Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task

Timothy T. Rogers, Charles Kalish, Bryan R. Gibson, Joseph Harrison and Xiaojin Zhu
Departments of Psychology and Computer Science
University of Wisconsin-Madison
Madison, WI 53706 USA

Abstract

Recent empirical studies of semi-supervised category learning—where learners only occasionally receive information about a given item's category membership—have yielded contradictory results, with some studies showing strong effects of unlabeled experience and others little or no effect. We report two experiments designed to help understand this heterogeneity. In both, participants performed a two-category classification task with novel stimuli varying along two psychologically separable dimensions. In semi-supervised conditions, participants categorized and received feedback on 32 "labeled" items intermixed with a large number of "unlabeled" items. In the supervised-only condition, participants viewed the same labeled trials intermixed with a large number of filler trials. Without time pressure participants learned the task equally well in both conditions. When required to respond very rapidly, however, participants performed substantially better in the semi-supervised condition. The discrepant results may indicate a role for selective attention in human semi-supervised learning.

Keywords: semantic interference, visual lexical decision, dual-task, single-system view

Introduction

Most theoretical and computational approaches to human category learning consider fully supervised learning: for every training experience, the learner has access to a representation of the stimulus and to the true category label (e.g. Nosofsky, 1986; Kruschke, 1992; Gluck and Bower, 1988; Anderson, 1991 and many others). Fully unsupervised approaches—where the learner never has access to the true category label but must learn to group items into categories on the basis of their similarity—are less common but have also appeared in the literature (e.g. Fried and Holyoak, 1984). Neither approach seems fully adequate, however, for explaining human categorization. Although a great deal of natural experience is unsupervised—we continually encounter objects in the world without a "teacher" telling us what kind of things they are—we also certainly get a nontrivial amount of "labeled" experience, where a recognized authority provides the true class label either directly in an explicit teaching scenario or indirectly through use of the label in communication. Human category learning may, therefore, involve combining both labeled and unlabeled sources of information—that is, human category learning may be semi-supervised.

The question of how best to combine labeled and unlabeled data has been a topic of considerable investigation in machine learning, where it has been formally shown that, for some kinds of learning problems, a learner can converge much more quickly on an accurate representation of the category structure by combining labeled and unlabeled observations (Chapelle, Zien, and Scholkopf, 2006; Zhu and Goldberg, 2009). In cognitive psychology, the empirical question of how experience with both labeled and unlabeled items might influence category learning has rarely been studied. Some well-known computational approaches to category learning suggest ways in which labeled and unlabeled observations might combine to influence knowledge of category structure (e.g. Nosofsky, 1986; Schyns, 1991; Love et al. 2004), but these ideas have not been linked to the formal analyses offered by machine learning and have not been a focus of much empirical work.

We are aware of only two studies designed to assess whether category learning is influenced by unlabeled experiences, and these come to opposing conclusions. On the positive side, Zhu and colleagues (2007) studied performance in a 1-dimensional 2-category learning task. After learning a category boundary with a small amount of supervised training (ie training with corrective feedback), participants subsequently classified a large number of items with no feedback. These "unlabeled" items were sampled from a bimodal distribution with a trough that was displaced to one side or the other of the original learned category boundary. The authors found that, following the unlabeled experience, participants shifted their mental category boundary toward the trough of the unlabeled distribution. This finding suggests that people expect category boundaries to align with low-density regions in the unlabeled feature space, and use unlabeled observations to adjust their representations of category structure accordingly.

In contrast, Vandist and colleagues (2009) studied a binary classification task with stimuli that varied in two psychologically separable dimensions (the orientation and spatial frequency of Gabor patches). Participants viewed a number of labeled examples intermixed either with additional unlabeled examples or with unrelated filler items. Unlabeled items were sampled from a bimodal distribution in which the trough aligned with the true category boundary. The authors found no difference in the rate of learning or overall performance between these conditions—suggesting that the unlabeled items provided no overall benefit in learning the category structure, even though the distribution of these items was consistent with the to-be-learned boundary.

