Inductive Reasoning Revisited: Children's reliance on category labels and appearances

Jonathan J. Loose (J.J.Loose@exeter.ac.uk)
School of Psychology
University of Exeter,
Perry Rd, Exeter, EX4 4QG, UK

Denis Mareschal (D.Mareschal@bbk.ac.uk)
Center for Brain and Cognitive Development,
Department of Psychology,
Birkbeck College, University of London,
Malet St., London, WC1E 7HX, UK.

Abstract

Previous studies of children's inductive reasoning have attempted to demonstrate that label information is preferred to perceptual similarity as the basis for inductive inference (Gelman and Markman, 1986; Gelman and Markman, 1987; Gelman, 1988). A connectionist model of the development of inductive reasoning predicts that this will only be true when the perceptual variability of category exemplars is high (Loose and Mareschal, 1997). We report three studies investigating the model's predictions. Study 1 demonstrates that patterns of categorization can depend on perceptual variability. In study 2 we develop a set of stimuli with differing variability but equal discriminability. Study 3 demonstrates that young children's patterns of reasoning are more affected by the presence of category labels when the inference is from an exemplar of a more perceptually variable category. This study also demonstrates that the basis of inference is not explicable in terms of the ease of the ability to categorize of the stimuli. Implications for the original model are discussed.

General Introduction

Studies of the basis of children's inductive reasoning have been used to support the notion that even young children's representations are abstract and conceptually sophisticated, as opposed to the historic view that they are perceptually grounded and limited (Inhelder and Piaget, 1964; Gelman and Markman, 1986). A connectionist model of the inductive reasoning paradigm used by Gelman and Markman demonstrates that their results are replicable by a system which does not utilise complex taxonomic representations (Loose and Mareschal, 1997). The model relies on the fact that there is greater variability inherent in perceptual information than in label information—which by its very nature serves to uniquely identify classes of objects. On this basis, learning to make useful inferences in the environment naturally leads to a reliance on category labels. However, this will only be the case if reliable information about the category cannot be extracted simply from the perceptual world. Labels may not be required when category instances are more homogeneous in appearance. Thus, the connectionist model makes a prediction regarding the decision to treat similarity of appearance or shared labels as the basis of inference. The prediction is that the basis of inference is mediated by the variability in appearance of the exemplars which were used in the formation of the category.

An empirical study of the effect of category variability on adult inferences within a taxonomic domain suggests that category variability is important for adults, and also suggests a parallel between the effect of development on conceptual systems and the effect of greater learning within a particular domain (Loose and Mareschal, 1998). This study, and others (e.g. Hampton, 1995) shed some doubt on the assumed target of the developmental process being a global preference for label information (at least within the domain of natural kinds) The question remains as to whether such effects will also be demonstrated by pre-school children. In the studies described here, we investigate the effect of category variability on inferences made by pre-schoolers—the population sampled for the original studies. The questions of interest are (a) whether the basis of young children's inferences is affected by variability, and (b) whether or not performance on inference tasks is best viewed as a simple product of categorization processes.

The rest of the paper unfolds as follows. First, two studies of categorization are described. These studies serve to validate the stimuli used in the final inference study. Different kinds of responses, to categories of different variability, lead to a possible explanation of the effect of category variability in inference. An inference study that investigates the relationship between category variability and the accuracy of children's inductive inference (holding category discriminability constant) is then reported. The results of this study are discussed with respect to the conclusions of the categorization studies. Finally, implications for the development of the connectionist model are briefly discussed.

Experiment 1: Categorization of objects drawn from populations with different perceptual variabilities.

This first study was designed to examine whether or not children could explicitly categorize the algorithmically generated stimuli to be used in the final inference study. Two stimulus categories were used, and are described below. Given that the stimuli were generated algorithmicly, we know what the differences in variability between the different categories used. It is important to know whether participants can explicitly detect the differences between exemplars of different categories. It is also important to know that participants adequately represent the range of the various dimensions along which objects of a particular category can vary.

