Shultz, T. R., Thivierge, J. P., & Laurin, K. (2008). Acquisition of concepts with characteristic and defining features.
In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 531-536).
Austin, TX: Cognitive Science Society.

Acquisition of Concepts with Characteristic and Defining Features

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

Jean-Philippe Thivierge (jthivier@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University, 1101 East Tenth Street
Bloomington, IN 47408 USA

Kristin Laurin (klaurin@artsmail.uwaterloo.ca)
Department of Psychology, University of Waterloo, 200 University Avenue West
Waterloo, ON N2L 3G1 Canada

Abstract

Concepts with defining features have sufficiently rigid boundaries that examples of one concept are rarely confused with another concept, whereas probabilistic concepts have vague boundaries, producing frequent misclassifications. It has been argued that, although neural learning might account for probabilistic (or fuzzy) concepts, it is incapable of accounting for acquisition of concepts with defining features. Simulations reported here with encoder networks call this view into question. These sibling-descendant cascade-correlation networks account for a variety of phenomena with concepts positioned along the probabilistic to defining-features continuum. Classification of exemplars by our networks was affected by both isolation of the concept in semantic feature space and by dispersion of its examples around a prototype. The detection of defining features cured networks from the effects of conceptual crowding and example dispersion. Simulations captured the developmental shift from characteristic to defining features in a natural way.

Keywords: Characteristic-to-defining shift; concept acquisition; defining features; neural networks.

Introduction

A major dichotomy has been proposed between concepts with characteristic versus defining features. Concepts with defining features have such rigid boundaries that exemplars of one are only rarely confused with another. In contrast, probabilistic concepts have vague and overlapping boundaries, often making exemplars difficult to classify.

It has been argued that, although neural learning might provide an adequate account of fuzzy concepts, it cannot account for the acquisition and representation of concepts with defining features, such as kinship terms (Armstrong, Gleitman, & Gleitman, 1983; Pinker, 1997). Some suggest instead dualistic theories – a neural theory for probabilistic concepts and a symbolic theory for defining-feature concepts (Pinker & Prince, 1996). Moreover, there is an interesting developmental shift from emphasis on characteristic to defining features (Keil & Batterman, 1984).

The dual-theory approach has a number of problems that have not been solved. One problem is that a person would not likely know which concept to access, the defining-feature one or the probabilistic one, for any particular example or context. Such confusion would be particularly acute early in concept acquisition when neither version of a concept is fully in place. Another problem with the dual view is difficulty in accounting for the developmental shift from probabilistic to defining features. If defining and characteristic versions of concepts reside in different systems that are represented in incompatible formats governed by distinct principles, how do the two systems interact and how does one eventually overtake the other? A similar issue played out in the area of past-tense acquisition, where a dual view could not explain migration between regular and irregular verb classes over the history of English, a migration that was naturally simulated within a homogenous neural network (Hare & Elman, 1995).

In contrast to the dual-system approach, there may be a continuum of concept crispness, perhaps covered by a more unified theory. A concept could have many features, including some that are defining and others that are probabilistic. It would be parsimonious for both sorts of features to be represented in the same fashion, with the likelihood of a feature appearing in any exemplar varying probabilistically. A defining feature appears in examples with a probability of 1, whereas a characteristic feature appears with a still substantial probability of less than 1.

It is reasonable to suppose that exemplar classification is affected by both the isolation of concepts in a multidimensional feature space and by the dispersion of exemplars in this space. That is, fuzzier concepts would be characterized by residing in a relatively crowded region of feature space and by having more widely dispersed examples; crisper concepts by residence in a relatively isolated region of feature space and by limited dispersion of examples. Exemplars of an isolated concept are unlikely to be confused with those of other concepts, whereas exemplars of crowded concepts could easily be misclassified, particularly if they are loosely dispersed around their prototypes. This notion of conceptual crowding is consistent with the idea that concepts are not randomly distributed in a multidimensional feature space, but rather may cluster in regions (Pinker & Prince, 1996), sometimes called consequential regions (Shepard, 1987).

