Development of Prototype Abstraction and Exemplar Memorization

Irina Baetu (irina.baetu@mail.mcgill.ca)
Department of Psychology, McGill University, 1205 Penfield Avenue
Montréal, QC H3A 1B1 Canada

Thomas R. Shultz (thomas.shultz@mcgill.ca)
Department of Psychology and School of Computer Science, McGill University, 1205 Penfield Avenue
Montréal, QC H3A 1B1 Canada

Abstract

We present a connectionist model of concept learning that integrates prototype and exemplar effects and reconciles apparently conflicting findings on the development of these effects. Using sibling-descendant cascade-correlation networks, we found that prototype effects were more prominent at the beginning of training and decreased with further training. In contrast, exemplar effects steadily increased with learning. Both kinds of effects were also influenced by category structure. Well-differentiated categories encouraged prototype abstraction while poorly structured categories promoted example memorization.

Keywords: exemplar memorization; prototype abstraction; category structure; neural networks; sibling-descendant cascade-correlation.

Introduction

One of the most fundamental abilities is learning to group things into categories. This faculty allows us to classify new examples and make useful predictions concerning their properties. Two general classes of models have been proposed to account for phenomena in concept learning: prototype and exemplar models. Prototype models claim that experience with items that belong to a given category results in the formation of a summary representation of all the items observed (Posner & Keele, 1968; Reed, 1972). Subsequent categorization of a new item is then based on a comparison between the prototype and the new item. Thus, the more similar a particular instance is to the abstracted prototype, the more likely it is to be classified as a category member (Homa & Cultice, 1984; Homa, Sterling, & Trepel, 1981). In contrast, exemplar models claim that all the observed items are remembered and that the categorization of a new item involves a comparison with items that are stored in memory (Hintzman, 1986).

There is ample evidence in favor of both prototype (Homa, et al., 1981; Posner & Keele, 1968) and exemplar models (Medin & Schaffer, 1978; Palmeri & Nosofsky, 2001), suggesting that both processes are used during category learning. What is more, the relative contribution of each mechanism to categorization might vary across development, as well as during training on a novel task. Early in development, categorization seems to be based on prototype representations while exemplar representations seem to increase with age (Hayes & Taplin, 1993; Mervis & Pani, 1980). There is also evidence that people are more likely to rely on prototypes at the beginning of a categorization task, and as training progresses they rely more on memorized exemplars (Horst, Oakes, & Madole, 2005; Minda & Smith, 2001; Smith & Minda, 1998). These studies are consistent with a shift from early prototype use to later exemplar memorization.

In addition to the amount of experience with a categorization task, category structure also influences which type of information is most used. Better-structured categories can be represented as separate clusters in psychological space, whereas poorly structured categories overlap with each other (Figure 1). Smith and Minda found that better structured categories encourage the early prototype formation, while poorly structured categories discourage it, and may even strongly disadvantage the use of prototypes (Smith & Minda, 1998). Their findings are consistent with a number of other studies (Homa, et al., 1981; Horst, et al., 2005; Reed, 1978).

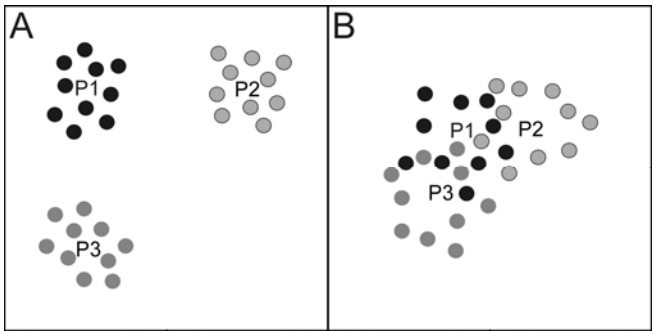

Figure 1: Hypothetical representations of three concepts. P1, P2 and P3 represent three prototypes and the circles represent examples of each concept. A: prototypes are relatively far from each other and examples are tightly clustered around their respective prototype, yielding concepts that are easy to distinguish. B: prototypes are close to each other and examples are more widely dispersed around their respective prototype, resulting in overlapping concepts that are difficult to distinguish.