In this paper we investigate some of the factors that might explain the different results obtained by these studies. Though both groups focused on semi-supervised learning, there were several key differences in their experiments: (i) Where Vandist et al. used stimuli varying in two psychologically separable dimensions, Zhu et al. employed visually complex shapes varying along a line in a multidimensional feature space. (ii) Where Vandist et al. provided participants with many labeled items, Zhu et al. trained participants with 10 repetitions each of just 2 individual tokens (ie one exemplar of each category). (iii) Vandist et al. employed a task requiring participants to integrate two separable dimensions (ie the category boundary was oblique in the 2D feature space) whereas Zhu et al. employed a simple 1D category learning task. (iv) Vandist et al. provided participants with ongoing labeled training experiences, whereas Zhu et al. performed a short block of supervised learning followed by a long block of unsupervised trials. (v) Vandist et al. compared performance in a semi-supervised condition to performance in a fully-supervised condition, whereas Zhu et al. compared two different semi-supervised conditions.

Thus there are several potential hypotheses as to why different results were obtained in the two studies. We report two experiments designed to narrow the range of possible hypotheses by capitalizing on the positive characteristics of both Zhu et al.'s (2007) and Vandist et al.'s (2009) original designs. Like Vandist and colleagues, our experiments (i) employ stimuli that vary along two obvious and psychologically separable dimensions, (ii) compare a semi-supervised condition to a matched supervised condition, and (iii) provide participants with ongoing exposure to labeled data. Like the experiment described by Zhu et al., (i) our stimuli were more object-like, (ii) participants in the semi-supervised condition received relatively few labeled trials (8%), and (iii) the boundary to be learned did not require integration of the two dimensions. In Experiment 1 we show that, under these conditions, people seem relatively insensitive to unlabeled learning experiences. Experiment 2 then tests a more explicit hypothesis about the conditions under which unlabeled experiences influence performance.

Experiment 1

Method

Participants. 50 undergraduate students from UW-Madison participated in Experiment 1 for course credit or monetary compensation. All had normal or corrected-to-normal vision.

Materials and Design. The stimuli were derived from classic work by Nosofsky (1986). They consisted of circles bisected by an oblique line, and varied in radius (ie circle size) and in the precise angle of the bisecting line. Like the dimensions employed by Vandist et al. (2009), size and line orientation are two psychologically separable dimensions—that is, it is possible to attend selectively to one dimension without processing the other. In our stimuli, circle radius varied from 50 to 120 pixels while line orientation varied from 0 to 90 degrees (measured from the horizontal).

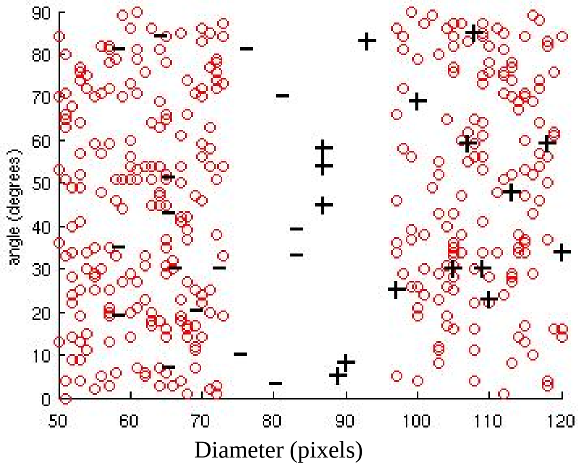

Figure 1. Example of the distribution of labeled (black) and unlabeled (red) items for one participant. Plus signs show labeled items from Category A, minus signs show labeled items from Category B.

Pilot testing with a fully-unsupervised procedure showed a general bias for classifying these stimuli according to the angle dimension—only 35% of participants made unsupervised categorization decisions based on size. Consequently our experiments involved learning to classify these items according to their size. Items larger than or equal to 85 pixels in radius were designated class A while those smaller than 85 pixels were designated class B.

The experiment included two between-subjects conditions. In the *semi-supervised* (SS) condition, participants viewed a total of 32 labeled items—items for which feedback was provided—sampled from a uniform distribution over the space. These were intermixed with 400 unlabeled examples sampled from a bimodal distribution that was uniform along the angle dimension but had a substantial gap along the size dimension (see Figure 1). Thus the gap in the unlabeled distribution provided a potential cue to orientation and location of the true category boundary. In the *supervised-only* (SO) condition, participants viewed the same 32 labeled items as in the semi-supervised condition. In this case, however, these items were intermixed with filler trials in which participants viewed the word "left" or "right" on the screen and pressed the corresponding mouse button. Labeled trials were ordered so that 8 appeared in each block of 100 unlabeled/filler trials. Subjects in the SS and SO conditions were yoked so that each SO participant viewed exactly the same labeled items in exactly the same sequence and at exactly the same time as a participant in the SS condition. Thus the only difference between conditions was whether the trials interspersed among labeled examples consisted of unlabeled examples or of filler. After experience with the labeled and unlabeled/filler trials, both groups categorized, without feedback, 36 items forming an evenly-spaced "grid" in the stimulus space. Performance was assessed as the mean proportion correct in each successive block of 8 labeled items and on the unlabeled grid items.

