Estimating the strength of unlabeled information during semi-supervised learning

Brenden M. Lake (brenden@mit.edu)
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology

James L. McClelland (mcclelland@stanford.edu)
Department of Psychology, Stanford University

Abstract

Semi-supervised category learning is when participants make classification judgements while receiving feedback about the right answers on some trials (labeled stimuli) but not others (unlabeled stimuli). Sporadic feedback is common outside the laboratory, and it is important to understand how people learn in this setting. While there are numerous recent studies, the strength and robustness of semi-supervised learning effects remain unclear, particularly when labeled and unlabeled stimuli are dispersed across learning. We designed an experiment, using simple unidimensional category learning, that allows us to measure the relative contribution of labeled and unlabeled experience. Based on an analysis of this task, we find that an unlabeled stimulus is worth more than 40% of a labeled stimulus.

Keywords: categorization; semi-supervised learning

People organize perceptual knowledge into categories such as types of cheese, types of cars, or types of animals. When acquiring these categories from experience, people must learn to associate stimuli with the appropriate category labels. Consider this anecdote about acquiring cheese categories.

> A man attending a food festival tries several samples of cheese. After tasting the samples, he looks at the accompanying labels on each cheese, remarking that the Gruyère is his favorite. Several days later at a party, he enjoys several pieces of cheese from a platter with no accompanying labels. Although uncertain, he thinks he recognizes the taste of Gruyère he recently experienced at the food festival. He suspects a certain cheese is Gruyère but he is unsure.

Does the man use this unlabeled encounter at the party to further refine his understanding of cheese categories? If so, he is engaged in *semi-supervised category learning*, defined as learning from both labeled (feedback) and unlabeled (no-feedback) encounters with objects. This is in contrast to supervised learning, which uses only labeled examples, and unsupervised learning, which uses only unlabeled examples, to learn categories.

Semi-supervised learning algorithms are studied in machine learning, largely because labeled data is often more difficult and expensive to obtain than unlabeled data (Zhu, 2005; Chapelle, Scholköpf, & Zien, 2006). For example, if one is classifying web pages into categories based on content, labeled data would likely be collected by hand while unlabeled data could be harvested from the internet automatically (see Nigam, McCallum, and Mitchell (2006) for related application). Humans face a similar problem when learning perceptual categories, where the amount of unlabeled encounters with objects often exceeds the amount of labeled encounters. Despite the sparsity of labels, the human ability to learn new concepts is remarkable, and perhaps this ability can be partially explained by effective semi-supervised learning.

Recent work has investigated whether people perform semi-supervised learning, but many studies are limited by having separate supervised (feedback) and unsupervised (no-feedback) phases of learning. In two studies, participants were first shown a small number of labeled examples (typically one or two), followed by a large set of unlabeled examples (Zhu, Rogers, Qian, & Kalish, 2007; Zhu et al., 2010). These studies showed that people are sensitive to both the distributional structure of the subsequent unlabeled experience (Zhu et al., 2007) and the stimulus order (Zhu et al., 2010). Other studies showed labeled and unlabeled data together in a questionnaire format. Stromsten (2002) presented participants with a labeled example of a fish simultaneously with either 0, 8, or 29 unlabeled fish examples, finding a overall difference in classification performance. Gibson, Zhu, Rogers, Kalish, and Harrison (2010) presented participants with a few labeled examples that reside within dense clusters of unlabeled stimuli. People propagated the labeled information along the cluster, but only if neighboring stimuli were highlighted during categorization.

By using separate supervised and unsupervised phases, the studies mentioned do not address semi-supervised learning when feedback is dispersed across learning. People may behave differently when feedback is available on some trials and not others, with the two types intermixed. If additional labeled information is anticipated, people might ignore or fail to learn from the more ambiguous unlabeled information. A recent study by Vandist, De Schryver, and Rosseel (2009) is consistent with this position, finding no evidence for semi-supervised learning when feedback was intermittent. Participants engaged in an information-integration task, defined as a task where participants have to combine perceptual information from two underlying stimulus dimensions simultaneously to obtain optimal performance (Ashby, Queller, & Berretty, 1999). Unlabeled trials were drawn from the exact same distribution as the labeled trials, and one group received unlabeled stimuli while another group received unrelated filler events instead. Learning progressed at similar rates, providing no evidence of semi-supervised learning. In a similar design utilized in Rogers, Kalish, Gibson, Harrison, and Zhu (2010), there was no evidence for semi-supervised learning, which the authors suggest arose from selective at-

tention to an irrelevant stimulus dimension. When they imposed a response deadline to disrupt this attention, there was an effect of the unlabeled stimuli.

