A Connectionist Account of Interference Effects in Early Infant Memory and Categorization[1]

(In *Proceedings of the 19th Annual Cognitive Science Society Conference*, NJ:LEA, (1997). pp 484-489)

Denis Mareschal
Department of Psychology
Exeter University
Perry Rd., Exeter
EX4 4QG, UK
d.mareschal@exeter.ac.uk

Robert M. French
Psychology Department, B32
Université de Liège
4000 Liège,
Belgium
rfrench@ulg.ac.be

Abstract

An unusual asymmetry has been observed in natural category formation in infants (Quinn, Eimas, and Rosenkrantz, 1993). Infants who are initially exposed to a series of pictures of cats and then are shown a dog and a novel cat, show significantly more interest in the dog than in the cat. However, when the order of presentation is reversed — dogs are seen first, then a cat and a novel dog — the cat attracts no more attention than the dog. We show that a simple connectionist network can model this unexpected learning asymmetry and propose that this asymmetry arises naturally from the asymmetric overlaps of the feature distributions of the two categories. The values of the cat features are subsumed by those of dog features, but not vice-versa. The autoencoder used for the experiments presented in this paper also reproduces exclusivity effects in the two categories as well the reported effect of catastrophic interference of dogs on previously learned cats, but not vice-versa. The results of the modeling suggest connectionist methods are ideal for exploring early infant knowledge acquisition.

Introduction

Memory and categorisation lie so deeply at the heart of human cognition that they are to be found in even very young infants. Newborns can remember visually presented information over long retention intervals (Slater, 1995) Very young infants can also be shown to separate complex visually presented stimuli into distinct categories (Quinn & Eimas, 1996). Categorisation is a means of reducing the load on memory (Rosch, 1975). It remains intimately related to memory at all ages. Although a number of connectionist models of adult memory and categorisation have been published (e.g., Shanks, 1991; Knapp & Anderson, 1984) no attempts have been made to extend these models (or to devise new models) in order to account for the particularities of both infant memory and categorisation. The one partial exception is Quinn & Johnson (1996) who model hierarchy effects in the acquisition of concepts during infancy.

In this paper, we present a simple connectionist model of memory and categorisation in early infancy that targets behaviors specific to that age range. In particular, we focus on:(a) the ability to categorize complex visual stimuli, (b) the asymmetry effect in early categorisation, and (c) interference effects in early memory. By using a mechanism that provides a good account of adult performance, this model underscores the continuity that exists between early (pre-linguistic) memory and categorisation abilities and that of mature adults.

Quinn and Eimas have shown an unexpected asymmetry in category learning in young infants. (e.g., Quinn, Eimas, & Rosenkrantz, 1993; Eimas, Quinn, & Cown, 1994). When 3- to 4-month-old infants are shown different photographs of either cats or dogs they can form perceptual categories of either groups of pictures. Infants who are first shown a number of different photographs of cats and are then shown a picture of a dog along with a novel picture of a cat will be much more attentive to the dog than the novel cat. This is interpreted as showing that the infants have formed a category of Cat that excludes dogs. In sharp contrast, infants who are first shown different photographs of dogs and are then shown a picture of a cat along with a novel dog will not be preferentially attentive to either picture. This is interpreted as showing that infants have formed a category of Dog that *includes* cats. Hence infants show an exclusivity asymmetry in the development of some perceptual categories. Here, the Dog category does not exclude cats whereas the Cat category excludes dogs.

Another unexpected finding has to do with infant memory. Although infants clearly show long-term retention of visual stimuli, under some conditions the presentation of intervening material during the retention interval leads to catastrophic interference with the original material completely eradicated (e.g., Cohen & Gelber, 1975; Deloache, 1976; McCall, Kennedy, & Dodds, 1978). Interference effects decrease with age but continue well into later infancy (Rovee-Collier & Boller, 1995). This corresponds to the period during which infants are improving their categorisation abilities (Quinn & Eimas, 1996). As a result of improved categorisation abilities one would expect the load on memory to decrease and, hence, the developmental profiles of these two skills would appear coupled. The model in this paper constitutes an explicit proposal of how memory and categorization are linked in this domain.

