the 10-Labeled Condition except that all of the items in the corners (40 in total) were presented along with category labels.

40-All-Labeled Condition. In this condition, all antennas were labeled in the training phase, meaning that this condition was fully supervised. However, to hold other aspects of the task consistent with the other conditions, 240 broken TVs without antennas (sham trials) took the place of the unlabeled examples, giving participants in this condition the same number of training trials as those in other conditions, and similar temporal spacing between labeled items to participants in the 40-Labeled Condition.

	Labeled	Unlabeled	Sham	Total
Unlabeled	0	280	0	280
10-Labeled	10	270	0	280
40-Labeled	40	240	0	280
40-All-Labeled	40	0	240	280

Table 1: Summary of the four training conditions. All participants viewed 280 items in the training condition. *Labeled* and *Unlabeled* here denotes a TV with an antenna, which were either working (labeled) or broken (unlabeled). A *sham* TV consisted of a broken TV set without an antenna (examples are provided along the top).

Regardless of condition, labeled items in the training phase always came from the corners of the space as depicted in Figure 3, which meant they were always non-diagnostic with respect to the best category rule.

The test phase was identical for all four groups and, following Z&M (2009), involved the presentation of 50 broken TVs sampled from the same distribution as used during training. Participants were asked to predict the channel based on the antenna setting.

All remaining arbitrary aspects of the design (e.g., which dimension served as the bimodal dimension) were counterbalanced across conditions.

Procedure The experiment was administered on standard Macintosh computers using an in-house data collection system written in Python[1]. Participants were tested over a single one-hour session.

The instructions emphasized that all of the antennas were in good working order and purposefully tuned either to CH1 or CH2. Participants were also told that, as a result, the antennas as a whole constituted two categories of items. Although many of the TVs were broken, being broken had nothing to do with the setting of the antenna or the potential to pick up one of the two channels. Broken TVs were missing some information, but were otherwise not different from the others. To confirm that they had understood the instructions, participants were given a brief quiz and misconceptions were addressed.

Next, participants observed 80 randomly generated antennas in quick succession (100ms each), giving information about the range of values for each of the two stimulus dimensions (angle and length).

On each trial of the training phase, participants viewed a new TV (which was broken, working, or a sham), and after 500ms were prompted to press the space bar to continue. After the button press, the stimulus remained on the screen for 500ms. Between trials the screen was blank for an inter-stimulus interval of 500ms.

The test phase consisted of 50 trials in which participants saw a broken TV drawn from the same distribution as the training trials. On each test trial, participants viewed a new broken TV and were asked to press a button on their keyboard to indicate whether they believed the antenna was tuned to CH1 or CH2. After each trial, a thank you message (along with the original stimulus) remained on the screen for 1000ms. No feedback was given. The next trial followed after 500ms.

Results

Accuracy Analysis In our first analysis, we considered whether participants correctly applied the category labels in the unambiguous regions of the space (i.e., the corners) in the 10-Labeled, 40-Labeled, and 40-All-Labeled conditions. Responding was significantly above chance in all conditions: 10-

[1] Available at `http://www.pypsyexp.org`

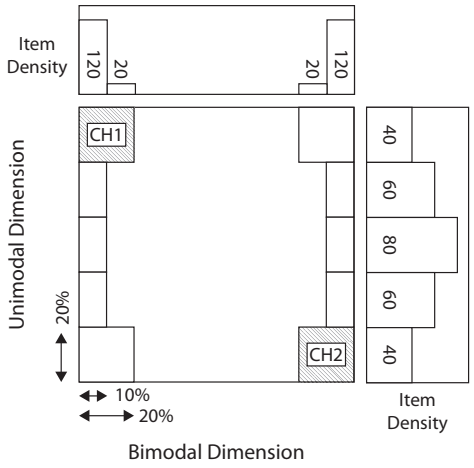

Figure 3: A graphical depiction of the distribution of items in the task. The two dimensions correspond to the size or angle of the stimuli (counterbalanced across participants). Each columnar distribution held half of all items. Items were placed into the bins to achieve the sampling distribution in the right-hand histogram, assuring that items in the center were more common than items at the ends of the distribution. Most of the bins took up 10% of the variability on the bimodal dimension and 20% on a unimodal dimension. The exceptions were the bins in the corners, which were given 20% of the space on both dimensions to assure that they were symmetrical with respect to one another. Labeled items were only placed into two corner bins opposite one another, marked CH1and CH2.