We designed an experiment that allows us to measure the relative contribution of labeled versus unlabeled experience during learning with intermittent feedback. In our task, the unlabeled stimuli provide additional information, beyond simply more samples from the stimulus distribution as in Vandist et al. (2009) and Rogers et al. (2010), which allows us to estimate the strength of the unlabeled impact. Zhu et al. (2007) estimated this quantity and found that an unlabeled stimulus was worth a surprisingly small fraction (about 5%) of a labeled stimulus. In a statistical analysis of our data, we find a much larger contribution of 40% to 100% based on the 95% posterior interval and discuss the implications for theories of categorization.

Experiment

Participants were assigned to one of two groups, where both groups received exactly the same labeled items but different unlabeled items. The labeled items suggest a particular category boundary, and the unlabeled distributions were designed to shift this boundary in opposing directions (Zhu et al., 2007). Consequently, a difference in classification boundary between the two groups would demonstrate semi-supervised learning.

Method: Semi-Supervised Learning

Participants 40 subjects from Stanford University and the surrounding community participated in this experiment for a payment of $6 to $8 depending on performance. Participants were told that overall accuracy in the task would determine payment.

Design Participants made a sequence of categorization judgements. After each judgement, feedback was or was not presented. Participants were randomly assigned to either a "left shift" or "right shift" group. These groups received the same distribution of feedback items but different distributions of no-feedback items.

Stimuli Stimuli were horizontal lines that varied only in length. Each line belonged to one of two categories, C (shorter lines) or N (longer lines). The distribution of stimulus items is shown in Figure 1. Feedback occurred for about 25% of the trials. Each training item was presented twice per block (either twice unlabeled or once labeled and once unlabeled) as illustrated in Figure 1. Both the left shift and right shift groups saw the exact same feedback items once per block (black bars). However, the groups differed in no-feedback items (grey bars), which were shifted to be smaller in length (left shift) or larger in length (right shift). If participants perform semi-supervised learning and integrate the no-feedback trials into their category representations, their classification boundaries should likewise be shifted either to the left or right.

Training consisted of 5 blocks of 48 trials each. In addition, 9 transfer items were added to the last two blocks, presented once per block without feedback. Hence there was no separate testing period during the experiment. These transfer items were designed to probe the region near the category boundary to see if the two participant groups differed in their category representations.

Procedure Participants were informed that they would see lines, varying only in length, that belonged to one of two categories corresponding to the keys "C" vs. "N" on the keyboard. They were instructed that sometimes they would be told whether their answer was correct or incorrect, and other times they would be told nothing, regardless of whether their answer was correct or incorrect. Participants were self-paced during the task. Upon entering a response, feedback trials displayed a "Correct!" or "Wrong!" message. The feedback and the stimulus remained on the screen for 2 seconds. For no-feedback trials, the stimulus stayed on the screen for 2 seconds without any text. One participant was removed from the analysis for near chance performance.

Method: Supervised Learning

For comparison, we ran a fully-supervised experiment that was exactly analogous to the semi-supervised experiment. The method was identical except where noted.

Participants and Design 21 subjects participated in this experiment. Participants were assigned to either a "left shift" or "right shift" group.

Stimuli The stimuli were identical to the semi-supervised version, as illustrated in Figure 1, except that all training items received feedback (there is no distinction between the black and grey bars in the figure). The purpose of the experiment was to see how much the boundary differed between the left and right shift groups when there was full feedback. As in the semi-supervised experiment, nine transfer items were intermixed in the last two blocks, presented once per block without feedback.

Procedure Instructions were the same, except participants were told that "Most of the time you will be told by the computer whether your answer was correct or incorrect" instead of "Sometimes." This was changed since only the transfer items were without feedback in this experiment. One participant was removed for using an aid to help measure the lines.

Results

Performance reached high accuracy. For the third training block, the semi-supervised group's mean accuracy was 97% correct (Figure 2a) and the supervised group's was 99% correct (Figure 2b). Participants were clearly categorizing the no-feedback items into the correct categories in most cases, before the transfer items were introduced in the next block.

