Semi-supervised learning: A role for similarity in generalization-based learning of relational categories

John D. Patterson (jpatter4@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East
Binghamton, NY 13905 USA

Abstract

Research on semi-supervised category learning has been sparse despite its representativeness of naturalistic category learning and potential applications. Most of the semi-supervised literature to date has focused on establishing the phenomenon. These efforts have produced mixed results and have explored a relatively limited set of learning circumstances. In the current work, we contribute a novel investigation of semi-supervised learning by extending the paradigm to relational category learning and evaluating the role that item similarity plays in the effectiveness of unsupervised learning opportunities. Our results show first-ever evidence of semi-supervised learning in the induction of relational categories and, further, that the similarity between supervised and unsupervised examples critically dictates whether benefits of unsupervised exposures accrue. We conclude with implications and future directions.

Keywords: semi-supervised learning; relational categories; similarity; classification learning; transfer

Introduction

A central goal of human category learning research is to understand what influences the quality, nature, and utility of the category representations we acquire. Research in pursuit of this goal has attained a respectable degree of breadth – ranging from the effect of category structure (e.g., Shepard, Hovland, & Jenkins, 1961), to learning mode (e.g., *classification vs. observational*: Estes, 1994; Levering & Kurtz, 2015; *classification vs. inference*: Jones & Ross, 2011; Yamauchi & Markman, 1998), to whether learning benefits more from blocking or interleaving categories during training (e.g., Carvalho & Goldstone, 2017).

Despite the literature's impressive breadth in many respects, it has been lopsidedly deep when it comes to the issue of supervision. A clear majority of studies investigate learning under full supervision. Within the context of the canonical classification learning task, supervised learning (SL) refers circumstances under which learners receive both a complete example and its associated, experimenter-defined class label on each learning trial. While SL has enjoyed much attention in the literature, a comparatively small amount of non-SL research has been conducted. Further, the majority of non-SL studies have pertained to fully unsupervised learning tasks such as 'free classification' or 'restricted classification' (Garner, 1974) where the learner must construct her own basis for what things go together (e.g., Medin, Wattenmaker, & Hampson, 1987).

The value of the SL and UL research programs is clear. There are many real-world circumstances – such as pedagogical settings – in which we are informed of a target's class membership. Additionally, there are also many times when we must group things together on our own without external indication about how they should be organized. In an ecological sense, however, each of these programs of research is estranged from the reality of how we learn most categories – that is, through some supervised experiences that are nested in a much broader context of unsupervised experiences. The integration of supervised and unsupervised learning experiences is known as semi-supervised learning (SSL).

The human SSL literature is a nascent, but critical, area for research. Besides serving to basic research interests, the study of SSL is highly relevant to educational applications. Given that a key goal of education is to provide learners with a set of learning experiences that enable their continued learning (Bransford & Schwartz, 1999), research that elucidates how to structure supervised learning to maximize subsequent unsupervised learning is in direct service of this goal. It should also be noted that SSL research has been an important topic in machine learning. As such, research on human SSL holds the potential to positively influence the development of novel machine learning algorithms.

What is known about human SSL? Given the recency of this research area, the findings are somewhat sparse. In fact, most of the literature hitherto has been devoted to establishing the existence of the phenomenon – that category representations formed from supervised experiences are impacted by unsupervised ones or vice versa (Gibson, Rogers, & Zhu, 2013; Kalish, Rogers, Lang, & Zhu, 2011; Lake & McClelland, 2011; McDonnell, Jew, & Gureckis, 2012; Vandist, De Schryver, & Rosseel, 2009; Vong, Navarro, & Perfors, 2016; Zaki & Nosofsky, 2007; Zhu, Rogers, Quian, & Kalish, 2007). Although it presents as uncontroversial that we do capitalize on both supervised and unsupervised learning opportunities, the literature on this topic has been somewhat mixed. Several studies have demonstrated evidence of SSL in category learning (e.g., Kalish et al., 2011; Zhu et al., 2007; Vong et al., 2016) while others have failed to find any compelling evidence (e.g., Vandist et al., 2009; McDonnell et al., 2012). Thus, studies with a novel take on the SSL effect are warranted.