The aim of this paper is to present a unified model able to simulate prototype and exemplar processes during concept learning. This unified model captures prototype and

exemplar effects with the same mechanism, as opposed to implementing two separate processes. We intend to demonstrate that it is possible for a unified mechanism to capture prototype and exemplar processes to different degrees depending on category structure and amount of training. We present here simulations with sibling-descendant cascade-correlation (SDCC) networks (Baluja & Fahlman, 1994), which offer several demonstrated advantages including automatic network construction, rapid and strong learning, and psychological and neurological plausibility (Shultz, 2003, 2006; Shultz, Mysore, & Quartz, 2007; Shultz, Thivierge, & Laurin, 2008). At the start, SDCC networks are composed of only input and output units. During training, examples were presented to the networks as specific activation patterns in the input layer. In encoder fashion, the networks gradually learned to reproduce this pattern on the output layer by changing the strength of the connections between the units and by recruiting and organizing new hidden units as needed.

In such networks, a relatively small number of units can store a large number of representations, with each representation being a specific pattern of activation across the units. These representations are relatively distributed, as opposed to being localized in single units. Because of its distributed nature, a network is likely to represent similar items as similar patterns of activations on the hidden units. The connection weights between the units reflect all trained items; thus, they represent something similar to a prototype, or an average of the trained concepts. Even if the networks are never presented with the category prototype, they are likely to falsely recognize it because it is so similar to many of the trained items. In addition, because the networks retain some specific information about the trained items, they show a familiarity effect when presented with old items, which is typical of exemplar models (Shultz, *et al.*, 2008).

The networks exhibit a prototype effect if they perform better when presented with examples that are similar to the hypothetical prototype (typical examples) than when they are presented with examples that are less similar to the prototype (atypical examples). We also tested whether the networks memorized some of the features of the trained examples. If our networks become more familiar with the trained examples and perform better when presented with old rather than new examples, regardless of distance from the prototype, then they reveal an exemplar effect.

We studied the impact of category structure and amount of training on prototype and exemplar effects. We manipulated category structure by changing the similarity between the prototypes of the trained categories and the similarity between each example and its prototype. Better-structured categories have more dissimilar prototypes and examples that are more similar to the prototype of their category (in other words, examples that are more tightly clustered around their prototype). To study the impact of training experience, networks were presented with varying numbers of training trials.

Method

As in past work (Shultz, *et al.*, 2008), we trained SDCC networks in encoder mode. Encoder networks learn to encode the input signal onto the hidden units, and then decode that hidden unit signal back onto the output units. Because error is computed as the sum-squared difference between input and output activations, this can be construed as self-supervised learning, without an externally-provided category name as target output. This type of learning occurs when people are not given information about category membership; hence, they can freely create concepts based on their observation of the examples (Homa & Cultice, 1984). In contrast, learning with category labels is much simpler and quicker. In typical encoder fashion, there were no input-output connections in our networks because such connections would have made the learning too simple.

Also as in Shultz *et al.* (2008), we trained the networks with examples belonging to four concepts. Each example varied on ten binary dimensions. A prototype was first constructed by randomly assigning values of 0.5 or -0.5 to each dimension. We refer to it as the prototype of the loner concept because it was relatively isolated from the other three concepts. Another 10-dimensional vector orthogonal to the first one was randomly selected (the normalized inner product between these two vectors was zero). From this orthogonal vector, three prototypes were created by randomly flipping one, two or four values. Flipping a value means reversing its sign. These three prototypes were much closer to each other in the 10-dimensional space than to the loner vector. We refer to them as the trio.

Nineteen examples were created from each prototype by flipping one or several values depending on the condition. Fifteen of these examples were used for training the networks, while four were used only during the test. Out of the fifteen trained examples, ten were closer to the prototype than the other five, i.e. they were created by flipping fewer values. We refer to the examples that were created through fewer flips as the close examples, and to the other ones as the far examples.

For each of the four concept prototypes, we manufactured examples by flipping 1, 2, 4, or 8 values of the prototype, randomly selected without replacement, depending on condition and subject to three additional constraints: (a) each example had a unique combination of features to flip, ensuring example uniqueness, (b) each feature was flipped in at least one example, and (c) no feature was flipped in every example. This last constraint ensured that no defining features were inadvertently created.