For participants assigned to the semi-supervised learning condition, the no-feedback items produced a clear influence

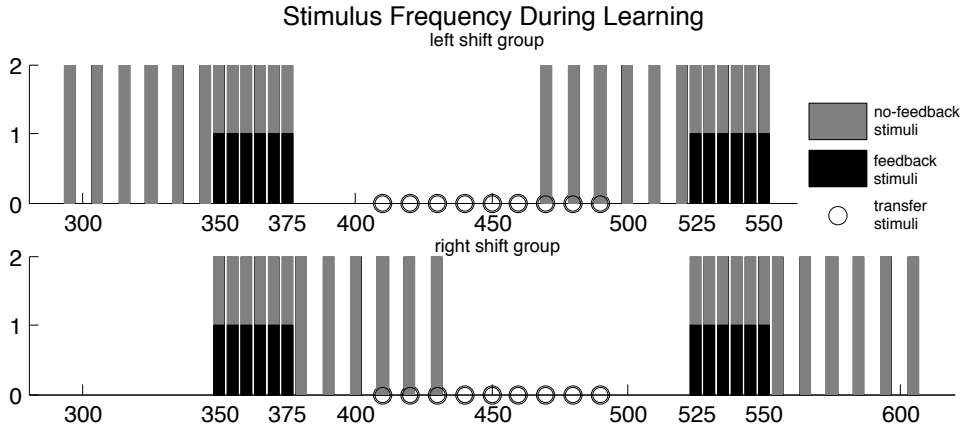

Figure 1: Frequency distribution for a block of learning. Participants were assigned to either the left shift group or the right shift group. All training items were presented twice per block. Each item was either unlabeled twice (all grey bars) or once unlabeled and once labeled (half grey and half black bars). Participants in both shift groups received the same feedback items, but no-feedback items were either left shifted or right shifted. Transfer stimuli that were presented once per block in the last two blocks are indicated by open circles.

on the categorization of novel transfer items. Figure 2c shows the responses to transfer stimuli during the last two blocks of the experiment, for the left shift and right shift groups. The groups differed in how they categorized the transfer stimuli (410 to 490 pixels) in the expected direction. The difference between groups in overall number of "N" responses for transfer items was significant (one-tailed t(37) = 5.7, $p < .001$, left shift M = 14.6 of possible 18, SD = 2.0 and right shift M = 9.3 of 18, SD = 3.6).[1]

Participants assigned to the fully supervised task showed an even larger shift, as expected (Figure 2d). The difference between shift groups in overall number of "N" responses for transfer items was significant (one-tailed t(18) = 7.8, $p < .001$, left shift M = 15.0 of possible 18, SD = 2.5 and right shift M = 5.9 of 18, SD = 2.7).

Bayesian estimate of unlabeled influence

To estimate the influence of the unlabeled data, we conduct a Bayesian analysis of the psychometric function for the semi-supervised experiment. This is a statistical analysis of the data, not a model of category learning. The variable of primary interest is λ that modulates the contribution of the unlabeled information, where $\lambda = 0$ denotes no influence and $\lambda = 1$ denotes an equal weight of unlabeled and labeled influence. There are two other unknown variables, response bias β_N and perceptual noise ϕ which flattens the decision curve. We assume uniform priors $\lambda \sim$ Unif(0,1), $\beta_N \sim$ Unif(0,1), and $\phi \sim$ Unif(0,100) measured in pixels.

[1]For any given participant, three transfer items were identical in length with three no-feedback training items. Thus, during the last two blocks when transfer items were shown, these items were presented three times each per block without feedback while the other transfer items are presented only once. For the analysis, these items were included only once, with one replication per block randomly designated as the transfer trial.

We use the notation s_i to denote a transfer item with a categorization response $y_i \in \{C, N\}$ which denote the two response keys "C" and "N." We model the response probability of a transfer item as

$$P(y_i = N | s_i, \lambda, \beta_N, \phi)$$

$$= \frac{\beta_N \text{Normal}(s_i | \mu_N, \sigma_N^2)}{\beta_N \text{Normal}(s_i | \mu_N, \sigma_N^2) + (1 - \beta_N)\text{Normal}(s_i | \mu_C, \sigma_C^2)},$$

using Gaussians to model the category densities.