While a few studies have examined factors that impact whether and to what degree SSL occurs – such as the effect of category structure and label ambiguity (Vong et al., 2016)

or the impact of the ratio of supervised to unsupervised examples (Vandist et al., 2009) – the literature's predominant focus on establishing the phenomenon has left it relatively limited in scope. One way in which the literature shows reduced scope is in its exclusive usage of continuous, feature-based categories as the target of learning. Although the study of feature-based categories is integral to our understanding of SSL and category learning generally, much of the category knowledge we possess is not reducible to a feature-based understanding. Instead, a plethora of the categories we are knowledgeable about, such as *positive feedback loop* and *reciprocity*, are abstract and reliant on relationships rather than features. Accordingly, mounting attention has been devoted the study of relational categories (Gentner & Kurtz, 2005; Markman & Stilwell, 2001) – categories whose members belong based on a shared set of relations (i.e., a common relational structure). It should be noted that relational category membership is based on deep, relational commonalities and members of the same category can be quite featurally distinct.

The present investigation represents the first time SSL has been studied using relational categories. It should be noted that studying SSL using relational categories carries some unique benefits, relative to feature-based studies. For one, many of the concepts that are targeted in formal educational settings are relational in nature (e.g., Newton's laws, the concepts of evolution by natural selection; Goldwater & Schalk, 2016). Given this, research on relational category learning holds the potential for high translational value. As a second point, relational categories have been characterized as being rule-like in nature (Gentner & Kurtz, 2005). That is, if a target possesses the requisite set of relationships, and the viewer identifies those relationships, then an item can be classified relatively unambiguously as a member of the category (in a way that is perhaps akin to the 'classical view' of category learning; Murphy, 2002). This may be contrasted with research in the feature-based realm, where membership is demonstrably graded in most cases. In effect, relational categories may more uniformly lend themselves to classifications that can be interpreted as 'correct', which may increase both the impact that unsupervised classifications have on learning and the probability of demonstrating evidence of SSL.

In addition, as an exploratory factor, we also sought to elucidate the role that similarity between the supervised and unsupervised item sets plays in whether SSL occurs. Towards this goal, we included high- and low-similarity SSL groups in our design. We operationalized similarity based on surface characteristics and the spatial orientation of the category-defining core. In the high-similarity group, items on both supervised and unsupervised trials shared a common (rock) domain and spatial orientation – that is, they shared literal similarity. In the low-similarity group, items encountered during unsupervised trials came from a different surface domain (mobiles) and had their relational core reflected over the X-axis. That is, these items were analogically similar to the supervised set. To be clear, although the unsupervised stimuli in the two SSL groups differed on the surface level, the deep, structural aspects of the target categories were equally preserved in both sets. We note one study that has examined the effect of similarity on SSL of feature-based categories (Vong et al., 2016). However, given the difference between the role of surface similarity in feature-based versus relational category learning (predictive vs. non-predictive of membership, respectively), this investigation of similarity will serve as an informative contribution.

In the following between-subjects experiment, learners engaged in one of three classification learning conditions. The supervised-only control engaged in three blocks of supervised classification trials. The two SSL conditions were just like the control, except additional unsupervised classification blocks were inserted after each supervised block. In the high-similarity condition, subjects classified items that were literally similar to the supervised set during unsupervised blocks. In the low-similarity condition, subjects classified items that were only analogously similar to the supervised set during the unsupervised blocks. To be expressly clear, the SSL groups received many more stimulus exposures than the supervised-only group (all of which were unsupervised). Typically, an exposure imbalance is a methodological shortcoming. However, the fundamental question at stake – the SSL effect, for which evidence is currently mixed – is whether unsupervised exposures add anything at all to what is learned through supervision. As such, an exposure imbalance is an integral part of the question and manipulation. Following training, all conditions engaged in a common assessment sequence that consisted of a within-domain test followed by an across-domain transfer test. If learners do indeed integrate supervised and unsupervised experiences, then we should observe a benefit for one or both SSL groups over the supervised-only control. If similarity dictates the degree to which supervised and unsupervised experiences are integrated, then we should see performance differences between the two SSL groups.