Out of the four examples that were used only during the test, two were close and two were far from the prototype. The networks were also tested on four of the trained examples, two that were randomly selected from the close examples, and the other two, from the far examples. Thus, testing consisted of presenting the networks with eight examples: two close trained examples, two far trained examples, two close test examples, and two far test examples. An exemplar effect is established if the networks

perform better on the trained examples than on the new test examples. Superior performance on the close examples versus the far ones demonstrates a prototype effect.

We manipulated the structure of the categories, which was determined by two factors. First, the number of flips that were applied to the vector orthogonal to the loner to create the trio was varied. Applying fewer flips means that the three concepts are closer to each other, while performing more flips means that the concepts are more distinct from one another. Second, we varied the number of flips applied to the loner and the trio to create examples. Fewer flips indicate that the examples are more tightly clustered around their prototype, while more flips imply a more dispersed distribution of the examples. These two manipulations affect the overall distinctiveness of the concepts. The concepts are more separate from one another with more prototype flips and fewer example flips.

Three levels of category structure were defined. The number of flips applied to the vector orthogonal to the loner to create the trio was 4 (Condition Easy), 2 (Condition Intermediate), or 1 (Condition Difficult). The number of flips applied to each prototype to create the close examples was 1 (Condition Easy), 2 (Condition Intermediate), or 4 (Condition Difficult). Finally, the number of flips applied to each prototype to create the far examples was 2 (Condition Easy), 4 (Condition Intermediate), or 8 (Condition Difficult).

The three conditions may be conceptualized as three levels of difficulty of a categorization task. Condition Easy was the easiest task because the examples were tightly distributed around their prototype and the concepts were well-differentiated. Condition Difficult was the hardest task because the examples were widely dispersed around their prototype and the concepts overlapped. Condition Intermediate was an easier task than Condition Difficult, but harder than Condition Easy. The concepts overlapped less than in Condition Difficult, but they were not as well differentiated as in Condition Easy.

To study the influence of training experience, the networks were trained for different numbers of epochs, varying from 5 to 700. An epoch is a training period during which a network is exposed to all trained examples once in random order. The networks were trained for 5, 10, 25, 50, 75, 100, 200, 300, 400 or 700 epochs. Twenty networks were trained for each number of epochs in each of the three conditions, for a total of 600 networks.

Results

We reserve a detailed discussion of all our findings for a longer paper and we describe here only some of the most important results. We chose network error as the dependent measure, error being defined as the sum of the squared differences between inputs and outputs. Because network error is the difference between the input and output patterns, it reflects familiarization with the examples – how well the networks recognize the examples. Thus, lower network error indicates a higher level of familiarization with the examples.

As training progressed, the mean network error decreased in all three conditions, reflecting the networks' increased familiarity with the examples. At the end of training, the mean error for the trained examples approached zero. The mean error for the new test examples was higher than the error for the trained examples, although it had decreased considerably during training. This indicates that the networks learned the trained examples very well, and at the same time generalized their acquired knowledge to the test examples never seen in training..

The most central findings of the simulations are illustrated in Figures 2 and 3. The figures show the prototype and exemplar effects in each condition as a function of the number of epochs.

Figure 2 shows the prototype effect calculated separately for the trained and for the new test examples. We calculated the prototype effect for each network by subtracting the mean error for the close examples from the mean error of the far examples. Thus, the prototype effect on the trained examples is the difference between the error for the far-train examples and the close-train examples. The prototype effect on the test examples is the error difference between the far-test and the close-test examples. A positive difference indicates a prototype effect, that is, smaller error for the examples that are more similar to the prototype.

Figure 3 illustrates the exemplar effect calculated separately for the far and the close examples. We calculated the exemplar effect by subtracting the mean error for the train examples from the mean error of the test examples. The exemplar effect on the close examples is the error difference between the close-test and the close-train examples. The exemplar effect on the far examples is the error difference between the far-test and the far-train examples. A positive difference indicates an exemplar memorization effect, which means that the error is smaller for the trained examples than for the test ones; or, in other words, that the networks are more familiar with examples that have already been encountered than with novel examples.