Because this work deals with category-based inferences,

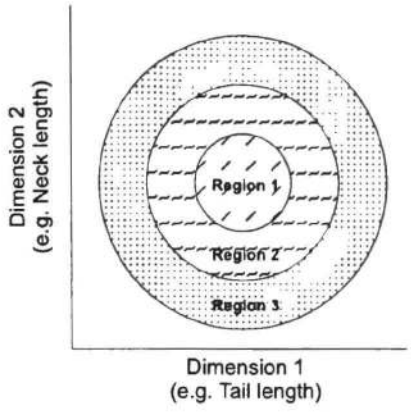

Figure 1: *Regions of feature space from which different 'category exemplars' were drawn.*

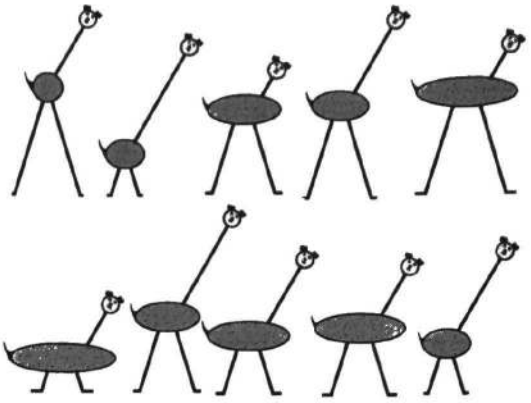

Figure 2: *Training stimuli used to familiarize children with the high variability category.*

it is crucial that the participants can categorize the stimuli. This point is made all the stronger by the fact that most other studies of this type have utilized stimuli which are previously known to the child and which can be clearly categorized (e.g. Gelman & Coley, 1990).

Method

Design & Materials The study has a within-subjects, single factor design. The dependent variable is category discriminability—a measure of the likelihood that stimuli will be accurately judged as instances of a particular category, or as being outside of that category. The only factor in this study is category variability. Variability has two levels, high and low, corresponding to categories formed from exemplars constructed across a wider or narrower range of a fixed set of dimensions. Subjects were 10 pupils in a primary reception class[1], six females and four males. Mean age was 4 years 11 months (4;11).

The materials for the study consisted of six color picture sets of artificial animals—two training sets plus four test sets. The training sets consisted of a low variability group and a high variability group. The test sets consisted of three levels of variability in order to provide pictures of animals inside/outside the high/low variability categories. The variability of pictures outside the low variability category is the same as that for pictures inside the high variability category—thus there are three variability levels, and four sets of stimuli.

The stimuli were constructed systematically such that differences between 'animals' consisted of controlled changes along a set of known dimensions. The overall look of the stimuli is similar to those used by Younger (1990) in her studies of categorization in infants and kindergarten children. Examples of the kind of stimuli used are given in figure 2. The reasons for having stimuli that looked like this were firstly so that they would be easily controlled, manipulated and generated, and secondly that they could be proposed as natural kinds from, "another world," which would promote the idea that children should reason about them as if they were biological kinds—albeit of a fictional nature.

In this study, the term 'category variability' refers to the spread of values across the different dimensions used to derive each category exemplar. The term 'category' is thus used in a non-standard way, since we used a nested category structure—that is, (referring to figure 1), region 1 is one category, while region 2 is a second category. Region 3 is necessary to provide exemplars outside the region 2 category. Contrasting categories with the same central tendency is consistent with a view of category representation which explicitly incorporates information from individual exemplars as well as central tendency (Smith and Medin, 1981).

The actual dimensions that were used for generation of the stimuli were body length, neck length and tail size (for the low variability category only) plus leg length and body lightness (high variability category). The high variability stimuli were also drawn across a broader range of each of these dimensions[2].

Procedure The study utilized familiarization and categorization phases. The familiarization process was motivated by our previous research. It minimizes the time taken with each child to avoid the risk of episodic learning.

Both categories were presented to each subject in random order, with the following procedure:

After spending a short time putting the child at ease, the experimenter presented one of the target sets, chosen at random. The child's familiarization was a guided process. First, the child was asked to point to each animal on the sheet, and then to 'count around' the animals. Finally, the child was asked to find exemplars with particular features—for example, "the animal with the longest neck". These questions focussed attention on the salient dimensions of the stimuli. Having done this, the child was told, "These animals are actually called 'wugs', and on the planet which they come from, there are lots of wugs. *But!...* on the planet there are also some other

[1]The ages of UK primary reception class children are equivalent to US pre-schoolers.