The rest of this paper unfolds as follows. We begin with a brief discussion of how infant preferential looking behaviors can be mapped onto the performance of a connectionist network. This argument is based on Solokov's (1963) classic model of habituation of the orienting reflex in which an internal representation of an external stimulus is constructed and embellished each time the organism encounters the same stimulus in the

[1] Both authors have contributed equally to the development of this paper.

same context. Next, we discuss the properties of the input stimuli. This is a critical step since both infants (Younger, 1985) and connectionist networks (e.g., Rumelhart & McClelland, 1986) are known to categorize based on the correlational structure of the input features. Then, the model's performance is presented with respect to the asymmetric categorisation and memory interference effects. We also describe a novel prediction of the model: an asymmetric interference effect. Finally, these results are discussed with respect to both infant cognition and connectionist modeling.

The Model

Infant categorisation tasks rely on preferential looking techniques based on the finding that infants direct attention more to unfamiliar or unexpected stimuli. The standard interpretation of this behavior is that the infants are comparing the input stimuli to an internal representation of that stimulus (e.g., Solokov, 1963; Charlseworth, 1969; Cohen, 1973). As long as there is a discrepancy between the information stored in the internal representation and the visual input the infant continues to attend to the stimulus. While attending to the stimulus the infant updates its internal representation. When the information in the internal representation is no longer discrepant with the visual input, attention is switched elsewhere. When a familiar object is presented there is little or no attending because the infant already has a reliable internal representation of that object. In contrast, when an unfamiliar or unexpected object is presented, there is a lot of attending because an internal representation has to be constructed or adjusted. The degree that the novel object differs from existing internal representations determines the amount of adjusting that has to be done, and hence the duration of attention.[2]

We used an autoencoder to model this process. Learning in an autoencoder consists of developing an internal representation of the input (at the hidden unit level) that is sufficiently reliable to reproduce all the information in the original input (Cottrell, Munro, & Zipser, 1988). Information is first compressed into an internal representation and then expanded to reproduce the original input. The successive cycles of training in the autoencoder are an iterative process by which a reliable internal representation of the input is developed. The reliability of the representation is tested by expanding it and comparing the resulting predictions to the actual stimulus being encoded.

We suggest that during the period of captured attention infants are actively involved in an iterative process of encoding the visual input into an internal representation and then assessing that representation against the continuing perceptual input. This is accomplished by using the internal representation to predict what the properties of the stimulus are. As long as the representation fails to predict the stimulus properties, the infant continues to fixate the stimulus and to update the internal representations. Similar interpretations have been suggested elsewhere (Mareschal, Plunkett, & Harris, 1995; Munakata, McClelland, Johnson, & Siegler, 1994).

There are several implications to this modeling approach. Looking time is monotonically related to the network error. The greater the error, the longer the looking time. Stimuli presented for a very short time will be encoded less well than those presented for a longer period. However, prolonged exposure after error (attention) has fallen off will not improve memory of the stimulus. The degree to which error (looking time) increases on presentation of a novel object depends on the similarity between the novel object and the familiar object. Presenting a series of similar objects leads to a progressive error drop on future similar objects. A prototype of the set of objects leads to lower error than individual objects. All of this is true of both autoassociators (where output error is the measurable quantity) and infants (where looking time is the measurable quantity).

The results reported below are based on the performance of a standard 10-8-10 feedforward backpropagation network. The learning rate was set to 0.9 and momentum to 0.9. A Fahlman offset of 0.1 was also used. Networks were trained for a maximum of 250 epochs or until all output bits were within 0.2 of their targets. This was meant to reflect the fact that in the Quinn and Eimas studies infants were shown pictures for a fixed duration of time rather than using a proportional looking time criterion. Results are averaged over 50 replications.

Twelve items from one category were presented sequentially to the network in groups of two (i.e., weights were updated in batches of two). This was meant to capture the fact that pairs of pictures were presented to the infants during the habituation phase. After exposure to the twelve patterns, the networks were tested on an item of the same category and an item of the other (unseen) category.