We performed an ANOVA on the error differences shown in Figure 2 with the within-network factor Train vs. Test Examples and the between-network factors Number of Epochs and Condition. We performed a similar ANOVA on the error differences shown in Figure 3. All main effects and interactions were reliable in both analyses, minimum $F(9, 570) = 4.54$, $p < .001$. We analyzed these effects separately for each condition, and found that all main effects and interactions were significant, minimum $F(9, 190) = 2.49$, $p = .010$, except the main effect of Epoch in Condition Difficult in Figure 2, $F < 1$. Hence, we describe the results without referring to more detailed statistical tests because all the effects we discuss are licensed by these significant main and interactive effects.

Category Structure

The difficulty of the task had a sizeable impact on the prototype effect (Figure 2). The prototype effect was quite

large in Condition Easy and somewhat smaller in Condition Intermediate. This effect was reversed in Condition Difficult as demonstrated by the negative difference scores; networks' error was higher for the close examples than for the far ones. The close examples in Condition Difficult shared a high degree of similarity, causing the networks to easily confuse them with each other. Thus, examples that shared a high degree of similarity with their prototype no longer had an advantage over ones that did not. This finding is consistent with Smith and Minda's (1998) psychological results. They found a reversed prototype effect with poorly structured categories. Thus, the prototype effect diminished and even reversed as the difficulty of the task increased.

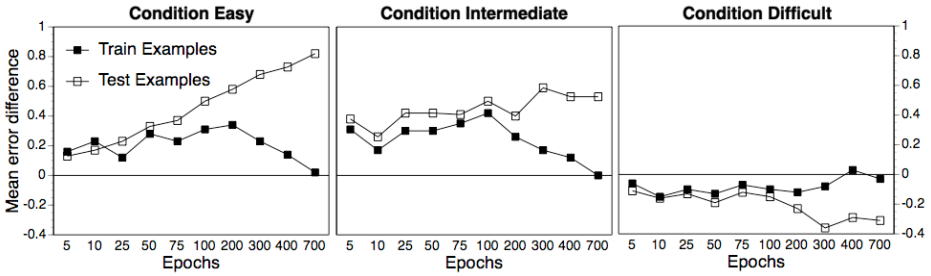

Figure 2: Prototype effect on the trained and the new test examples.

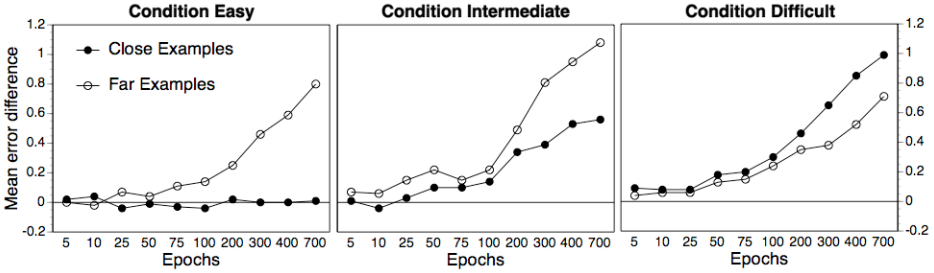

Figure 3: Exemplar effect on the examples that were close and those that were far from the prototype.

In contrast, the exemplar effect increased with the difficulty of the task (Figure 3), which is also consistent with psychological data (Minda & Smith, 2001; Smith & Minda, 1998). The networks relied more on exemplar memorization as the task became increasingly difficult and the prototype representation no longer provided useful information for discriminating the categories.

Amount of Training

The exemplar effect increased with the number of epochs in every condition (Figure 3), simulating Smith and Minda's psychological results (Minda & Smith, 2001; Smith & Minda, 1998). The prototype effect on the trained examples, on the other hand, decreased with the number of epochs in Conditions Easy and Intermediate, but was less affected by the number of training epochs in Condition Difficult. The decreasing prototype effect for the trained examples is consistent with Smith and Minda' results with trained examples. They did not test new examples in their experiments. Our networks make another novel prediction, namely that the prototype effect should increase with training for new test examples, especially if the categorization task is easy (left panel of Figure 2).