If participants use the gap in the unlabeled distribution to form their mental category boundary, their accuracy on the labeled items should increase more rapidly, and their performance on the final grid should be better overall, than participants in the control condition.

Procedure The experiment was carried out on PCs running the DMDX software package under Windows XP. The 50 participants were randomly assigned to either the SS or SO condition with 25 participants in each. Participants in both groups were told that they would view a series of objects and that each belonged to one of two categories. Their job was to learn to classify the objects correctly by pressing one of two buttons on the mouse. Participants in both conditions were told that they would only occasionally get feedback indicating whether their choice was correct, but that they should do their best to categorize all of the items regardless. Participants in the SO condition were additionally told that categorization trials would be interspersed with button-pressing trials in which they would view the word "left" or "right" and must press the corresponding mouse button. The principal dependent measure was the mean proportion correct for each successive block of 8 labeled items and for the 36 unlabeled grid items.

Results

Figure 2 (top) shows means and standard errors of the accuracy for each block of 8 labeled items and for the final unlabeled grid in the two conditions. A repeated measures ANOVA treating time (each block of 8 labeled items plus final grid) as a within-subjects factor and learning condition (SS / SO) as a between-subjects factor revealed a significant main effect of time with performance improving overall (F(5,192) = 5.36, p < 0.001), but no effect of learning condition (F(1,48) = 0.29, p = 0.59) and no interaction between these (F(4,192) = 0.51, p = 0.73).

Performance overall was highly variable, with some participants learning fairly well and others not at all. In fact performance on the final grid was bimodal in both groups, with one subgroup choosing correctly on 67% or more of the grid trials and the other group at chance. We therefore classified each participant as a "learner" or a "nonlearner" based on grid performance, with learners showing accuracy greater than 66%. The number of learners in each condition was comparable (13/25 in the semi-supervised group, 12 /25 in the control group), suggesting that the unlabeled items did not produce a greater likelihood of learning the correct boundary.

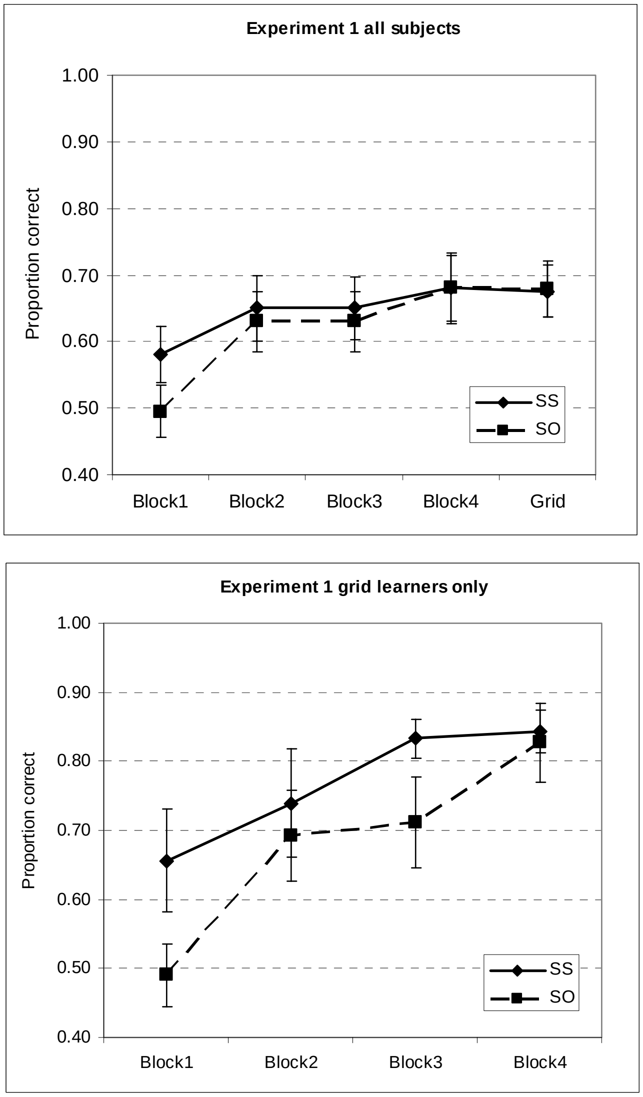

Figure 2. Top: Mean proportion correct for labeled items and grid for all participants in Experiment 1. Bottom: Mean proportion correct for labeled items in each block across participants who performed above criterion on the final grid. Error bars indicate the standard error of the mean.