The Data

The data were obtained from measurements of the original Cat and Dog pictures used by Eimas and Quinn. There were 18 dogs and 18 cats classified according to the following ten traits: head length, head width, eye separation, ear separation, ear length, nose length, nose width, leg length vertical extent, and horizontal extent. Although it is difficult to say for certain which features the infants are using during categorisation, it is well known that infants segregate items into categories on the basis of clusters of correlated attributes of different values (Younger, 1985; see Quinn and Johnson, 1996 for a detailed justification these input features). The feature values were normalized to be within 0 and 1.

Each feature is assumed to be normally distributed. Figure 1 shows the probability distributions of the 10 traits for both cats and dogs. Some of the traits are very similar in terms of their means and distribution of both cats and dogs (e.g. head length and head width). Others, especially

[2] This process can be interrupted at any point by the intervention of a more salient event. See Hood (1995) for thorough review of what determines infant selective attention.

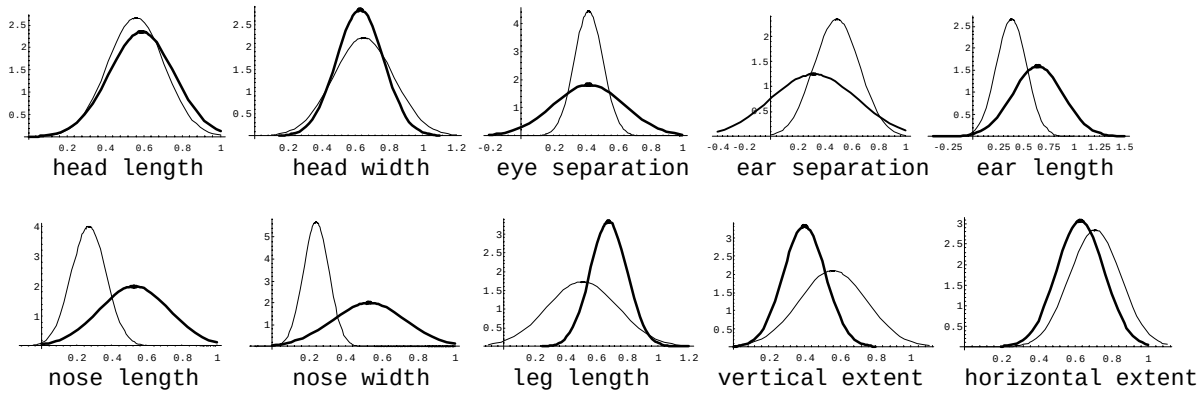

Figure 1. The frequency distributions of values of the ten defining traits for both populations of 18 dogs and 18 cats. The dark line indicates the distribution for the "dog" category.

nose length and nose width, are very different and will provide the crucial explanation of the unexpected attentional asymmetries reported by Quinn, Eimas, & Rosenkrantz (1993) and Eimas, Quinn, & Cown (1994).

Consider a single trait, for example, "nose width." Figure 2 shows the probability distributions for this trait for both dogs and cats. The (normalized) mean nose width for the dog population is 0.53 with a standard deviation (σ) of 0.2, whereas the mean for the cat population is 0.24 with a much smaller standard deviation of 0.07. Consequently, the nose width of virtually all cats in the population will fall within two standard deviations of the nose-width mean for dogs. On the other hand, the nose width of the majority of dogs *does not* fall within 2σ of the nose-width mean for cats. The result, in short, is that at least for this trait, all cats could be exemplars of dogs, whereas most dogs *could not* be cats.

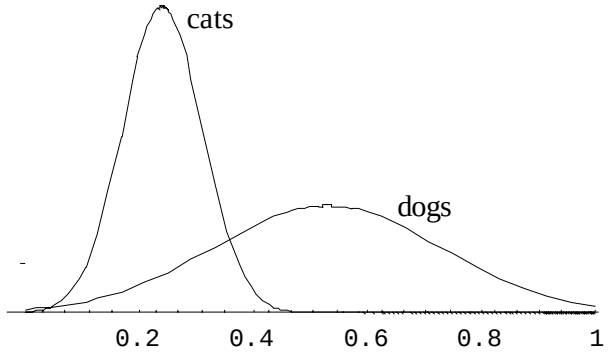

Figure 2. The distributions of the "nose-width" trait for the population of cats and dogs shown to infants by Quinn, Eimas & Rosenkrantz, 1993.