Networks became increasingly familiar with trained examples because they could memorize them. As training progressed, networks' recognition of trained examples relied more on individual memories, and less on their similarity to the prototype (just as with Smith and Minda). In contrast, novel examples had not been memorized. Hence, recognition of novel examples relied solely on their similarity to the prototype, and this prototype effect increased during training presumably because the prototype representation became increasingly well-defined.

Interaction Between Exemplar and Prototype Effects

The prototype effect was greater for new test examples than for old, trained ones (Figure 2). This finding seems realistic because only the trained examples could be memorized. Furthermore, the exemplar effect was stronger on the far examples than on the close examples in Conditions Easy and Intermediate (Figure 3, left and middle panels). Features of atypical instances were better remembered than those of typical instances. This presumably occurred because there was less interference between the memories of the atypical examples than between the similar memories of the typical examples. This is consistent with Light, Kayra-Stuart and

Hollander's (1979) finding that adults' recognition memory is better for atypical rather than typical faces. Similar results were found by Going and Read (1974) and Cohen and Carr (1975).

In Condition Difficult (right panel of Figure 3), however, the exemplar effect was larger on the close examples than on the far ones. The close examples were disadvantaged by their similarity to their prototype (because of the overlap between the categories); hence, these examples may have been the ones that benefited most from exemplar memorization. Reitman and Bower (1973) reported a similar effect with adult participants who were trained on an easy or a difficult categorization task. Following training, participants were given a recognition test. The results for the easy task were similar to Light *et al*.'s (1979) psychological results and our simulations in Conditions Easy and Intermediate: recognition performance was better for atypical examples. In contrast, their results for the difficult task were reversed: recognition performance was better for typical examples, matching our simulations in Condition Difficult.

Thus, prototype and exemplar effects seem to complement each other, each process having a stronger influence on the examples that are not favored by the other.

Discussion

We demonstrated that a unified model can capture both prototype and exemplar effects. The networks abstracted concept prototypes and at the same time remembered some features of the trained examples.

Networks also successfully simulated the prototype-to-exemplar trend as the learning task increased in difficulty (Minda & Smith, 2001; Smith & Minda, 1998). Our networks also showed an increase in the size of the exemplar effect from Condition Easy to Condition Difficult, as the concepts became more poorly structured. At the same time, the prototype effect substantially decreased and even reversed as difficulty level increased. For better-structured concepts (Conditions Easy and Intermediate), the exemplar effect was greater farther away from the prototype; for poorly structured concepts (Condition Difficult), the exemplar effect was greater closer to the prototype. As we mentioned earlier, this is consistent with a number of psychological studies.

The networks also exhibited a shift from prototype use to exemplar memorization during training. We observed an increase in the exemplar effect and a decrease in the prototype effect on the trained examples. Better memorization with more training makes perfect sense, as memorization depends on the amount of experience. A possible reason for the decrease in the use of prototype information for the trained examples is that it is less needed as the examples are better remembered. This is consistent with psychological studies reviewed earlier (Hayes & Taplin, 1993; Horst, *et al*., 2005; Mervis & Pani, 1980; Minda & Smith, 2001; Smith & Minda, 1998).

Other studies, however, reported that exemplar information is used earlier in development, and the ability to abstract a prototype emerges later (Fisher & Sloutsky, 2005; Sloutsky & Fisher, 2004; Tighe, Tighe, & Schechter, 1975). Fisher and Sloutsky (2005), for instance, found that younger children's memory for trained items was significantly better than that of older children and adults, suggesting that the latter relied more on an average prototype representation.

It is important to note a key difference with these studies. The studies finding an exemplar-to-prototype shift used concepts with defining features, while those that found a prototype-to-exemplar shift did not (and neither did our simulations). Defining features are present in all examples that belong to a category, and only in those, allowing perfect categorization performance. For example, Tighe *et al.* (1975) used a word classification task in which names of animals belonged to one category, while body parts belonged to another. Following this classification task, adults were less likely to correctly recognize a previously encountered example than children. Tighe *et al.* proposed that adult participants used the defining feature as an encoding device and learned less about the other features of the words. In contrast, children are less likely to use defining features (Keil & Batterman, 1984), which may result in better memorization of the probabilistic features.