A psychologically-documented case of exemplar dispersion concerns the difference between jars and bottles (Malt,

Sloman, Gennari, Shi, & Wang, 1999). For both Spanish and English speakers, different jars were tightly distributed around a *jar* prototype, whereas different bottles were quite widely dispersed around a *bottle* prototype. For instance, pill bottles, milk bottles, and water-cooler bottles each possess quite different feature values, exhibiting a diversity not found among jars.

We hypothesized that defining features could protect an otherwise fuzzy concept from the problems posed by conceptual crowding and example dispersion. Even if a concept resides in a crowded neighborhood and has widely dispersed examples, its exemplars might be categorized correctly if defining features are discovered. The discovery of defining features, specifying necessary and sufficient conditions for concept membership, projects a crowded concept into a possibly less crowded region of feature space by providing new and useful dimensions to that space.

We report here on simulations of concept learning using a variant of cascade-correlation (CC) encoder networks. CC networks resemble ordinary back-propagation (BP) networks in propagating activation from one layer to the next and learning the values of their connection weights. But CC networks differ from ordinary BP networks by: (a) training weights only one layer at a time, thus avoiding the biologically improbable back-propagation of error signals, (b) growing as well as learning, by recruiting new hidden units as needed, thus simulating the process of synaptogenesis, (c) also employing cross-connections that bypass hidden-unit layers, and (d) using curvature as well as slope information on the error surface in order to adjust connection weights more decisively (Fahlman & Lebiere, 1990). CC networks have been used to simulate a variety of phenomena in psychological development (Shultz, 2003; Shultz, Mysore, & Quartz, 2007).

As with standard BP learning, CC can be used in encoder mode, encoding inputs onto hidden units and then decoding hidden-unit representations onto output units (Shultz, 2003). This implements a kind of recognition memory that does not require a separate teaching signal specifying target output activations. Because encoder solutions can be trivial when direct input-output connections are permitted, we followed the convention of eliminating them in CC encoder networks. Computational and mathematical details of both the BP and CC algorithms can be found elsewhere (Shultz, 2003).

A variant of CC allows the algorithm to install a new hidden unit either on the highest hidden-unit layer (as a sibling) or on its own higher layer (as a descendant) (Baluja & Fahlman, 1994). Descendant units have input weights from all input units and previously installed hidden units. Sibling units do not receive inputs from their siblings previously installed on the same highest level. Sibling-descendant cascade-correlation (SDCC) has so far yielded similar psychological coverage to standard CC but with fewer connection weights, shallower networks, and greater topological variety (Shultz et al., 2007). An example of an SDCC network topology is shown in Figure 3. Four sibling candidates and four descendant candidates compete for being recruited. As with CC, SDCC recruits the candidate whose activations correlate most highly with network error. To counter the natural tendency to recruit a descendant unit (because of its extra weights), descendant correlations are penalized by a default multiplier of 0.8.

We report three simulations with SDCC encoder networks learning to recognize examples of crisp and fuzzy concepts. Simulation 1 examines the effects of concept crowding and example dispersion with probabilistic concepts, simulation 2 assesses the protective effects of defining features against these problems, and simulation 3 focuses on the characteristic-to-defining developmental shift.

Simulation 1: Crowding and Dispersion

The purpose of this first simulation was to test the idea that classification accuracy is disturbed by concept crowding and example dispersion.

Concept Coding

The concepts used in these simulations are abstract and have no particular semantic content, thus affording better experimental control and greater generality than with semantically-laden concepts.