Referring again to Figure 1, it is clear that in almost all cases the distribution for each Dog trait (represented by the dark line) subsumes the distribution for corresponding trait for cats. The narrower distributions for most Cat traits, on the other hand, do not subsume the range of values for the corresponding Dog traits. In other words, cats are possible dogs but the reverse is not the case: most dogs are not possible cats.

Specifically, when we examine all of the members of the two populations, we see that the values of all ten traits for 9 (i.e., 50%) of the members of the Cat category fall within a 2σ cut-off for those traits for the Dog category. Fully half of the cats in the population could be reasonably classified as dogs. In contrast, the smaller means and lower variances of a number of traits (especially, nose length and nose width) for cats compared to dogs means that only 2 of the 18 dogs could conceivably be classified as being members of the Cat category.

Results

The Development of Cat and Dog Categories

Like infants, these networks form a category of both Cat and Dog. Figure 3 shows the initial error score, the error score after twelve presentations of either cats or dogs, and the average error score (after training) for the 6 remaining exemplars in either the Cat or Dog category.

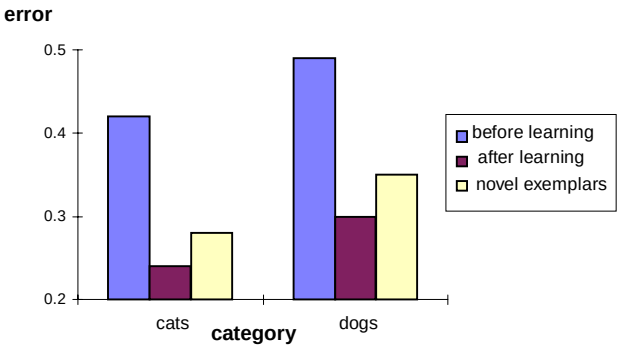

Figure 3. Mean network error when (a) presented with exemplars before learning, (b) presented with exemplars after learning, and (c) presented with novel exemplars after learning.

After learning, error is lower suggesting that the network has developed a reliable internal representation of cats or dogs. The generalization error rises slightly, showing that

the networks recognize these exemplars as novel. Infants are also able to distinguish individual exemplars within the category (Quinn, Eimas, & Rosenkrantz, 1993). However, the generalization error remains below the initial error suggesting that the new exemplars are indeed assimilated to the category formed by the networks.

The Exclusivity of the Cat and Dog Categories

Eimas and Quinn found that there was an asymmetry in the exclusivity of the Cat and Dog categories developed by infants. To summarize the discussion at the beginning of this paper, when infants are shown a series of photographs of cats, the subsequent presentation of a dog produces a large increase in attention (compared to the presentation of yet another cat). The opposite is not true. In other words, when an infant is shown a series of photographs of dogs, the subsequent presentation of a cat is essentially of no greater interest than the presentation of another dog. The modeling assumption that we have made is that network error and infant attention levels correlate: the higher the network error, the longer the looking time of the infant (Mareschal, Plunkett, & Harris, 1995; Munakata, McClelland, Johnson, & Siegler, 1994).

Figure 4 shows what happens when networks trained on cats are presented with a novel cat and a dog, and when networks trained on dogs are tested with a novel dog and a cat. When the networks are initially trained on cats, the presentation of a dog results in a large error score (corresponding to the results observed with infants in terms of a longer looking time). Dogs are not included within the categorical representation of cats. In contrast, when the networks are initially trained on dogs, the presentation of a cat result only in small increase in error suggesting that the cats have been included in the dog category.

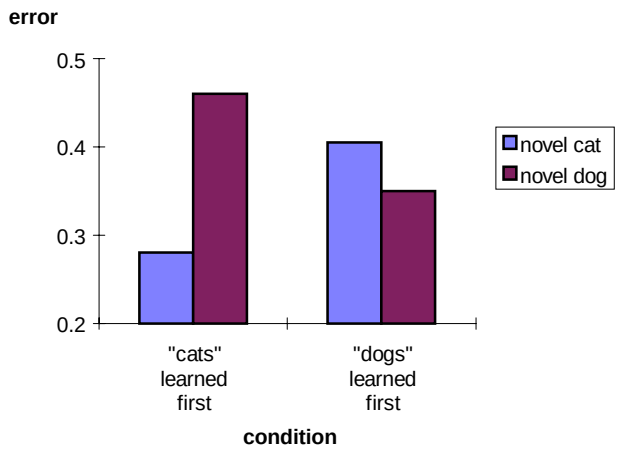

Figure 4. Asymmetric exclusivity of the Cat and Dog categories. When trained first on cats, a novel dog results in a large increase in error (as compared to a novel cat) but when trained first on dogs, a novel cat only produces a small increase in error (as compared to a novel dog).