Method

Participants

120 undergraduates at Binghamton University participated for partial fulfillment of a course requirement.

Materials

The stimuli used for supervised training – held constant across all participants/conditions – consisted of 24 unique rock arrangements, eight per category. These stimuli have been used in previous relational category learning research (see Kurtz, Boukrina, & Gentner, 2013; see also Patterson & Kurtz, 2015). Each arrangement was made up of rocks that varied in shape, size, color, and spatial location (see Figure 1). Each of these rock arrangements conformed to only one of the three following relational categories: *monotonicity* – embodied by a monotonic decline in rock height going from

left to right in the arrangement, *support* – defined by the presence of one rock being elevated by two other rocks below, and *mirrored stack* – characterized by the presence of two same-color rocks that were of similar size, shape, and stacked vertically. The artificial labels used for these categories in the experiment were Besod, Makif, and Tolar, respectively.

The stimuli used for unsupervised training consisted of two subsets – a high-similarity (rocks) set and a low-similarity (mobiles) set. The high-similarity set was made up of 15 unique rock arrangements that were similar to, but distinct from, those used for supervised training. The low-similarity set consisted of 15 mobile stimuli that were composed of different shapes connected by lines (as if dangling from above) that varied in size, shape, color, and spatial location. Besides coming from a domain with different surface characteristics, the spatial orientation of the category-defining core was reflected over the X-axis in the mobiles, relative to the rocks. Thus, performance on these items was hinged on successful analogical transfer. As with the supervised set, each stimulus in each of the unsupervised sets conformed to only one of the target categories, and both high- and low-similarity sets were balanced by category with five items per category in each set.

The stimuli used at test consisted of within-domain and across-domain transfer items. The within-domain test was used to assess for differences in mastery and near-generalization ability; it consisted of the 24 'old' rock arrangements from the supervised set and 12 'new' rock arrangements not previously experienced by any group. Given our main interest in learners' ability to generalize to new examples, each 'old' item occurred once in the set and each 'new' item occurred twice. The 'old' and 'new' items were randomly interspersed. The across-domain transfer items were also novel to all participants. These items were used to assess for differences in learners' abilities to transfer knowledge to surface-dissimilar examples of the categories. The 36 transfer items were an exact replica of the 24 'old' and 12 'new' rock arrangements from the within-domain test, except each item was mapped into one of eight different domains with different surface characteristics (see Figure 1 for examples). The frequency of each domain was equated across categories. As with the 'new' within-domain items, the across-domain items each occurred twice (totaling 72 trials) and the order of the examples was randomized.

Design and Procedure

Learning phase Participants were randomly assigned to one of the three learning conditions: supervised learning only control (SL-only; n = 38), high-similarity SSL (n = 39), or low-similarity SSL (n = 43). Prior to training, all subjects were given an archaeological cover story and condition-specific instructions that informed them of the upcoming tasks they would engage in – including the test; subjects were encouraged to learn, as best they could, both the class labels and what makes an item belong to a category.

A schematic illustrating the training procedure by condition can be seen in Figure 2. All three conditions engaged in three blocks of supervised classification training. In each block, subjects encountered a new random order of the same 24 rock arrangements. On each supervised learning trial, an item was presented in the center of the screen and remained visible for the duration of the self-paced trial. A query about the item's class membership was presented above the item and three response buttons (one for each category) below it. After registering a guess using the mouse, visual confirmation of the selection was shown, and evaluative feedback was given that included whether the response was correct and the correct category label of the presented item (e.g., Correct! This one is a Makif). Feedback was displayed in green or red for correct and incorrect responses, respectively. Following feedback, participants clicked the screen anywhere to proceed to the next trial. Time to make a classification and evaluate feedback were both unconstrained.