For each network, we began with a random ten-dimensional vector of binary values, either -0.5 or 0.5. Ten is a psychologically realistic number of features because when people were asked to list features of common object concepts, they produced between 6 and 15 features (Rosch, Mervis, Gray, Johnson, & Boyes-Braehm, 1976). This random vector was the prototype for a *loner* prototype concept occupying an isolated region of the ten-dimensional feature space. Next we randomly selected another ten-dimensional vector from those that are orthogonal to the loner prototype. Orthogonality was defined by having a normalized inner product (NIP) of 0 with the loner concept. We created three additional prototypes in a crowded region of the feature space by flipping either 1 or 2 values of the orthogonal vector, depending on experimental condition. The 1 or 2 values flipped were randomly selected without replacement. Values were flipped from -0.5 to 0.5, or from 0.5 to -0.5. This provided four concept prototypes, one of which was isolated and a trio which were close together in the ten-dimensional feature space. This was our manipulation of concept isolation vs. crowding.

For each of these four concept prototypes, we then created ten examples by flipping 1, 2, or 3 values of the prototype, randomly selected without replacement, depending on condition and subject to three further constraints: (a) each example had a unique combination of features to flip, ensuring that each example was unique, (b) each feature was flipped in at least one example, and (c) no feature was flipped in every example. This ensured that no defining features were inadvertently created and it constituted our manipulation of exemplar dispersion.

Network Training

We trained 20 SDCC encoder networks in each of the six experimental conditions (2 levels of prototype flips x 3 levels of example flips), each network starting with random

connection weights and random concept prototypes, for 700 or fewer epochs, an epoch being a pass through all of the training patterns. Training stopped early when a network's outputs were all within score-threshold of their target values for all training patterns. We used the conventional score-threshold value of 0.4, which provides for an uncertain range between -0.1 and 0.1.

These encoder networks learned to recognize examples, not to classify them into their respective prototype concepts. There was no explicit teaching signal to enable classification learning. This is realistic because category labels are not always provided in human concept learning.

Results

After learning, we assessed network classification ability by presenting all 40 training exemplars and computing the NIP between each output vector and each of the four concept prototypes. An exemplar was considered to be classified into the concept yielding the highest NIP. We tallied the number of correctly classified examples out of ten possible for each of the four concepts. The numbers of correct classifications were subjected to a mixed ANOVA where the between-network factors were prototype flips (1 or 2) and example flips (1, 2, or 3), and the within-network factor was concept (with 4 levels: loner plus the trio of close concepts). We report only the largest significant effects.

The importance of concept isolation was revealed by main effects of concept, $F(3, 342) = 67$, $p < .0001$, and number of prototype flips, $F(1, 114) = 39$, $p < .0001$, and an interaction between them, $F(3, 342) = 5.8$, $p < .001$. The means for these main and interaction effects are shown in Figure 1. Classification was more accurate with the loner concept than with the three crowded concepts. Classification was also better when the crowded concepts were relatively less crowded (2 prototype flips versus 1 prototype flip). The interaction reflects that fact that number of prototype flips affected performance on crowded-concept exemplars more than it affected performance on loner-concept exemplars. At both levels of prototype flips, as assessed by paired-samples *t*-tests, performance was better on the loner concept than on each of three trio concepts, $p < .0001$, none of which differed from each other. All means in Figure 1 were well above the chance level of 2.5, computed as ¼ of the number of examples per mean, $p < .01$. This is also true of all other means throughout the paper. Comparisons of obtained means to the theoretical mean of 2.5 were made using Dunnett's (1955) technique.

There was also a main effect of example flips, $F(1, 114) = 256$, $p < .0001$, and an interaction between example flips and concept, $F(6, 342) = 3.6$, $p < .002$. The relevant means, illustrating both crowding and dispersion effects, are shown in Figure 2. Classification accuracy decreased with dispersion of examples (number of example flips). The interaction reflects a larger influence of example dispersion with the crowded trio concepts than with the isolated loner concept. At each level of example flips, performance was better on the loner than on each of three trio concepts, $p < .0001$, none of which differed from each other.