The parameters of these categories (μ_N, μ_C, σ_N^2, and σ_C^2) are defined as functions of the unlabeled contribution variable λ. Essentially, these parameters are calculated as the mean and variance of the stimuli assigned to each category during training, where the labeled and unlabeled training stimuli have different contributions as determined by λ. More formally, let r_i denote a training item and let $z_i \in \{C, N\}$ denote its label as defined by the experimenter (which may or may not be observed by participants). The set $i \in L$ denote labeled items r_i with observed labels z_i, and the set $j \in U$ denote unlabeled items r_j with hidden labels z_j. The category parameters are then defined as

$$\mu_N = \frac{\sum_{i \in L} \delta(z_i, N) r_i + \lambda \sum_{j \in U} \delta(z_j, N) r_j}{\sum_{i \in L} \delta(z_i, N) + \lambda \sum_{j \in U} \delta(z_j, N)}$$

where the delta function $\delta(z_j, N) = 1$ when $z_j = N$ and $\delta(z_j, N) = 0$ otherwise, and the category variance is

$$\sigma_N^2 = \phi^2 + \frac{\sum_{i \in L} \delta(z_i, N)(r_i - \mu_N)^2}{\sum_{i \in L} \delta(z_i, N) + \lambda \sum_{j \in U} \delta(z_j, N)} + \frac{\lambda \sum_{j \in U} \delta(z_j, N)(r_j - \mu_N)^2}{\sum_{i \in L} \delta(z_i, N) + \lambda \sum_{j \in U} \delta(z_j, N)}.$$