[2]The stimuli were all placed on an (identical) brightly colored "sky" background to make the pictures a little more interesting.

	$\bar{x}$	s
Var:High, Test:Exemplars	3.20	1.03
Var:High, Test:Non-exemplars	1.60	1.07
Var:Low, Test:Exemplars	3.50	0.71
Var:Low, Test:Non-Exemplars	2.60	1.58

Table 1: *Number of correct identifications of objects as either "wugs" or not ($N = 10$ in each case).*

animals which look a bit like wugs—but they're not! They're something else! Those other animals that we saw were like that."

The child was then told that they were going to play a game—they must look at new pictures of animals, and decide which animals actually were 'wugs' and which were impostors. It was expected that by emphasizing that there would be animals in the test sets which would look like wugs but would not actually *be* wugs, participants would be more discriminating in their judgements. Without such instruction, there could be a tendency towards over-generalization, since the child had only seen positive exemplars of 'wugs' during the familiarization process.

The child was then presented with the appropriate test set of eight individual pictures for the target category being familiarized. The pictures were presented to the child in random order. Half of the pictures were of category exemplars, and half were taken from a more variable category with similar prototype. Each time a picture was presented, the child was encouraged to take a quick look at the sheet of known exemplars before deciding on the identity of the new animal. Children were not given time to explicitly compare each test picture with each known exemplar—the purpose of taking a quick look was to reinforce the child's original impression of the kinds of things that could count as category exemplars. This was a continuation of the familiarization process, saving time and allowing the study to be successfully performed with children of the target age.

The child's response was recorded as either correct or incorrect. Once this procedure had been completed for all eight members of the test set, attention switched to the other category which was not originally chosen. The procedure was repeated with the second target category. This new category of animals was given a new name. Exemplars were called 'keeches' rather than 'wugs'. Note that these names are nonwords, and have been used previously in similar studies, e.g. Florian (1994). Order effects were removed by presenting each of the target categories first in 50% of cases.

Results

Response accuracy is shown in Table 1. It is easier to detect non-exemplars from a low variance category than from a high variance category. It is possible that children might be biased to give a 'yes' or 'no' response irrespective of the question. This is accounted for by a recoding of the response data. This recoding takes advantage of a measure derived from signal detection theory (McNicol, 1972 cited in Monk & Eiser, 1980). $P(A)$—an approximation of the area under the ROC curve[3] is a bias free measure of discriminability. Given that there are only two response categories in this study, the measure is easily understood informally. We assume that I_i is the proportion of category exemplars judged as such (hits), O_i is the proportion of non-category exemplars judged as exemplars (false alarms), I_o is the proportion of category exemplars judged as non-exemplars (misses) and O_o is the proportion of non-category exemplars judged as such (all clears)[4]. These assumptions allow us to state $P(A)$ as in equation 1.

$$P(A) = I_i O_o + \frac{I_i O_i + I_o O_o}{2} \quad (1)$$

This equation has two components which correspond to different aspects of table 2. Correct judgements are found along the I_i and O_o diagonal. If all responses are in these cells, then the first component of $P(A)$ evaluates to 1. The second component reflects response bias. In this case, it will evaluate to 0—however, to the extent that subjects' judgements are the same irrespective of the correct answer, responses will be distributed across a single table row ($I_i O_i / I_o O_o$), leading to an increase in $\frac{I_i O_i + I_o O_o}{2}$, and a corresponding decrease in $I_i O_o$. Given the denominator of $\frac{I_i O_i + I_o O_o}{2}$, it can be seen that response bias will tend to reduce the overall result.

Thus, computing $P(A)$ provides a measure of discriminability ranging from 0—1 which is not subject to response bias.

		Correct Judgment	
		IN	OUT
Subject	IN	I_i	O_i
Judgment	OUT	I_o	O_o

Table 2: *Response bias and correct inference.*

The boundary of the low variability category was discriminated more clearly than the boundary of the high variability category (Mean discriminabilities 0.68/0.33). This difference is reliable ($N = 10$, $t = -5.314$, $p < 0.001$). Thus, despite equivalent objective differences between exemplars/non-exemplars of each category, the boundary of the less variable category is still more discriminable.