Finally we investigated the effect of time and learning condition on accuracy for the 4 blocks of labeled items considering just those participants who performed to criterion on the grid items. These data are shown in Figure 2 (bottom). Though learners in the SS condition appeared to perform marginally better, this effect was not statistically reliable. A repeated-measures ANOVA showed a reliable main effect of time (F(3,69) = 10.9, p < 0.001) but no effect of condition (F(1,23) = 2.2, p = 0.15) and no interaction (F(3, 69) = 1.0, p = 0.40).

In sum, we obtained no evidence for semi-supervised learning in this experiment: though unlabeled items were selected from a distribution with a prominent gap that aligned well with the true category boundary, experience with this distribution did not significantly impact the overall rate of learning, the mean accuracy, or the number of participants who learned successfully.

Experiment 2

Consistent with the observations of Vandist and colleagues (2009), Experiment 1 showed little effect of unlabeled experience on category learning. What then accounts for the strong effects of unlabeled experience previously observed by Zhu et al. (2007)? Experiment 2 tested one hypothesis: perhaps the difference is observed because, in both the current work and in Vandist et al.'s (2009) experiment, the stimuli were composed of two psychologically separable dimensions. A classic tradition of research in *concept attainment* has shown that, for such stimuli, people often adopt a "win-stay-lose-shift" strategy (Bruner, Goodnow and Austin, 1956). That is, they formulate a hypothesis about the relevant dimension for categorization, then make their decision based solely on that dimension until they receive evidence that their hypothesis is wrong, at which point they shift to a new hypothesis. If feedback is very sparse, participants may focus on the dimension they believe to be relevant to the exclusion of other dimensions. That is, participants may not attend to the competing dimension at all on many trials, and so may be exposed to very little information about the distribution on this dimension. Especially for our stimuli, where pilot studies suggest that participants are biased to attend to the irrelevant dimension (angle), such strategic/attentional effects might seriously attenuate any influence of unlabeled experience.

To test this hypothesis, we conducted a second study identical to Experiment 1 in all but one respect: in Experiment 2, participants were required to respond within a deadline of 600ms. With this requirement of a very rapid response, participants have little time to focus their attention on one dimension or the other. Consequently, we predicted that the distribution of unlabeled examples would have a more significant impact on category learning in this paradigm.

Method

Participants 50 undergraduate students who did not participate in Experiment 1 were recruited for this study in return for course credit. All participants had normal or corrected-to-normal vision.

Materials and Designs The materials and design were identical to Experiment 1, except that participants in both groups were told that they would need to respond to each item as rapidly as possible.

Procedure Participants were randomly assigned to one of the 2 conditions, with 25 participants in each group. The procedure was identical to Experiment 1 with the following exceptions. First, each stimulus appeared onscreen for 125ms and was then replaced by a visual mask composed of hash marks. Participants were given 600ms from the onset of the mask to make their response. If the participant did not respond within this window, the computer indicated that the response was too slow. On labeled trials that did not meet deadline, the computer indicated that the response was too slow and also presented the correct category label. In both conditions, the deadline was imposed on both labeled trials and on unlabeled/filler trials.

Results

Trials that did not meet deadline were discarded from the analysis; these included just 5% of trials on average. Thus most participants were able to respond within the time-window on the majority of trials. For the remaining trials, we computed the mean accuracy on each successive block of 8 labeled trials and on the final unlabeled grid. Results are shown in Figure 3.