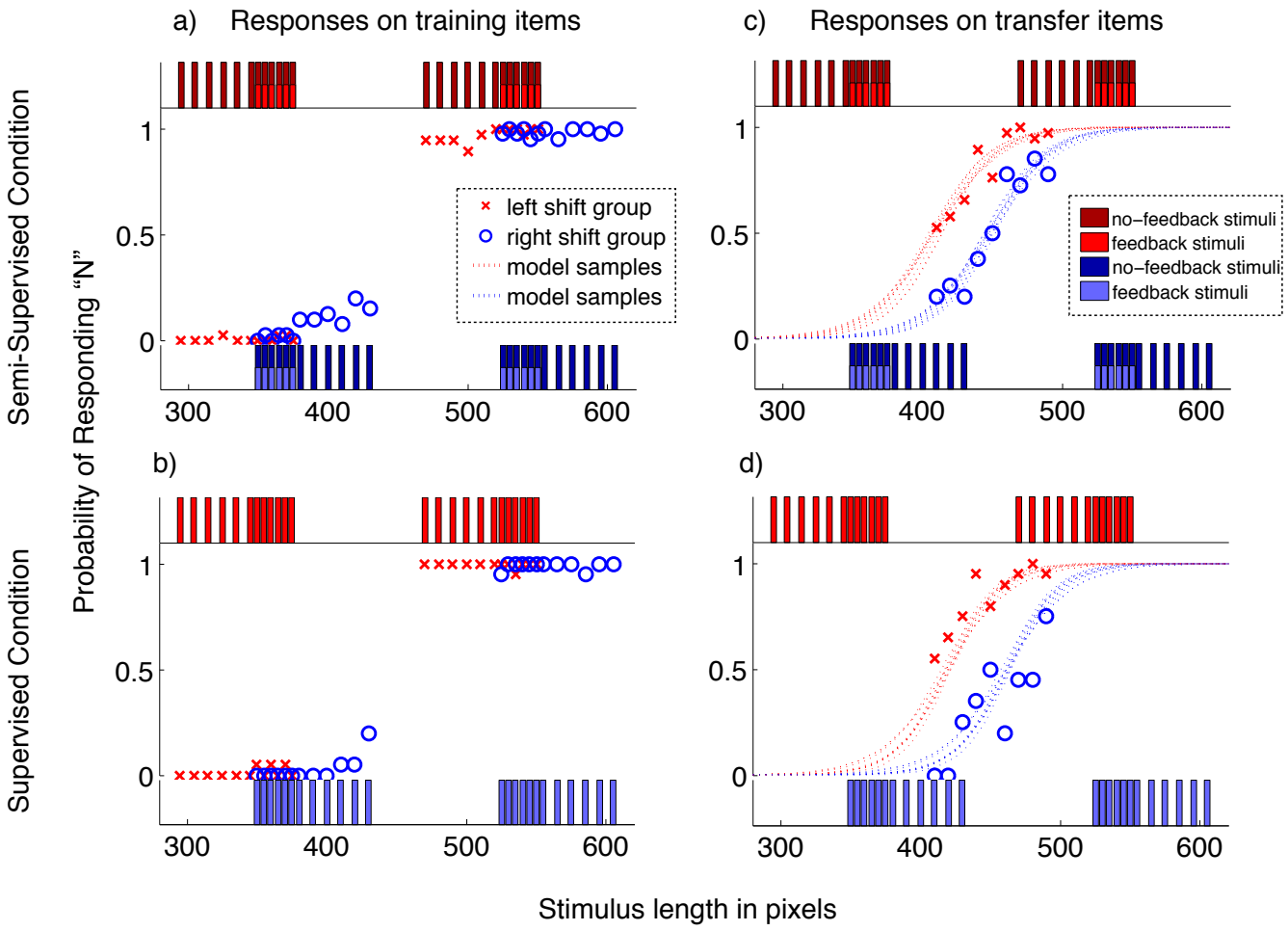

Figure 2: Categorization of stimuli aggregated across participants. Part a) and b) show the responses during the third block of learning (the block before the transfer items were introduced). Part c) and d) show the categorization of transfer items (lengths 410 through 490) during the last two blocks of training, which were intermixed with the training items (colored histograms). Part a) and c) show semi-supervised participants, and b) and d) show supervised participants. The plots shows the probability of responding "N" for each stimulus during these blocks, aggregated across participants. Participants in the left shift and right shift groups showed different categorization profiles. For the semi-supervised condition, these two groups only differed in the no-feedback stimuli they saw. This indicates that no-feedback stimuli influenced the learned category representations. The dotted curves shows 10 posterior samples of psychometric functions, see "Bayesian estimate of unlabeled influence."

The parameters μ_C and σ_C^2 are defined similarly. This analysis does assume that training stimuli r_i are correctly categorized by the participants, which leads to a conservative estimate of λ.[2]

We aggregate the transfer stimuli across participants to form vectors y of category responses ($y_i \in \{C, N\}$) with the corresponding stimulus vector s. The posterior of the parameters given the data is thus

$$p(\lambda, \beta_N, \phi|y, s) \propto p(\lambda)p(\beta_N)p(\phi)\prod_i P(y_i|s_i, \lambda, \beta_N, \phi).$$

From this joint posterior, we compute the marginal posterior distributions $p(\lambda|y, s)$, $p(\beta_N|y, s)$, and $p(\phi|y, s)$ which are shown in Figure 3. This computation is done by approximating the continuous posterior with a discrete grid and summing over the other variables. The distribution of λ is of primary interest, since it signifies the contribution of the unlabeled data. By simulation, the posterior mean for λ is 0.725 and the 95% central interval is $[0.413, 1]$ (Gelman, Carlin, Stern, & Rubin, 2004). Given the modeling assumptions, there is a 97.5% chance that an unlabeled example is worth more than 40% of a labeled example.

For an analogous analysis in the supervised case without λ, we find a 95% central posterior interval for β_N is $[0.55,\ 0.68]$ and for ϕ is $[47.5,\ 61]$

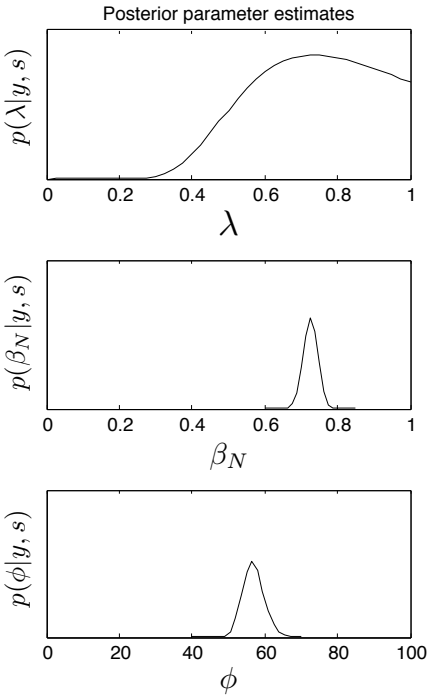

Figure 3: Posterior parameter estimates for semi-supervised learning. Parameter λ controls the ratio of unlabeled to labeled influence, β_N is the response bias, and ϕ is the perceptual noise in pixels.