Discussion

The results of study 1 suggest that the more perceptually diverse a category is, the more an object must be perceptually different from category members before it is recognized as not being a category member. This is an interesting finding,

[3]The 'Receiver Operating Characteristic' curve plots the probability of false-positive identifications against true-positive identifications under different conditions of noise and signal strength.

[4]Thus, for example, for this study I_i is the number of 'wug' responses given to genuine 'wug' stimuli, divided by the total number of genuine 'wug' stimuli presented (always 4 in this study).

suggesting something like a "Weber's law for categorization". The general idea of such a law would be that the required (perceptual) distance between an exemplar of a category and an object which is not an exemplar so that the category distinction is clearly noticed becomes larger with the perceptual variability of the category. It is helpful to define a distinct new here. If we define a 'just noticeable category discrimination', or JNCD, as the distance in feature space between the edge of a category and the nearest point which is reliably judged to be outside of the category, then we can make the simple claim that the JNCD will grow with the perceptual variability of the category.

It may be that the differences in the JNCDs of categories with different variabilities explains why higher variability categories promote more label based inductive reasoning. Therefore, it would be interesting to modify our stimuli to take account of the JNCD, and see if we are then able to find an effect of category variability on inductive reasoning performance.

Experiment 2: Categorization of stimuli with equivalent JNCD

The second categorization study was almost identical to the first, involving ten more pupils from the same primary school class (mean age 5;0). The difference between this study and study 1 is that a new set of stimuli were generated accounting for the effect of differing JNCD with increased variability. Stimulus modifications are described below.

Revised Materials

In order to be able to increase exemplar differences without making some stimuli extremely large, an appropriate extra dimension of variability was added—that of texture. It has been demonstrated that young children's classifications of natural kinds are extremely sensitive to texture. Thus the texture dimension should have a disproportionately large effect on children's judgements (Jones et al., 1991). A texture scale was taken from a computer painting program, and applied to the pictures in the same way that the other dimensions had been. This would serve to place extra 'out of category' information in the non-exemplar pictures.

Further to this, the pictures of animals intended as instances drawn from *outside* the *high* variability category were also modified such that they included more feature information which was further outside the boundaries of anything used in the high variability category. This explicit modification of only the high variability category is required to remove the effect of differing JNCDs.

Results

Discriminability ratings were computed for each subject as in the previous study. Mean discriminability ratings in this study were 0.82/0.73 for the low/high variability categories respectively. The difference in means in this second study is not reliable ($N = 10, t = -1.231, p = 0.25$). Thus, there is no evidence of differing discriminability between the more and less homogeneous categories. This result also demonstrates that the modified stimuli account for the effect of differing JNCDs.

Discussion

The difference in discriminability between the two categories has been removed by the modifications to the stimuli. This is confirmed by a comparison between the results from the two studies.

A two way mixed analysis of variance was performed, comparing the discriminability ratings between the previous studies. The within subjects factor was category variability (two levels, low and high), the between subjects factor was the study from which the results came. The discriminability of exemplars of both categories has improved from the first to the second study. This would be expected from the fact that we have added another dimension to all stimuli (texture). There is also a greater improvement in the high variability category. This would be expected from the extra modifications to stimuli drawn from the high variability category. The main effect of stimulus group is significant ($N = 20, F(1, 18) = 9.311, p = 0.007$), as is the main effect of category variability ($N = 20, F(1, 18) = 17.778, p = 0.001$). Importantly, the interaction is also significant ($N = 20, F(1, 18) = 5.133, p = 0.036$). Thus the effect of category variability is moderated by the stimulus set used. In this case, there is only a reliable difference between the two categories when the original stimuli were used.

These preliminary categorization studies have achieved two things. First, if our model's prediction is correct—that an important factor in the choice to make inferences on the basis of appearance or label information is the perceptual variability of the set of exemplars which form the category—then our first study suggests one possible explanation. It may be that some kind of psychophysical law applies such that the greater the perceptual variability of a category, the larger the "just noticeable category difference." This would lead to a potential re-interpretation of at least some of the comparisons made in previous studies on the basis that the notion of "perceptual similarity" is not an absolute measure, and therefore should not be compared across stimuli without regard for JNCD. Secondly, we now have a set of stimuli which allow us to conduct a study of inductive inference. These stimuli have taken into account the effect of changing JNCD, since categories are not significantly different in their discriminability despite having different levels of perceptual variability.