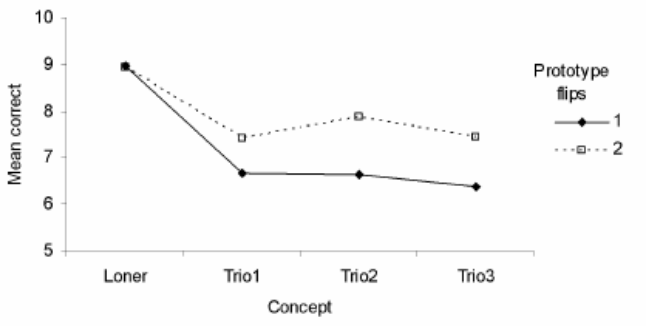

Figure 1: Concept crowding effects.

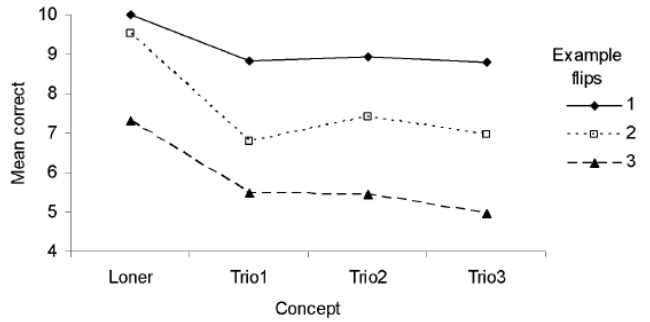

Figure 2: Concept crowding and example dispersion effects.

Most of the 120 networks recruited all of their hidden units on the same level. Five networks created more than one level of hidden units and each of them had two layers of hidden units. Figure 3 shows the topology of a network that recruited three hidden units on one layer and four on a second layer. The arrows in Figure 3 indicate full connectivity between layers. There is also a bias input unit (always with an input of 1.0 and not shown in the figure) connected with trainable weights to all downstream units.

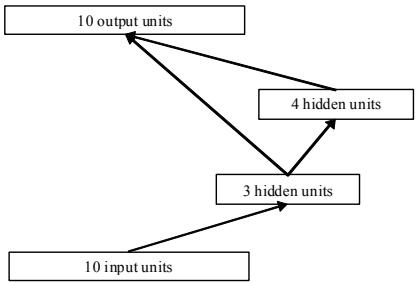

Figure 3: Topology of an SDCC network in the condition with 2 prototype flips and 1 example flip.

Discussion

This simulation documented the expected importance of both concept crowding and example dispersion for classification of exemplars of probabilistic concepts. Classification was more accurate for an isolated concept than for crowded concepts

and more accurate for tightly dispersed examples than widely dispersed examples. Isolated concepts were less affected by example dispersion than more crowded concepts were. Goldstone (1996) reported a similar example-dispersion effect in experiments where people were trained in classification, rather than in recognition memory as here.

Simulation 2: Effects of Defining Features

Simulation 2 was designed to test the idea that defining features would protect otherwise probabilistic concepts against the effects of concept crowding and example dispersion obtained in simulation 1.

Concept Coding and Network Training

After coding concepts as in simulation 1, we gave each concept two additional binary features that uniquely defined the concept. Defining features for the four concepts were coded as (-0.5 -0.5), (-0.5 0.5), (0.5 -0.5), or (0.5 0.5), respectively, by random assignment without replacement of these four binary codes to the four concepts. These defining features were present in each example, and were unaffected by the probabilistic feature flipping used for the other ten features. As in simulation 1, we trained 20 networks in each of the six conditions for 700 or fewer epochs.