The Asymmetric Interference Effect

This section examines the effect of learning a second category during the retention interval. The network was sequentially trained on twelve exemplars (6 pairs) of either cats or dogs. It was then tested a first time (T1) with a novel exemplar of the same category. (i.e., when trained with cats it was tested with a novel cat; when trained on dogs the network was tested with a novel dog). Following this, the network was trained on 4 exemplars (2 pairs) of the complementary category. If the network had initially been trained on cats it was presented with four dogs. If it had originally learned dogs, the network was presented with 4 cats. Finally, the network was tested a second time (T2) with the same novel exemplar as in the first test session. The difference in the network's performance in T2 as compared to T1 is a measure of the amount of interference (or forgetting) that has occurred as a consequence of learning the intervening exemplars.

Figure 5 shows the difference between the network's performance at T2 and T1, when (a) the original category is Cat and the intervening category is Dog, and (b) when the original category is Dog and the intervening category Cat. Learning dogs during the intervening period has a large detrimental effect on the prior learning of cats. In stark contrast to this, learning cats during the intervening period has little or no detrimental effect on the prior learning of dogs. This finding echoes the category exclusivity dissociation of the previous section and reflects the distribution of means and variances of the input attributes. Although these experiments have yet to be done on children, the model makes the clear prediction that in infants learning dogs after having first learned cats will cause far more forgetting of the originally learned cats than vice-versa.

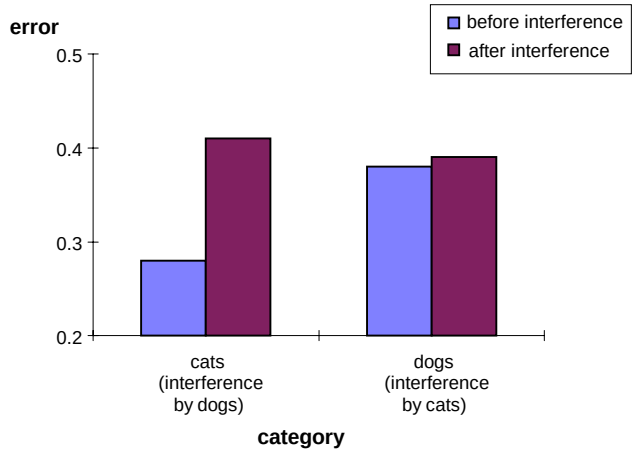

Figure 5. Network performance with novel exemplars before and after learning an interference category.

The Effect of Learning Closely Related Animals

One possible explanation for the asymmetry effect is that the Cat category is particularly susceptible to interference in the presence of *any* new category. To test this we first had the network sequentially learn 6 pairs of cats and then trained the network on 4 examples (2 pairs) of lions. Although lions are more similar to cats than are

dogs, they nonetheless form a distinguishable category and, as such, could interfere with the prior learning of the cat category, assuming this latter category did indeed turn out to be exceptionally susceptible to interference.

This, however, did not turn out to be the case. As Figure 6 shows, there is only a very slight increase in the error for novel cats after the lions have been learned. This is due to the fact that, in the lion and cat data used for the experiment, all of the lions fall within the Cat category (i.e., the value of each trait is within 2σ of the mean value of the same trait for the "lion" category) and vice-versa. This permits the prediction of very little interference in the model. As can be seen in Figure 6, this is indeed the case.