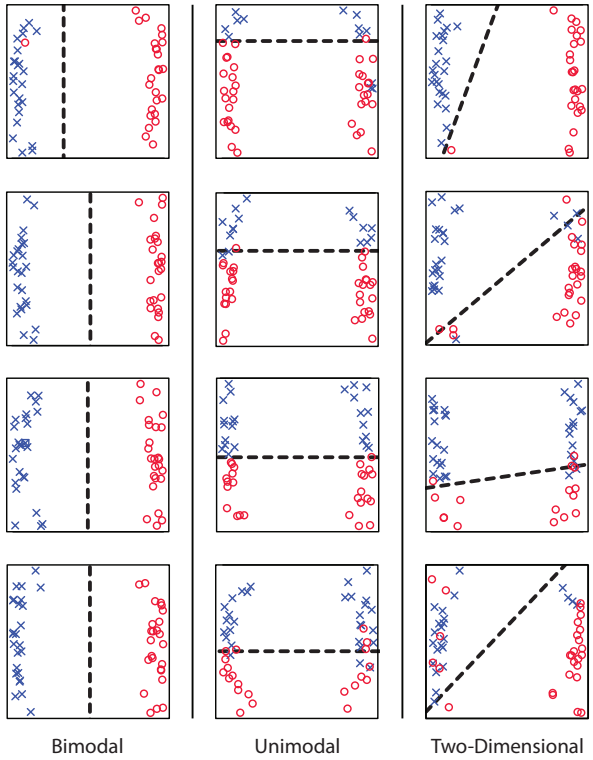

Figure 4: Performance of randomly selected subjects in the test phase. Four subjects best fit by each model were chosen at random. Each point represents an item that was presented to that participant during the test phase. The color and shape (red circles or blue crosses) indicate the response given by the participant to that item (i.e., CH1 or CH2). The dashed black line indicates the discriminant boundary that best fit the pattern of responding.

Labeled ($\bar{X}$ = .95, SD = .18, $t(29)$ = 13.43, $p < .001$), 40-Labeled ($\bar{X}$ = .92, SD = .16, $t(29)$ = 13.63, $p < .001$), and 40-All-Labeled ($\bar{X}$ = .98, SD = .049, $t(29)$ = 54.50, $p < .001$). Two participants reversed the rule (that is, consistently answered CH1 when CH2 was correct, and vice versa). Overall, subjects were familiar enough with the labeled training examples to generalize consistently to highly similar test items.

Strategy Analysis Following Z&M (2009), participants' classification strategies were characterized using the General Linear Classifier (or GLC, Ashby, 1992). Four potential models were fit to each subject. The first two assumed that participants formed a rule that could be characterized as a criterion on a single perceptual dimension: a *Bimodal* strategy was characterized by a criterion on the bimodal perceptual dimension, and a *Unimodal* strategy was characterized by a criterion on the unimodal dimension. The *Two-Dimensional* (2D) strategy was characterized by the use of a classifier integrating information from both stimulus dimensions (i.e., a diagonal boundary through the space). These three strategies formalize the strategies described in Figure 2. Finally, a *Flat* model was fit, which assumed that all responses were random and did not depend on the observed stimulus. The best-fitting model was chosen for each participant by comparing the BIC values for each of these models (which penalizes more flexible models such as our 2D model). Two research assistants visually rated the response patterns as well (see Figure 4 for examples), and showed high agreement with the model classification (90% and 91%)[2].

Four participants were best fit by the Flat model. Because there were few of them and their strategies were difficult to interpret, they were excluded from the analysis and replaced. A summary of the fits of the remaining participants can be found in Figure 5.

Replicating the findings of Z&M (2009), participants in the Unlabeled Condition showed a strong preference for classifying stimuli along the bimodal dimension (20/30, binomial test with $H_0 : k/n = 1/3$, $p < .001$). No analogous preference was found in the fully supervised (40-All-Labeled) condition (13/30, $H_0 : k/n = 1/3$, $p = .25$). A direct comparison of the fully unsupervised (Unlabeled) and fully supervised (40-All-Labeled) conditions showed a reliable difference (Fisher's exact test, $p < .05$). There was a slight preference for the Bimodal strategy over the Unimodal strategy in the 40-All-Labeled Condition (presumably from learning about the distribution at test). However, since this is a control condition

[2]Raters mostly disagreed in cases where subjects were best fit by the 2D model, but deviated from a 1D rule on only a few trials. Lacking a more principled method of choosing an objective criterion for classification, we deferred the outcome of model comparison via BIC.