Results

After learning, we assessed network classification ability as in simulation 1. The numbers of correct classifications were subjected to a mixed ANOVA that included the results of simulation 1 and an additional between-network factor representing the presence or absence of defining features. All four main effects (defining features, prototype flips, example flips, and concept) were significant, $p < .0001$, and defining features interacted with each of the other three variables. As shown in Figure 4, without defining features, the familiar signatures of concept crowding were evident. Performance on examples from the crowded trio concepts was worse than performance on the isolated concept. With defining features, performance was nearly perfect for every concept, interaction of defining features by concept $F(3, 684) = 45$, $p < .0001$.

There was also an interaction between defining features and prototype flips, $F(1, 228) = 25$, $p < .0001$, the means for which are given in Figure 5. Without defining features, performance was worse with one prototype flip than with two prototype flips. With defining features, this effect of concept crowding was eliminated.

Defining features also dampened the effects of example dispersion, as revealed by the interaction between defining features and number of example flips, $F(2, 228) = 114$, $p < .0001$, shown in Figure 6. Without defining features, performance deteriorated with an increase in example flips. But with defining features, this effect of example dispersion was diminished.

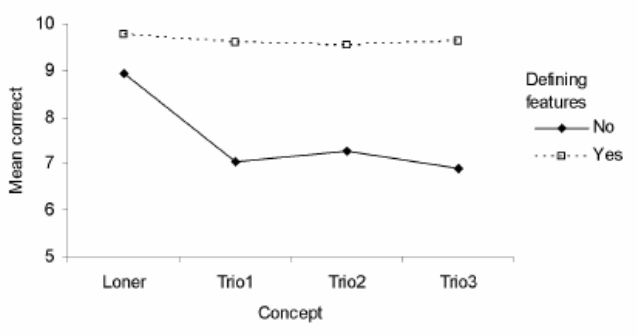

Figure 4: Defining features protect against concept crowding, implemented by concept position.

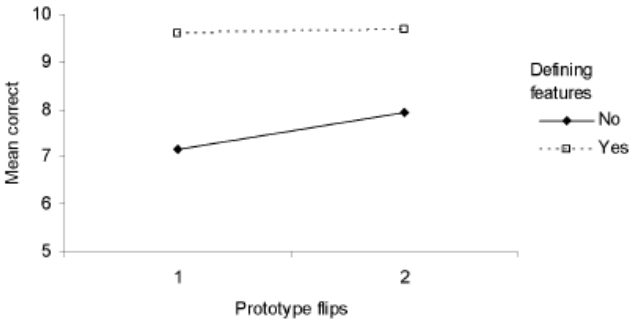

Figure 5: Defining features protect against concept crowding, implemented by prototype flips.

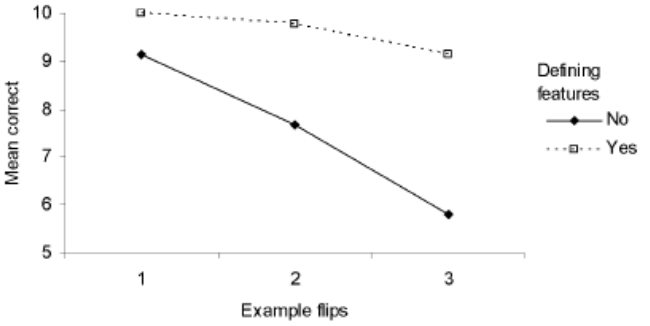

Figure 6: Defining features protect against example dispersion.

Discussion

This simulation showed that defining features protected otherwise probabilistic concepts against the typical effects of concept crowding and example dispersion. Introduction of defining features allowed the fairly precise classification that was supposed to be impossible for neural networks. By treating defining features in the same manner as probabilistic features, neural networks can use defining features effectively to produce nearly perfect classification of examples. Defining features project concepts into new dimensions and thus less-crowded regions of feature space.

There was not a systematic difference in perfect performance between concepts with and without defining features, as would presumably be predicted by dual theories.

Instead, there was perfect performance in isolated probabilistic concepts with tightly dispersed examples, and less than perfect performance in all defining-feature concepts with more widely dispersed examples. It would be interesting to see how well these simulations predict human results in these respects.