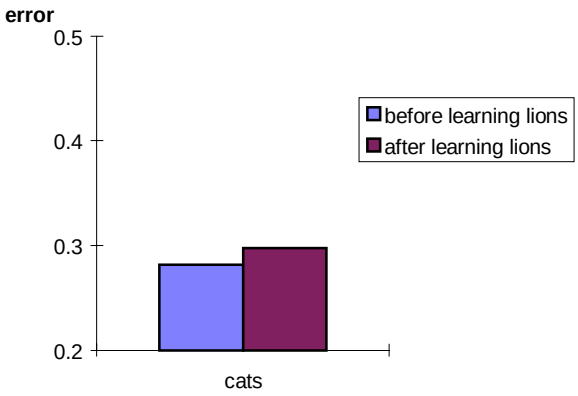

Figure 6. After learning an initial set of cats, being exposed to a set of lions interferes only very slightly with the ability of the network to recognize novel cats.

Discussion

In this paper we have presented a simple connectionist model of early infant memory and categorisation. An autoencoder was presented with data measured directly from photographs of cats and dogs that had been presented to infants. Like the infants, the networks categorized the pictures appropriately into Cat or Dog categories. The categories showed the same asymmetric exclusivity effect found in infants. When trained on dogs, the category formed included cats, whereas when trained on cats, the category formed excluded dogs. Learning sequential categories showed asymmetric interference. Learning cats after having learned dogs did not interfere with the prior knowledge of dogs. In contrast, learning dogs after having learned cats resulted in catastrophic interference with the knowledge of cats.

The asymmetry in both categorisation and interference in the networks was found to arise from the distribution of features describing the stimuli. Although the means of the traits for the cats and the dogs were different, the variance of the dog features was much larger. Most, Cat features fell within 2 standard deviations of the Dog features. In this sense, the cats are subsumed by the Dog category. Learning about cats does not disrupt the knowledge of dogs. However, learning about dogs exposes the network to feature values that are outside the range experienced by the network as it learned about cats. These new feature values lead to a change in internal representation in order to accommodate the new information. Hence, learning about dogs disrupts the representation of cats.

The model provides a more precise mechanistic account of the categorisation asymmetry than that suggested by Quinn, Eimas, and Rosenkrantz (1993). They suggested that the failure to learn a Dog category that excluded cats was due to the greater variability of the Dog category. We suggest that this asymmetry arises from the asymmetrical overlap of the trait distributions and not just the variance of the distribution itself. It is not just the fact the there is greater variability in the Dog category, but also that the values of the Cat features are subsumed by those of the Dog features whereas the reverse is not true.

The work reported in this paper goes beyond simply capturing a quirk of infant performance. It suggests a link between performance in infant categorisation and memory tasks. The same asymmetry observed in categorisation also appears in interference tasks. This is a strong prediction of novel infant behavior. Note that some indirect evidence already exists. Cats have been found to interfere with lions, but lions do not interfere with cats in categorization tasks (Quinn, Eimas, & Cowan, 1994).

No mechanistic accounts of the interference effects in infant visual memory have been proposed (Rovee-Collier & Bollet, 1995). The only suggestion is that interference will occur when the intervening stimulus is similar to the original material and is encoded by the infant. The modeling work we present suggests that the interference effects arise from the mechanisms involved in categorising multiple stimuli in an associative system with distributed representations. This proposal is corroborated by the fact that techniques such as interleaving reminder examples during the second (interference) learning phase reduces interference both in connectionist networks (Robins, 1995) and in infants (Quinn, Eimas, & Cowan, 1994; Rovee-Collier & Bollet, 1995).

Future modeling work needs to account for the development of an immature system that is susceptible to interference (such as infant memory) to one that is not susceptible to interference (such as adult memory). We are currently pursuing this avenue.

Finally, this model is an attempt to synthesize a range of idiosyncratic infant behaviors under one mechanism. It is a simple model but the principles accounting for the observed phenomena are true of most distributed associative systems. A more sophisticated model would produce the same interference and asymmetry results.

In terms of connectionist modeling, this work suggests that catastrophic interference is an integral part of early cognition. Rather than brandishing it as a failure of connectionist systems, it should be viewed as a necessary feature of any system wishing to model human memory and categorisation across the whole range of ages.