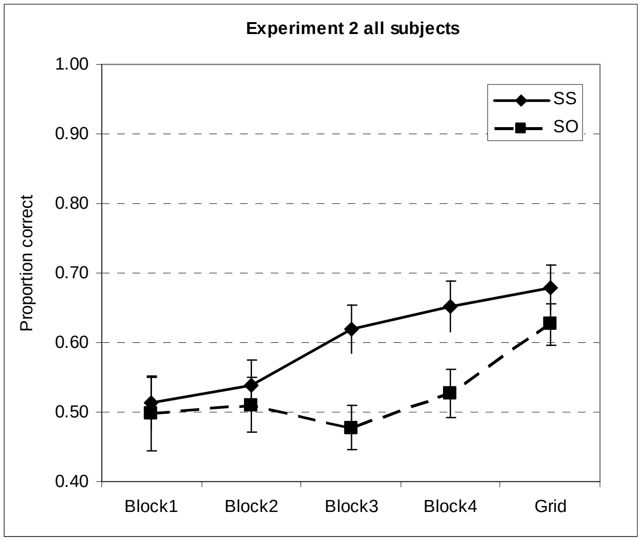

Figure 3. Mean proportion correct across all participants in Experiment 2 for labeled items in each block and grid. Error bars indicate the standard error of the mean.

In contrast to Experiment 1, participants in the semi-supervised condition showed greater accuracy across all blocks and on the final grid. A general linear model treating time (4 successive blocks of 8 labeled items + grid) as a within-subjects factor and learning condition (SS versus SO) as a between-subjects factor revealed reliable main effects of both factors (for time, F(4,192) = 6.8, p < 0.001; for learning condition, F(1,48) = 4.32, p < 0.05) and no interaction between them (F(4, 192) = 1.2, p = 0.32).

As previously we also computed the number of participants who performed to a criterion of 67% or better on the final grid in each condition. In the SS condition, more than half the participants exceeded this criterion (13/25) whereas less than a third did in the SO condition (8/25). These odds are different with likelihood p<0.08 according to a one-tailed test of the log odds ratio.

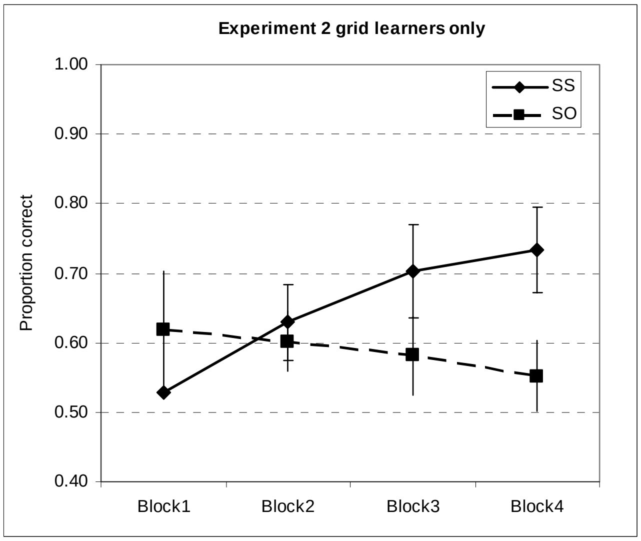

Figure 4. Mean proportion correct in Experiment 2 for participants who performed above criterion in the final grid. Error bars indicate the standard error of the mean.

Finally, we again considered mean accuracy over successive blocks of labeled items in just the participants who performed to criterion according to their grid accuracy. In these participants performance was much better in the SS than the SO group, with accuracy on labeled items improving from 50% to 73% for learners in the SS group but not exceeding chance on any block in the SO group. A general linear model of these data showed no reliable main effect of time or learning condition but these factors did interact significantly (F(3,60) = 2.8, p < 0.05) . Inspection of Figure 4, which plots these data, explains the absence of any main effect and the interaction: performance did not improve significantly at all for the 8 participants in the SO group who performed above criterion on the final grid, but did improve substantially for those participants in the SS group. Consistent with these observations, oneway repeated-measures ANOVAS conducted separately for the two groups showed significantly different accuracy across blocks for learners in the SS condition (F(3,36) = 4.6, p < 0.009) but not in the SO condition (F(3,24) = 0.3, p = 0.83).

In sum, when responses were speeded, providing little time for strategic control of attention, participants in the SS condition performed more accurately overall, were marginally more likely to learn to criterion, and learned labeled items more rapidly than participants in the SO condition.

Discussion

In two experiments we assessed whether the ability to learn a simple 2D binary classification task is influenced by unlabeled experiences. In the first experiment, where participants responded with no time pressure, we observed little evidence that unlabeled data matter: participants performed equally well, were equally likely to learn, and learned equally rapidly regardless of whether they received unlabeled learning items. In the second experiment, which was identical in all respects except that participants were pressured to respond rapidly, we observed a very different pattern: in this case, experience with unlabeled items led to better overall performance, a greater likelihood of learning to criterion, and more rapid learning compared with supervised learning only. Like Vandist et al (2009), we found little evidence that unlabeled data influence category learning when response times were unconstrained. When responses were speeded, we replicated Zhu et al.'s (2007) finding that unlabeled data can produce substantial effects. What accounts for these different patterns?