Discussion

We tested participants in a paradigm where two groups received exactly the same labeled items but different unlabeled distributions (see Zhu et al., 2007). Labeled and unlabeled trials were randomly intermixed and presented in sequence. By the end of learning, there was a difference in classification boundary between the two groups, indicating that the participants performed semi-supervised learning. The experiment allowed us to measure the relative contribution of labeled and unlabeled experience. Through a Bayesian analysis of the psychometric function, we find that parameter λ, the ratio between the unlabeled and labeled contribution, has a posterior mean of $\lambda \approx 0.725$. The central posterior interval suggests there is a 97.5% chance that an unlabeled example is worth more than 40% of a labeled example.

This is in contrast to a past study by Zhu et al. (2007), where a similar parameter was found to be $\lambda \approx 0.06$, meaning about 5% of a labeled example. Also Zhu et al. (2010) fit a similar parameter in an exemplar model and found $\lambda \approx 0.2$. There are several important factors that could underly this difference between past studies and our own. First, the Zhu et al. studies showed participants the labeled stimuli first and then the unlabeled stimuli afterwards. Second, the ratio of labeled to unlabeled stimuli was much higher in our study. To better understand the strength of semi-supervised learning, future work should manipulate these factors systematically and further explore the situations most representative of natural learning settings.

Studying semi-supervised learning in scenarios that do not support strictly unsupervised learning is another interesting avenue for future work. For instance, the information integration task (Ashby et al., 1999) results in successful learning when supervised but unsuccessful learning when unsupervised. Could semi-supervised learning lead to successful learning in this task, above and beyond learning supported by just feedback examples? Vandist et al. (2009), who tested semi-supervised learning in the information integration task, suggest the answer is no. But in their design, the unlabeled stimuli provided no new information beyond simply more samples from the stimulus distribution. Future work should explore the information integration task when the unlabeled distribution provides additional information, as in our experiment, that results in large decision curve shifts. Given the clear effect in our study, it would be interesting to see whether unlabeled stimuli influences learning in the information integration task.

Future work should also investigate computational models that support semi-supervised category learning. There are a variety of candidate models that do not make strong dis-

[2]The assumption that the distribution parameters are derived from the intended labeling of the training stimuli r_i is supported by the high overall accuracy in the third training block (97% correct). In a hypothetical case that a participant has not assigned a particular unlabeled stimulus to the correct category but the model has, parameter fitting to response data would underestimate rather than overestimate the influence parameter λ, since this misassigned stimulus is contributing to the decision curve shift in the model, but not the subject.

tinctions between supervised and unsupervised learning (e.g., Anderson, 1991; Love, Medin, & Gureckis, 2004; Vallabha, McClelland, Pons, Werker, & Amano, 2007; Lake, Vallabha, & McClelland, 2009; Zhu et al., 2010). Theories of categorization must account for semi-supervised learning, given that this type of learning has substantial impact.

Acknowledgments

We thank Frank Jäkel and Juan Gao for numerous discussions, and Peter Battaglia and Stanford's Parallel Distributed Processing Lab for their helpful comments. This research was supported in part by Stanford University's Symbolic Systems Summer Internship Program, AFOSR, and the Air Force Research Laboratory, under agreement number FA9550-07-1-0537. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

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

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

Chapelle, O., Scholköpf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. MIT Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2 ed.). Chapman and Hall/CRC.

Gibson, B., Zhu, X., Rogers, T., Kalish, C., & Harrison, J. (2010). Humans learn using manifolds, reluctantly. In *Advances in Neural Information Processing Systems (NIPS)* (Vol. 24).

Lake, B. M., Vallabha, G. K., & McClelland, J. L. (2009). Modeling unsupervised perceptual category learning. *IEEE Transactions on Autonomous Mental Development*, *1*(1), 35-43.

Love, B. C., Medin, D. L., & Gureckis, T. (2004). Sustain: A network model of category learning. *Psychological Review*, *111*(2), 309-332.

Nigam, K., McCallum, A., & Mitchell, T. (2006). Semi-supervised text classification using EM. In O. Chapelle, B. Scholköpf, & A. Zien (Eds.), *Semi-supervised learning.* MIT Press.

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

Stromsten, S. (2002). *Classification learning from both classified and unclassified examples*. Unpublished doctoral dissertation, Stanford University.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Science*, *104*(33), 13273-13278.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: The impact of feedback in learning the information-integration task. *Attention, Perception, and Psychophysics*, *71*(2), 328-341.

Zhu, X. (2005). *Semi-supervised learning literature survey* (Tech. Rep. No. 1530). Computer Sciences, University of Wisconsin-Madison.

Zhu, X., Gibson, B., Jun, K.-S., Rogers, 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).*

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. In *Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07).*