In summary, we present a simple connectionist model of memory and categorisation in early infancy. The model underscores the continuity that underlies the development

of memory and categorisation by using mechanisms known to model adult performance. The model makes a strong prediction concerning asymmetric interference effects in infancy, and suggests that catastrophic interference is a necessary part of any model that intends to capture the whole range of human memory and categorisation abilities

Acknowledgements

This work was funded by a collaborative research grant awarded to both authors by the British Council and the Belgian CGRI as well as by Belgian FNRS Grant No. D.4516.93 and PAI Grant No. P4/19. We would like to thank Paul Quinn for providing copies of the original stimuli used to test infants and for providing helpful comments on an earlier draft of this paper.

References

Charlesworth, W. R. (1969). The role of surprise in cognitive development. In D. Elkind & J. Flavell (Eds.), *Studies in cognitive development. Essays in honor of Jean Paiget*, pp. 257-314, Oxford, UK: Oxford University Press.

Cohen, L. B. (1973). A two-process model of infant visual attention. *Merrill-Palmer Quarterly, 19*, 157-180.

Cohen, L. & Gelber, E. R. (1975). Infant visual memory. In L. Cohen & Salapatek (Eds.), *Infant perception: From sensation to cognition*, Vol.1 (pp. 347-403). New York: Academic Press.

Cottrell, G. W., Munro, P., & Zipser, D. (1988). Image compression by backpropagation: and example of extensional programming. In N. E. Sharkey (Ed.), *Advances in cognitive science, Vol. 3*. Norwood, NJ: Ablex.

Deloache, J. S. (1976). Rate of habituation and visual memory in infants, *Child Development, 47*, 145-154.

Eimas, P. D., Quinn, P. C., & Cowan, P. (1994). Development of exclusivity in perceptually based categories of young infants, *Journal of Experimental Child Psychology, 58*, 418-431.

Hood, B. M. (1995). Shifts of visual attention in the human infant: A neuroscientific approach. In C. Rovee-Collier & L. P. Lipsitt, *Advances in infancy research, Vol. 9* (pp. 163-216). Norwood, NJ: Ablex.

Knapp, A. G. & Anderson, J. A. (1984). Theory of categorisation based on distributed memory storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 616-637.

Mareschal, D., Plunkett, K., & Harris, P. (1995). Developing object permanence: A connectionist model. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the 17th annual conference of the Cognitive Science Society* (pp. 170-175). Mahwah, NJ: LEA

McCall, R. B., Kennedy, C. B., & Dodds, C. (1977). The interfering effect of distracting stimuli on infant's memory. Child Development, 48-79-87.

Munakata, Y., McClelland, J. L., Johnson, M. N., & Seigler, R. S. (1994). Now you see it now you don't: A gradualistic framework for understanding infant success and failures in object permanence tasks. Technical Report, PDP.CNS.94.2, Carnegie Mellon University, Pittsburgh, USA.

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants, *Perception, 22*, 463-475.

Quinn, P. C., & Eimas, P. D. (1996). Perceptual organization and categorization in young infants. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research* (Vol. 10, pp. 1-36). Norwood, NJ: Ablex.

Quinn, P. C., & Johnson, M. H. (1996). The emergence of perceptual category representations during early development: A connectionist analysis. In G. W. Cottrell (Ed.), *Proceedings of the 18th annual conference of the Cognitive Science Society*(pp. 638-643). Hillsdale, NJ: LEA.

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal, *Connection Science, 7*, 123-146.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233.

Rovee-Collier, C. & Boller, K. (1995). Interference or facilitation in infant memory? In F. N. Dempster & C. J. Brainerd (Eds.), *Interference and inhibition in cognition* (pp. 61-104). San Diego, CA: Academic Press.

Rumelhart, D. & McClelland, J. (1986). *Parallel Distributed Processing*. Cambridge, MA: The MIT Press.

Slater, A. (1995). Visual perception and memory at birth. In C. Rovee-Collier & L. P. Lipsitt, *Advances in infancy research, Vol. 9* (pp. 107-125). Norwood, NJ: Ablex.

Shanks, D. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 433-443.

Solokov, E. N. (1963). *Perception and the conditioned reflex.* Hillsdale, NJ: LEA.

Younger, B, A. (1985). The segregation of items into categories by ten-month-old infants, *Child Development, 56*, 1574-1583.