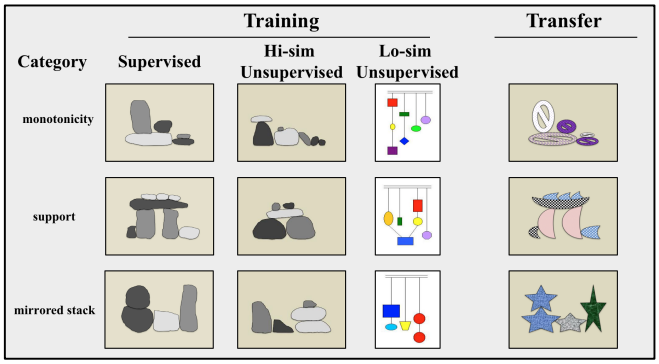

Figure 1: Sample stimuli for each category in each phase. Note: within-domain test items (not shown above) were also rock arrangements.

What distinguished the three conditions from one another was the type of task that followed each of the supervised learning blocks. The two SSL groups received blocks of unsupervised classification trials – three blocks in total. In the high-similarity SSL group learners made two passes through the 15 rock arrangements – a random order each time. Given the mixed results in the literature, we opted for this more heavy-handed approach of making two passes through the set. Thus, each unsupervised block consisted of 30 trials. The structure of each unsupervised trial was identical to that of the supervised trials except no feedback was given following learners' classification decisions. The low-similarity SSL group was just like the high-similarity SSL group except they made two randomized passes through the 15 mobile stimuli instead of the rock stimuli.

In contrast to the SSL groups, the SL-only control group did not engage in unsupervised classification trials. These learners were instead given an 80 second break after each of the three supervised learning blocks. The duration of 80 seconds was chosen based on preliminary data that suggested this was roughly the amount of time SSL learners

would need to complete each unsupervised block. To control for visual exposure, during each break, a distinct 'Where's Waldo' image was centered on the screen and a picture of Waldo was shown in the bottom corner of the screen. Both remained visible for the duration of the break. Learners were instructed to rest and were invited to play 'Where's Waldo' if they felt like it. An additional instruction was included for the Waldo-naïve that taught subjects how to play 'Where's Waldo' (by finding the guy shown in the bottom corner of the screen). After the break time was up, learners were shown where Waldo was for 10 seconds – so as not to leave any Waldo participators frustrated going into the next block

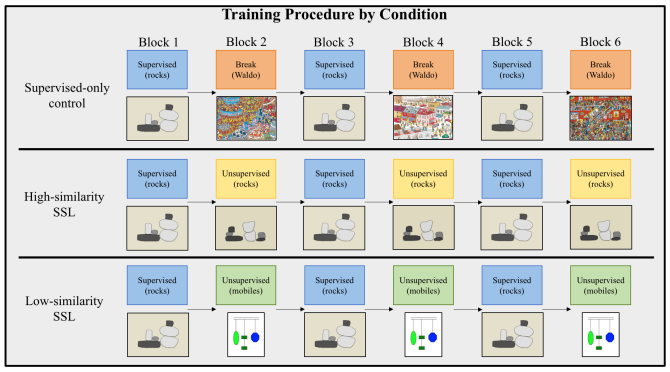

Figure 2: Block-by-block training procedure by condition.

Assessment phase After completing the three supervised classification blocks and the three interleaved blocks of either unsupervised classification (SSL groups) or break time (SL-only group), all conditions performed an identical assessment sequence. Participants were given a notification when they reached the test phase. Learners first received the within-domain test. Upon completion, learners were informed they would then be tested on a different set of items and subsequently began the across-domain transfer assessment. On each trial, in both the within- and across-domain assessments, subjects were asked to classify the presented item and were not given feedback.