Simulation 3: Developmental Shift from Characteristic to Defining Features

Here we examine whether these networks capture the characteristic-to-defining developmental shift. Keil and Batterman (1984) reported that, with increasing age, children shift from using characteristic features to defining features in those concepts that have defining features. As far as we know, this effect had never been computationally simulated, although Rogers and McClelland (2004) showed that neural networks can simulate the progressive differentiation of concepts (Keil, 1979) and the tendency for concepts to be organized around causal knowledge that explains correlations of features (Keil, 1991).

On the assumption that young children have less experience with examples than do older children, we simulated the characteristic-to-defining shift by comparing networks early in training to networks late in training. As in the studies with children, this was a cross-sectional, rather than a longitudinal, research design.

Concept Coding and Network Training

Concepts were coded as with the defining-features concepts of simulation 2, using 10 probabilistic features and 2 defining features. Twenty networks in each of the six conditions were trained up to a maximum of 100 epochs. These networks were then compared to the more fully trained networks of simulation 2.

Results

Numbers of correctly-classified examples out of 10 were subjected to a mixed ANOVA in which maximum epochs was added as a between-network factor. Number of defining features was dropped from the ANOVA because all concepts here had defining features.

The ANOVA yielded main effects of concept, $F(3, 684) = 22$, $p < .0001$, and maximum epochs, $F(1, 228) = 1173$, $p < .0001$, and an interaction between them, $F(3, 684) = 17$, $p < .0001$. The relevant means are shown in Figure 7. At 100 epochs, there was a crowding effect, with more accurate classification of exemplars of the loner concept than of the three crowded concepts. Here, as assessed by paired-samples *t*-tests, performance was better on the loner concept than on each of three trio concepts, $p < .0001$, none of which differed from each other. At 700 epochs, there was good protection against concept crowding due to eventual mastery of the defining features.

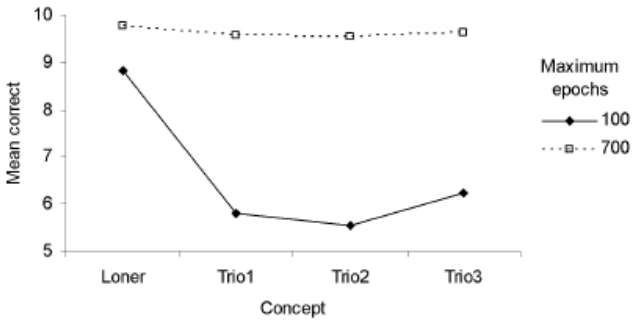

Figure 7: Characteristic-to-defining features shift for concept crowding.

Similar results were obtained for example dispersion. There was a main effect of example flips, $F(2, 228) = 131$, $p < .0001$, and an interaction of that with maximum epochs, $F(2, 228) = 38$, $p < .0001$. The means are presented in Figure 8. At 100 epochs, example dispersion produced classification errors, but there was relative protection against example dispersion at 700 epochs due to mastery of the defining features.

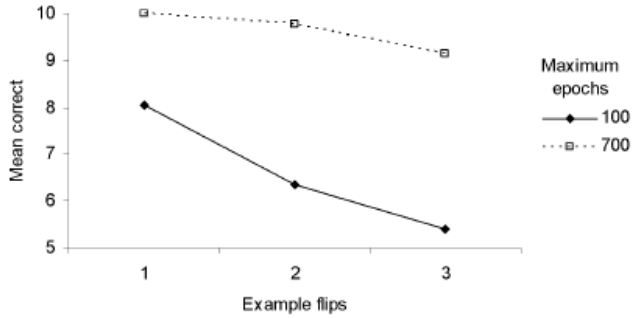

Figure 8. Characteristic-to-defining features shift for example dispersion

The mean number of hidden units recruited in the networks stopped at 100 epochs was 1.95, with all but 5 of the 120 networks recruiting two hidden units. All hidden units were installed as siblings, on a single level.