Experiment 3: Category-based inductive inferences

This study uses the previous stimuli to investigate whether the accuracy of inferences from exemplars of more/less perceptually variable categories will be affected in the same way by the addition of label information. The categories used are equally discriminable, despite having distinct perceptual variabilities.

Method

Subjects & Materials Twenty-eight children participated in the study, taken from three primary school reception classes in Exeter, UK. 14 males and 14 females participated. The mean age of all participants was 4 years 10 months. Participants were chosen at random from school classes. Care was taken to exclude children known to have learning difficulties.

Stimuli were as described above. In the high variability condition, "wugs"/"keeches" were drawn from regions 2/3 of figure 1. In the low variability condition, "wugs"/"keeches" were drawn from regions 1/2. Importantly, the modified stimuli from Experiment 2 were used here so that both low/high variability categories were equally discriminable. Examples of the non-perceptual properties used in the inference questions are, "very quiet and shy of people", and "live only where it is very cold."

Design & Procedure The study utilizes a 2x2 between subjects design. The first factor was the variability of the set of exemplars used to give an impression of the category of animals (two levels, high/low). The second factor was the presence or absence of stimulus labels (two levels, present/absent). The presence of labels was indicated by giving test exemplars different names (wug/keech) depending on whether they were category exemplars or not. When labels were absent, all pictures were described as "animals".

Each participant went through two phases in the study, familiarization and inference. After a short time spent putting the child at ease, s/he began the familiarization phase, and was shown a target category in the form of a single sheet of paper with a set of exemplars printed on it. All exemplars were presented along a single line, and without any white space around them. The variability of the exemplars was varied across conditions. All children were told that these were animals from "another planet" and that they were called "wugs".

The familiarization process was the same across all conditions. It consisted of ensuring that the child looked at each exemplar individually, and also attended to each of the salient dimensions. It also included the use of the terms "wug" and "keech" to describe the stimuli.

Having spent some time examining the target category and learning its name, the children entered the second phase—inference. During this phase, children in the *label* condition were told when they were looking at a wug, and when they were looking at a keech. Children in the *no-label* condition were always told/asked about this or that *animal*. In the rest of this section, the things that the children were told in each condition are contained within a single description, with the differing wording represented as: [label condition/no-label condition].

First, each participant was told that the 'planet' which the [wugs/animals] live on contains some animals that look like [wugs/these animals], but which are not. This was to give the children the idea that what they were about to see *might* be a *different kind* of alien animal. Since the wugs look distinct from anything the children have seen before, it is quite possible that without being told this, they would assume that everything they were shown was a wug.

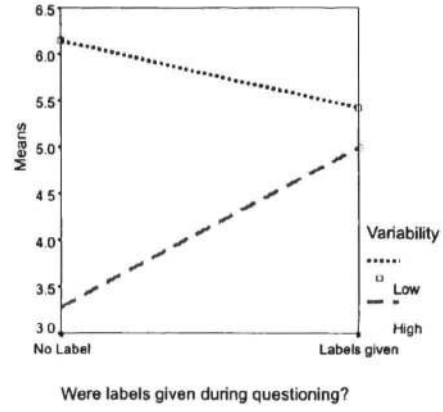

Figure 3: *Number of correct responses (/8) given by children in the different experimental conditions of the inference study.*

Next, the children were told that all [wug/animal] pictures on the target sheet had a particular non-perceptual property. For example, "All these [wugs/animals] can see in the dark." The child was then shown a further picture—one that had not been seen before, and was asked, "Do you think that this [(wug/keech)/animal] can [see in the dark]?". We make the assumption that an inference is valid *only* if the source and target objects are drawn from the same category. When children generalized a property *across* categories, or did *not* generalize a property *within* categories, then the response was considered *incorrect*. Responses were recorded as correct/incorrect.

Results

Figure 3 shows the mean number of correct inferences in each of the conditions.