Results

The accuracy data were modeled trial-wise using logistic regressions run in the R environment (R Core Team, 2015). Learning phase models included block, learning condition, and their interaction as predictors. Assessment phase models predicted accuracy with learning condition as the lone predictor. Adjusted means and standard errors can be seen in Table 1.

Learning Phase Performance

Unsupervised blocks – SSL groups We first look at performance during the unsupervised learning blocks to evaluate whether high and low-similarity SSL groups differed in their ability to classify the items they were presented. We note two core effects here. First, the effect of block was highly significant (Hi-sim: $\beta = 0.37$, $SE = 0.058$, $Z = 6.50$, $p < .001$; Lo-sim: $\beta = 0.54$, $SE = 0.069$, $Z = 7.95$, $p < .001$), which reflects that participants' accuracy on the unsupervised items increased as they progressed through each unsupervised block. Second, the effect of learning condition was also highly significant (Hi-sim > Lo-sim; $\beta = 0.57$, $SE = 0.052$, $Z = 11.15$, $p < .001$) – showing that learners were better able to make near generalizations, as opposed to more distant transfer. Last, we note a marginal interaction between block and learning condition that suggests that learners improved more in their unsupervised classifications across blocks when they received high-similarity items ($\beta = 0.17$, $SE = 0.09$, $Z = 1.91$, $p = .057$). To the extent that more accurate generalization contributes greater learning, these findings suggest the high-similarity SSL group should demonstrate higher performance than the low-similarity group in the subsequent metrics.

Table 1: Adjusted condition means and standard errors across all performance phases.

	Supervised Block 1	Unsupervised Block 1	Supervised Block 2	Unsupervised Block 2	Supervised Block 3	Unsupervised Block 3	Within-domain Test (old/new)	Across-domain Test
SL-only	.49(.02)	--	72(.01)	--	.80(.01)	--	.83(.01)/ .71(.02)	.74(.01)
Hi-sim SSL	.54(.02)	.68(.01)	74(.01)	.75(.01)	.80(.01)	.80(.01)	.86(.01)/ .76(.01)	.77(.01)
Lo-sim SSL	.53(.02)	.55(.01)	.75(.01)	.63(.01)	.78(.01)	.70(.01)	.83(.01)/ .73(.01)	.74(.01)

Supervised blocks – All groups On the supervised trials, block was the only reliable effect (SL-only: $\beta = 0.99$, $SE = 0.075$, $Z = 13.21$, $p < .001$; Hi-sim: $\beta = 0.84$, $SE = 0.074$, $Z = 11.41$, $p < .001$; Lo-sim: $\beta = 0.81$, $SE = 0.068$, $Z = 11.91$, $p < .001$). This demonstrates that learners became more adept at accurately classifying the exemplars across blocks. We note two additional trends that did not reach significance. First, there was a trend for high-similarity SSL over the SL-only group ($\beta = 0.10$, $SE = 0.061$, $Z = 1.66$, $p = .096$), which hints at a possible benefit to receiving high-similarity unsupervised trials in addition to supervised trials. Second, we found a trend for the interaction between block and learning condition for the SL-only and low-similarity groups ($\beta = -0.17$, $SE = 0.10$, $Z = -1.71$, $p = .086$). The interaction suggests a possibility that learners in the low-similarity SSL group were hindered in their ability to learn across supervised blocks, relative to the SL-only group. In sum, the supervised blocks provide only a weak suggestion that the unsupervised learning experiences exerted an effect on supervised training performance.