Interestingly, Shultz *et al.* (2008) successfully simulated this shift from probabilistic feature learning to the use of defining features using the same kind of networks presented here. To test the hypothesis that defining features affect exemplar memorization in the present work, we repeated the simulations for Condition Intermediate, but added two defining features to each example. Although exemplar memorization did not decrease with training (on the contrary, it increased), overall network error was higher in the simulations with defining features. These networks were less familiar with the trained examples than if they had been trained without defining features. This is consistent with Tighe *et al.*'s (1975) finding that adults, who use defining features more readily than children, exhibit poorer recognition performance. This explains why Tighe *et al.* and other researchers who also used defining features (Fisher & Sloutsky, 2005; Sloutsky & Fisher, 2004) found better memorization of exemplars in children than in adults.

To conclude, our simulations further decrease the gap between the numerous incongruent studies reported in the literature regarding the development of exemplar and prototype effects during category learning. Indeed, considering factors such as the structure of the categories and the presence of defining features, there is considerable, unexpected coherence in these mixed results. Most importantly, we have demonstrated that it is possible for a single mechanism to capture a gradual shift in concept processing depending on task difficulty and the amount of experience.

Acknowledgments

This research was supported by a postgraduate fellowship to IB and a grant to TRS from the Natural Sciences and Engineering Research Council of Canada.

References

Baluja, S., & Fahlman, S. E. (1994). *Reducing network depth in the cascade-correlation learning architecture. Tech Report CMU-CS-94-209*: School of Computer Science, Carnegie Mellon University.

Cohen, M. E., & Carr, W. J. (1975). Facial recognition and Von Restorff effect. *Bulletin of the Psychonomic Society, 6*(4), 383-384.

Fisher, A. V., & Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. *Child Development, 76*(3), 583-597.

Going, M., & Read, J. D. (1974). Effects of uniqueness, sex of subject, and sex of photograph on facial recognition. *Perceptual and Motor Skills, 39*(1), 109-110.

Hayes, B. K., & Taplin, J. E. (1993). Developmental differences in the use of prototype and exemplar-specific information. *Journal of Experimental Child Psychology, 55*(3), 329-352.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning Memory and Cognition, 10*(1), 83-94.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7*(6), 418-439.

Horst, J. S., Oakes, L. M., & Madole, K. L. (2005). What does it look like and what can it do? Category structure influences how infants categorize. *Child Development, 76*(3), 614-631.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior, 23*(2), 221-236.

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5*(3), 212-228.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207-238.

Mervis, C. B., & Pani, J. R. (1980). Acquisition of basic object categories. *Cognitive Psychology, 12*(4), 496-522.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning Memory and Cognition, 27*(3), 775-799.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 54*(1), 197-235.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3), 353-363.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Reed, S. K. (1978). Category vs item learning: Implications for categorization models. *Memory & Cognition, 6*(6), 612-621.

Reitman, J. S., & Bower, G. H. (1973). Storage and later recognition of exemplars of concepts. *Cognitive Psychology, 4*(2), 194-206.

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.

Shultz, T. R. (2006). Constructive learning in the modeling of psychological development. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and performance XXI*. (pp. 61-86). Oxford, UK: Oxford University Press.

Shultz, T. R., Mysore, S. P., & Quartz, S. R. (2007). Why let networks grow? In D. Mareschal, S. Sirois, G. Westermann & M. H. Johnson (Eds.), *Neuroconstructivism: Perspectives and prospects* (Vol. 2, pp. 65-98). Oxford, UK: Oxford University Press.

Shultz, T. R., Thivierge, J.-P., & Laurin, K. (2008). Acquisition of concepts with characteristic and defining features. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 531-536.

Sloutsky, V. M., & Fisher, A. V. (2004). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science, 15*(8), 553-558.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning Memory and Cognition, 24*(6), 1411-1436.

Tighe, T. J., Tighe, L. S., & Schechter, J. (1975). Memory for instances and categories in children and adults. *Journal of Experimental Child Psychology, 20*(1), 22-37.