Significant differences were apparent between the accuracies of the different groups ($N = 28$, $F(3, 24) = 4.423$, $p = 0.013$). The data show a significant main effect of category variance ($F(1, 24) = 8.097$, $p = 0.009$). There is no significant *main* effect of adding labels ($F(1, 24) = 1.750$, $p = 0.395$). There is, however, a significant interaction between variance and label ($F(1, 24) = 10.321$, $p = 0.046$)[5]. The effect of labelling on inference therefore depends on the variability of the category used. One-way analyses of the effect of labelling at different levels of category variance demonstrate a significant effect of labelling for high variance categories only ($N = 14$, $F(1, 12) = 5.760$, $p = 0.034$)[6].

Thus, subjects were more likely to make correct inferences when they were making inferences from a low as opposed

[5]Note that the interaction is ordinal, but there is a change in direction. The low variance target performance is made worse by the addition of a label—the high variance target performance is improved.

[6]For low variance categories, $N = 14$, $F(1, 12) = 0.620$, $p = 0.446$.

to high variability category. Overall, there is not a significant effect of adding labels. This is explicable in terms of the surprising (and unreliable) reduction in accuracy when labels were provided in the low variability condition, as opposed to the increase in accuracy when labels were provided in the high variability condition.

Discussion

These studies support the predictions of the connectionist model of the development of inductive inference (Loose and Mareschal, 1997). The variability of the set of exemplars from which a category is inferred is important in subsequent inductive reasoning with that category. The more variable the set of exemplars, the greater is the effect of adding category labels.

Study 1 demonstrates that there is a tendency to overgeneralize properties of more variable categories to a greater extent than properties of less variable categories. This might explain the findings of some inference studies, in that it is harder to find perceptual distinctions between categories on which to base inferences when those categories are more variable. However, experiments 2 and 3 demonstrate that there is more to inference than this. Experiment 2 demonstrates that we have produced a set of stimuli taking JNCD into account, since we find no evidence of a difference in discriminability between the two categories used in that study. Experiment 3 investigates inference using these revised stimuli. The study finds that the accuracy of inferences from the high variability category are more affected by the addition of label information than inferences from the low variability category.

The principle that perceptually more heterogeneous categories will be more affected by category labels in inference is predicted by the Loose & Mareschal model. However, depending on the interpretation of these results, a reconsideration of the model may be required.

If it turns out that the two categories are in fact not equally discriminable, then the simplest explanation is that the perceptually more homogeneous category promotes both categorization and inference. Previously reported simulations demonstrate that this interpretation is consistent with the model's performance (Loose and Mareschal, 1997). However, our results suggest that we should treat the two categories as *equally* discriminable. Thus, we are led to argue that the integration of new conceptual information into a perceptually heterogeneous category is actually *more difficult* than the integration of new conceptual information into a perceptually homogeneous category. The previously reported model does not seem to account for this specific finding. The model is currently being developed to investigate what "more difficult" might mean in this context.

References

Florian, J. (1994). Stripes do not a zebra make, or do they - conceptual and perceptual information in inductive inference. *Developmental Psychology*, 30:88–101.

Gelman, S. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, 20:65 – 95.

Gelman, S. and Coley, J. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-yr-old children. *Developmental Psychology*, 26:796–804.

Gelman, S. and Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23:183—209.

Gelman, S. and Markman, E. M. (1987). Young children's inductions from natural kinds: the role of categories and appearances. *Child Development*, 58:1532—1541.

Hampton, J. A. (1995). Testing prototype theory of concepts. *Journal of Memory and Language*, 34:686—708.

Inhelder, B. and Piaget, J. (1964). *The Early Growth of Logic in the Child: Classification and Seriation*. Routledge & Kegan Paul Ltd, London.

Jones, S., Smith, L., and Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, 62.

Loose, J. J. and Mareschal, D. (1997). When a word is worth a thousand pictures: A connectionist account of the percept to label shift in children's reasoning. In Shafto, M. and Langley, P., editors, *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pages 454–459, London. Lawrence Erlbaum Associates.

Loose, J. J. and Mareschal, D. (1998). Inductive reasoning tasks revisited: Adults don't rely on label information when inferring hidden properties. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*.

Monk, A. F. and Eiser, J. R. (1980). A simple, bias-free method for scoring attitude scale responses. *British Journal of Social and Clinical Psychology*, 19:17–22.

Smith, E. and Medin, D. (1981). *Categories and Concepts*. Harvard University Press, Cambridge, MA.

Younger, B. (1990). Infants' detection of correlation among feature categories. *Child Development*, 61:614–620.