Assessment Phase Performance

Within-domain Of critical interest to this investigation was performance at test. On the within-domain assessment (see Figure 3) – consisting of a mixture of the 'old' and 'new' rock arrangements – we found two reliable effects of condition. The reliable advantage for the high-similarity SSL group over the SL-only group ($\beta = 0.26$, $SE = 0.081$, Z

= 3.16, p = .002) indicates that classifying high-similarity category members without feedback confers an added benefit. In addition to this effect, we also observed a reliable advantage for the high-similarity SSL group over the low-similarity SSL group (β = 0.19, SE = 0.079, Z = 2.38, p = .02) – indicating a greater benefit was derived from experiencing near, as opposed to far, members during unsupervised exposure. No differences were found between the SL-only and low-similarity SSL groups.

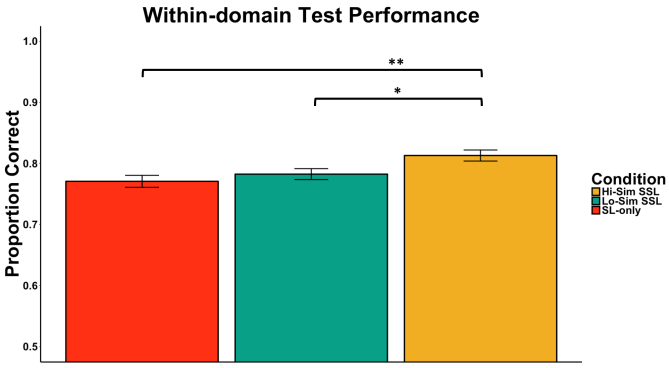

Figure 3: Within-domain performance. Error bars represent +/- 1SE.

We can also examine these data by breaking them down into their separate 'old' and 'new' subcomponents. On the 'old' items, we did not find any reliable effects. However, two trends mirrored the effects found in the overall assessment. First, the high-similarity SSL group exhibited a numerical advantage over the SL-only group that failed to reach significance (β = 0.22, SE = 0.13, Z = 1.72, p = .085). Second, the high-similarity group also showed a numerical advantage over the low-similarity group (β = 0.23, SE = 0.13, Z = 1.84, p = .066). On the 'new' items however, we found a reliable advantage of high-similarity SSL over SL-only (β = 0.29, SE = 0.11, Z = 2.71, p = .007) – suggesting that high-similarity SSL promotes further within-domain generalization.

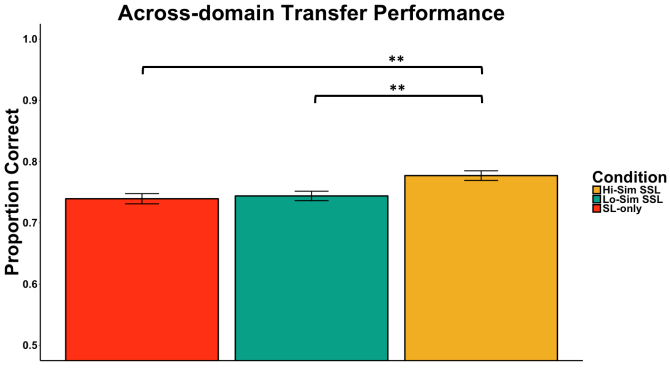

Figure 4: Transfer performance. Error bars represent +/- 1SE.

Across-domain Transfer The across-domain transfer test mirrored the within-domain results (see Figure 4). We saw a highly reliable advantage for the high-similarity SSL group over both the SL-only group (β = 0.21, SE = 0.063, Z = 3.27, p = .001) and the low-similarity SSL group (β = 0.18, SE = 0.061, Z = 2.98, p = .003). These findings provide further evidence that unsupervised learning experiences impact learners' category knowledge and improve their ability to accurately identify new examples of a category in a different domain. However, this benefit only accrues when the unsupervised experiences are similar to those encountered under supervision.