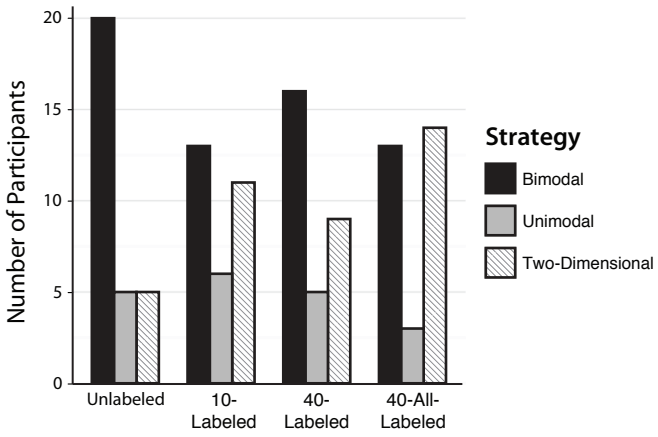

Figure 5: Histogram showing the trend in the number of subjects adopting each strategy across conditions. There is a general trend of an increase in the use of 2D rules in the presence of labeled items, as well as a drop in the use of a 1D rule on the bimodal dimension, but trends among the labeled conditions were weak.

the same bias can be evaluated in other conditions. Overall, the difference between the Unlabeled and 40-All-Labeled conditions was driven largely by the increase in the use of 2D rules in the 40-All-Labeled Condition. Note that while the distribution of labeled examples in the 40-All-Labeled Condition is logically consistent with all three strategies (Unimodal, Bimodal, 2D), a bias toward 2D rules is in line with the predictions of the optimal linear discriminant for the labeled examples.

Turning to the semi-supervised conditions, the distribution of strategies did not differ significantly between these two conditions (Fisher's exact test, $p = .79$). In addition, the proportion of participants using 2D rules did not vary between the conditions (Fisher's exact test, $p = .78$).

Combining the two semi-supervised conditions, we find an overall interaction between condition and the use of 2D rules (3 conditions × 2 strategies, Fisher's exact test, $p < .05$). However, the primary source of this effect seems to be the difference between the Unlabeled Condition and the other conditions. When we aggregate all the labeled conditions together, we see a greater use of 2D rules by the labeled conditions than the unlabeled condition (2 conditions × 2 strategies, Fisher's exact test, $p < .05$), while evidence for a parallel effect when aggregating the conditions had access to unlabeled training items together fell short of significance (2 conditions × 2 strategies, Fisher's exact test, $p = .07$). In summary, we found minimal evidence that the semi-supervised conditions were different from the all-labeled condition.

Discussion

Semi-supervised learning is a bit like the Higgs boson in particle physics. It is believed to occur (e.g., to allow word learning) and is strongly suggested by theories of human category learning (Anderson, 1991; Love et al., 2004), but has proven surprisingly difficult to observe. Our study represents yet another attempt to find laboratory support for this form of category learning. However, the patterns of generalization behavior exhibited at test during the two semi-supervised conditions most closely resembled the strategies of participants who learned in the fully supervised condition.

This result is striking for two reasons. First, unlike some of the previous work on semi-supervised learning, our experiment closely following existing protocols for studying unsupervised category learning in the literature in successfully replicating the results of Z&M (2009). In addition, given no other information participants in our study were willing to generalize according to the distribution of unlabeled examples. In the Unlabeled Condition, the most common strategy was to use a rule on the bimodal dimension. However, when labeled examples were included, participants responded similarly to the 40-All-Labeled Condition. This is the response pattern we would expect to see if subjects mostly failed to incorporate the unlabeled items into their representation of the category in the semi-supervised learning conditions. In this sense, our results join a growing chorus of studies which have failed to find semi-supervised learning except under very specific and limited circumstances (Gibson, Zhu, Rogers, Kalish, & Harrison, 2010; Rogers et al., 2010; Vandist et al., 2009).

In the following sections, we outline a number of possibilities about why semi-supervised learning has been so elusive in the lab.