One possibility concerns the extent to which participants can selectively attend to only some of the stimulus feature dimensions. Prior work has shown that, in categorization tasks where it is possible for participants to form an explicit categorization rule, learning depends importantly upon mechanisms of attention and cognitive control (Ashby and Maddox, 2005). In Zhu et al.'s (2007) work, stimuli varied along a line in a complex multidimensional feature space—therefore it was impossible for participants to selectively attend to information that was irrelevant to the category learning task. In contrast, in Vandist's et al.'s (2009) work and the current study, stimuli varied in two psychologically separable dimensions. If participants selectively attended to only one of these, so that distributional information about the unattended dimension was not available to the learning system, effects of unlabelled data might be attenuated or eliminated—producing the null result in Vandist's (2009) work and in Experiment 1.

On this hypothesis, the robust influence of unlabeled data in Experiment 2 was observed because participants lacked sufficient time to selectively attend to just one feature dimension. If, under speeded conditions, both stimulus dimensions are fully represented, then the unlabeled distribution should have a more robust impact on learning. On this view, it is not the speed of response that matters *per se*, but whether or not the learning system has access to all of the relevant distributional information. If this account is correct, it predicts that unlabeled data should have a stronger effect for multidimensional stimuli where the stimulus dimensions are not psychologically separable, even if response times are unconstrained. We leave this prediction to future work.

We further note that, because there are many factors that differentiate Zhu et al's (2007) study from that of Vandist and colleagues (2009), there remain several additional hypotheses about the difference in their findings. The current study isolates speed of response as an important mitigating factor, but other potentially important factors—including the orientation of the category boundary in the stimulus space, the ratio of labeled to unlabeled examples, and the temporal distribution of labeled examples over the learning session—should be parametrically explored in future work.

More generally, the question of whether or not people make use of unlabeled observations when learning categories has strong implications for theories of human conceptual knowledge. Many researchers have noted that even young children are able, with just a handful of learning experiences, to infer the extension of many category labels (Hall and Waxman, 2004; Keil, 1979; Markman, 1989). Once they reach the right age, most children need hear the word "horse" only once or twice before being able to make a reasonable guess about which objects in the world are horses and which not. This rapid learning from sparse data is sometimes held to indicate that children bring strong inductive biases to bear on word-learning (Xu and Tenenbaum, 2007).

Semi-supervised learning suggests a different explanation: Maybe children can learn from just a few labeled examples because they are marrying these sparse episodes to knowledge gleaned from a vast amount of unsupervised experience. If children assume that category labels tend to span relatively dense clusters in a conceptual feature space, and that category boundaries follow the low-density valleys in this space, then—to the extent that this assumption holds—they only need a small number of labeled experiences to work out which labels "go with" which clusters. This explanation frees theories of word-learning from having to rely too heavily on strong inductive biases to explain rapid word-learning abilities in children.

Acknowledgements

This work was supported in part by a grant from the Air Force Office of Scientific Research (AFOSR project number FA9550-09-1-0313) and in part by NSF project IIS-0916038.

References

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

Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56, 149-178.

Bruner, J. S., Goodnow, J. J. and Austin, G. A. (1956). *A Study of Thinking*. Hoboken NJ: John Wiley and Sons.

Chapelle, O., Zien, A. and Scholkopf, B. (2006). *Semi-Supervised Learning*. Cambridge, MA: MIT Press.

Fried, L. S. and Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10 (2), 234-257.

Gluck, M. A. and Bower, G. H. From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117(3), 227-247.

Hall, D. G. and Waxman, S. R., Eds. (2004). *Weaving a Lexicon*. Cambridge, MA: MIT Press.

Keil, F. C. (1979). *Semantic and Conceptual Development: An Ontological Perspective*. Cambridge, MA: Harvard University Press.

Kruschke, J. K. (1992). An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 22-44.

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

Markman, E. M. (1989) *Categorization and Naming in Children*: Cambridge, MA: MIT Press.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39-57.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, 15, 461-508.

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

Xu, F. and Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, 114, 245-272.

Zhu, X. and Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. San Rafael: Morgan and Claypool.

Zhu, X., Rogers, T. T., Qian, R., and Kalish, C. (2007). Humans perform semi-supervised classification too. Proceedings of *AAAI 2007.*