Discussion

There were two primary goals of this study. First, we sought to make a novel contribution to the SSL literature by extending the evaluation of SSL, as a phenomenon, to the realm of relational category learning. Consistent with our predictions, we found compelling evidence that unsupervised encounters exert a marked effect on the quality and portability of category knowledge. Though SSL was not found to affect accuracy on the better-learned, supervised training items, we note a prominent effect of SSL on the extension of category knowledge both to near members and more distant, surface-dissimilar members. To our knowledge, these results represent the first evidence of SSL in relational category learning by induction. Further, these findings corroborate those studies in the feature-based literature that validate SSL as a phenomenon (Lake & McClelland, 2011; Gibson et al., 2013; Kalish et al., 2011; Zaki & Nosofsky, 2007; Zhu et al., 2007).

We also found that the value of unsupervised exposures is not uniform. The second primary goal of this study was to gain insight into the potential role that superficial similarity plays in whether and to what degree SSL occurs. Our findings suggest that similarity is a critical determinant of SSL. Although the high-similarity SSL group showed clear benefits of unsupervised exposures, its low-similarity counterpart performed reliably worse at test and appeared to provide no additional value over SL alone. Understanding this finding will require additional research, however we offer three speculative interpretations. Perhaps the most obvious interpretation is that the analogical mapping was too challenging for learners to make. Without being accurately mapped/classified, it's hard to see how the items might benefit learning. However, the accuracy data seem to cast doubt on this as a full account; learners performed reliably above chance on even the first unsupervised block and achieved a respectable degree of accuracy by the final block. Another possibility is that, although learners performed reasonably well, the higher degree of error they faced (relative to high-sim learners) led to more inappropriate/inaccurate knowledge updates. Under this view, any benefits of unsupervised classifications might be corrupted by inaccurate guesses. Lastly, we note the possibility that learners may have for some reason down-weighted the validity of their unsupervised classifications

(despite achieving accuracy), thereby nullifying any benefit. For example, the distant mappings required in the low-similarity group may have contributed to lower confidence in their classifications. If confidence serves as a moderator to the amount that category knowledge is updated, this could explain the discrepancy between the two SSL groups. Future work should serve to distinguish these possibilities.

There remain many follow-up questions that will need to be addressed to gain a fuller understanding of relational SSL. Are certain aspects of our paradigm critical to whether SSL occurs? For one, we chose the classification learning mode because it requires the making of a committed guess on each unsupervised trial – which could increase the effect of unsupervised exposures. Additional work will need to establish whether the effect is resilient to other more passive learning modes. We also chose to block unsupervised examples together instead of interleaving them with supervised examples. One could imagine a benefit of making temporally close comparisons between supervised and unsupervised examples on subsequent trials. Lastly, it remains to be seen whether the SSL effect and the role of similarity are expertise dependent – do these patterns shift at different stages of learning? Future work will bring the supervision we seek. But, until then, we'll just have to go with our best guess.

References

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*(1), 61-100.

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699.

Estes, W. K. (1994). *Classification and cognition*. Oxford University Press.

Garner, W. R. (1974). *The processing of information and structure*. Psychology Press.

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin*. Washington, DC: American Psychological Association.

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

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin, 142*(7), 729-757.

Jones, E. L., & Ross, B. H. (2011). Classification versus inference learning contrasted with real-world categories. *Memory & Cognition*, *39*(5), 764-777.

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.

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1303–1310.

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

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*(2), 266-282.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*, 329– 358.

McDonnell, J., Jew, C., and Gureckis, T.M. (2012) "Sparse category labels obstruct generalization of category membership." *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*(2), 242-279.

Murphy, G. (2004). *The Big Book of Concepts*. MIT press.

Patterson, J.D., & Kurtz, K.J. (2015). Learning mode and comparison in relational category learning. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

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.

Vong, W. K., Navarro, D. J., & Perfors, A. (2016). The helpfulness of category labels in semi-supervised learning depends on category structure. *Psychonomic Bulletin & Review*, *23*(1), 230-238. doi:10.3758/s13423-015-0857-9

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*(1), 124-148.

Zaki, S. R., & Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition*, *35*(8), 2088-2096.

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 22, No. 1, p. 864). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.