Noticing "gaps" in the input? In our design, the distribution of labeled examples was systematically biased. One possibility is that learners eventually noticed the "gaps" in their input (i.e., that the labels only appeared with particular items) and thus inferred that these examples were somehow special or different. Such a hypothesis may be consistent with a rational learner who tries to determine which items should be clustered together (Griffiths, Sanborn, Canini, & Navarro, 2008). Under this view, an even smaller number of labeled examples (perhaps even one) may actually facilitate generalization (since the amount of data is enough to learn, but not enough to infer some systematic bias). We attempted to get at this issue by modulating the number of labeled training examples, and found no evidence of a trend. However, recent studies of one-shot learning suggest that often even a single labeled example can support robust generalization (Lake, Salakhutdinov, Gross, & Tenenbaum, 2011).

Overweighting of labeled examples Another interpretation, suggested by Zhu et al. (2010) and Lake and McClelland (2011), is that labeled items may simply be given more weight. Although this seems plausible, it would appear that the 10 labeled items in the 10-Labeled Condition outweighed the other 270 trials in the training phase, suggesting a weight for unlabeled items much lower than the previously reported estimate of around 40% (Lake & McClelland, 2011). Interestingly, subjects in our 10-Labeled Condition spent consider-

ably more time studying the labeled items, presumably raising their relative influence. One possibility is that the weight given to labeled items is actively adjusted by learners based on the task context.

Pedagogical sampling While assuming that labeled examples are given more weight might *describe* the lack of semi-supervised learning, it offers no specific proposal for why this should be the case. One possibility is that participants believed that the experimenter was providing information to teach them the category via the labeled examples (i.e., the labeled examples were pedagogically sampled). In this case, it may be reasonable to trust that the labeled items are particularly informative about the category distinction. For example, Shafto, Goodman, Gerstle, and Ladusaw (2010) have shown that adults adjust their inferences based on the intention of a speaker (either pedagogical or overheard). It is possible that participants in a lab-like setting often assume that training examples are presented pedagogically, causing them to downplay the relevance of unlabeled trials.

Is an explicit prediction required? A final hypothesis, there were minor differences between our task and previous work that may have influenced performance. For example, participants made observations and then simply pressed the space bar to acknowledge each item. By contrast, both Kalish et al. (2011) and Lake and McClelland (2011) asked participants to make a response on each trial. It is possible that making a response or prediction on each trial facilitates the integration of information across learning episodes. Consistent with this view is the fact that subjects 40-All-Labeled Condition showed evidence of learning from the items presented at test (where predictions were required)—recall that participants were slightly biased to respond according to the bimodal dimension, even though the only information about its bimodality was provided by the distribution of test examples. A similar effect may have carried over to the semi-supervised conditions as well.

Current work is exploring each of these possibilities. The hunt for semi-supervised learning continues.

Acknowledgments

We thank Seth Madlon-Kay and Dylan Simon for helpful comments and discussion in the development of this project. TMG was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

References

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

Ashby, F. G. (1992). Multidimensional Models of Categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Clapper, J. P., & Bower, G. H. (1994). Category Invention in Unsupervised Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 443–460.

Gentner, D. (1982). Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language development: lanugage, cognition, and culture* (pp. 301–334). Hillsdale, NJ: Erlbaum.

Gibson, B., Zhu, X., Rogers, T., Kalish, C., & Harrison, J. (2010). Humans learn using manifolds, reluctantly. In *Advances in neural information processing systems* (Vol. 24).

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as Nonparametric Bayesian Density Estimation.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, 1–13.

Lake, B., & McClelland, J. (2011). Estimating the strength of unlabeled information during semi-supervised learning. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Lake, B. M., Salakhutdinov, R., Gross, J., & Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society.* Cognitive Science Society. Austin, TX.

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

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*, 309–332.

Markant, D., & Gureckis, T. M. (2010). Category Learning Through Active Sampling. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Medin, D. L., & Schaffer, M. M. (1978). Context Theory of Classification Learning. *Psychological Review*, *85*, 207–238.

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

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*, 83–100.

Rogers, T. T., Kalish, C., Gibson, B. R., Harrison, J., & Zhu, X. (2010, May). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Shafto, P., Goodman, N. D., Gerstle, B., & Ladusaw, F. (2010). Prior expectations in pedagogical situations. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

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

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

Zhu, X. (2005). Semi-Supervised Learning Literature Survey. *Technical Report 1530*, 1–60.

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 *The 27th international conference on machine learning (ICML)* (p. 158).