Discussion

This simulation shows that these networks can naturally capture the developmental shift from using characteristic to defining features. Early in the learning of concepts with both defining and probabilistic features, effects of concept crowding and example dispersion are evident. Later, as the defining features are discovered, they protect classification performance against these probabilistic effects. This developmental change shows that defining features cure the earlier effects of crowding and dispersion rather than immunize against them.

In many natural concepts, the number of defining features, if there are any, is likely to be smaller than the number of probabilistic features. In a sea of probabilistic features, it may take considerable experience to eventually discover the defining features of a concept. The hypothesis of essentialism

is that concepts with defining features may have an empty place holder (Medin & Ortony, 1989). People may believe that a concept has an essence provided by its defining features, but they may not yet know what that essence is. Despite using probabilistic information to classify examples of the concept, people may remain willing to yield to more expert opinion about defining features.

The basis for this developmental shift in networks is the inherent tendency of neural networks to reduce as much error as possible. This tendency, in turn, derives from a weight-adjustment rule specifying that the amount of weight change is proportional to the derivative of error with respect to weight. Because probabilistic features are more numerous than defining features, probabilistic features naturally capture more initial attention from the weight-adjustment process. As the networks learn to reduce error as much as possible from attending to ubiquitous probabilistic features, they eventually attend to the small number of defining features, yielding more precise classification.

General Discussion

These simulations show that SDCC encoder networks can capture sensible phenomena regarding both crisp and fuzzy concepts. Simulation 1 revealed that classification errors increase as concepts become crowded in multidimensional feature space and as examples become more widely dispersed in that space. Simulation 2 demonstrated that the eventual discovery of defining features can protect networks against these concept-crowding and example-dispersion effects. Simulation 3 captured the developmental shift from emphasizing characteristic to emphasizing defining features, in concepts that possess defining features. Initially, networks learning concepts with defining features showed the crowding and dispersion effects seen with probabilistic concepts. Continued learning allowed eventual detection of the defining features and eliminated the earlier effects of concept crowding and example dispersion. As with the work of Rogers and McClelland (2004), our results suggest that artificial neural-network models can cover and explain important psychological findings on concept acquisition.

In current work, we are exploring whether our model can also capture some other, well-known psychological phenomena in the concept-acquisition literature, namely pattern completion and prototype and exemplar effects. Also of current interest is whether our model could account for the conflicted literature on the relative developmental order of exemplar and prototype effects.

Acknowledgments

This research was supported by a grant to TRS from the Natural Sciences and Engineering Research Council of Canada, and a Doctoral Fellowship to JPT from the Fonds Québéçois sur la Nature et les Technologies.

References

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*, 263-308.

Baluja, S., & Fahlman, S. E. (1994). *Reducing network depth in the cascade-correlation learning architecture.* (No. Technical Report CMU-CS-94-209). Pittsburgh, PA: School of Computer Science, Carnegie Mellon University.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50*, 1096-1121.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 524-532). Los Altos, CA: Morgan Kaufmann.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24*, 608-628.

Hare, M., & Elman, J. L. (1995). Learning and morphological change. *Cognition, 56*, 61-98.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective*. Cambridge, MA: Harvard University Press.

Keil, F. C. (1991). The emergence of theoretical beliefs as constraints on concepts. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition*. Hillsdale, NJ: Lawrence Erlbaum.

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*, 221-236.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40*, 230-262.

Pinker, S. (1997). *How the mind works*. New York: Norton.

Pinker, S., & Prince, A. (1996). The nature of human concepts: Evidence from an unusual source. *Communication & Cognition, 29*, 307-362.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rosch, E., Mervis, C., Gray, D., Johnson, D., & Boyes-Braehm, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 3*, 382-439.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

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