When More Evidence Makes Word Learning Less Suspicious

Gavin W. Jenkins (gavin-jenkins@uiowa.edu)
Jodi R. Smith (jodi-r-smith@uiowa.edu)
John P. Spencer (john-spencer@uiowa.edu)
Larissa K. Samuelson (larissa-samuelson@uiowa.edu)
Department of Psychology and Delta Center, E11 Seashore Hall
Iowa City, IA 52240 USA

Abstract

One challenging problem that children overcome in learning new words is recognizing the hierarchical category of a label. For instance, one object could be called a Dalmatian, a dog, or an animal. Xu and Tenenbaum (2007) proposed a Bayesian model to explain how 3.5 to 5-year-olds solve this ambiguity. They emphasized children's appreciation for "suspicious coincidences:" a label applied to three identical toys is interpreted more narrowly than a label applied to one toy. Xu and Tenenbaum did not investigate children's prior category knowledge, however. We replicated their "suspicious coincidence" effect and measured this knowledge. Unexpectedly, children with more category knowledge appreciated "suspicious coincidences" less. In a second experiment, repeatedly emphasizing novel labels caused all children to stop recognizing the "suspicious coincidence." These data are inconsistent with the Bayesian account and suggest the phenomenon is influenced by subtler aspects of prior knowledge and by task-specific details.

Keywords: Word Learning, Categorization, Bayesian Model

Introduction

A central issue in the study of word learning is how children acquire names for hierarchically nested categories. "Animal," "mammal," "dog," "Black Labrador," and "Rover" can all apply to the same referent. This presents a challenge to a young word learner, because when a child hears a label applied to a novel object, the correct interpretation is ambiguous in a hierarchically labeled system: does a novel label "fep" combined with an animal refer to the species, to the breed, or is it a proper name?

Additionally, some of the tools children usually use to decipher novel word-object mappings become less helpful in the case of hierarchically nested categories. Mutual exclusivity (Markman, 1989) is counterproductive in cases where two words both refer to the same object, but at different hierarchical levels (e.g., "animal" and a novel word for the same thing, like "Dachshund"). Any child relying on mutual exclusivity would fail to learn more than one hierarchical label for an object at a time. Golinkoff, et al.'s (1992) N3C constraint would be counterproductive for the same reasons: it rejects such overlapping labels by design.

To approach this problem, Xu and Tenenbaum (2007) recently suggested a Bayesian approach children might use to succeed at learning names for objects at multiple levels in a hierarchy. Specifically, Xu and Tenenbaum suggest that children recognize so-called "suspicious coincidences" when a label is applied to multiple, distinct exemplars that look very similar. For example, a child might hear the word "fep" applied to a Black Labrador dog. After just one labeling event, the word "fep" is ambiguous. Imagine that a few minutes later, however, the child again hears "fep" applied to a different Black Labrador. Now, the child can use Bayesian inference to suppress some possible interpretations: if "fep" refers to all animals, it would be a "suspicious coincidence" for the first two random examples that the child saw to both be examples of the same breed of dog. For the same reason, the evidence would also be suspicious if "fep" refers to only "dogs." It would not be suspicious at all, however, to see "fep" applied to two Black Labradors in a row if "fep" meant only "Black Labrador." Xu and Tenenbaum suggest that children can recognize when a label is applied to a "suspiciously" small subset of the possible objects it could refer to, and use this to infer the label carries a narrow meaning. If the child hears "fep" applied to Black Labradors a third and fourth time, the child's narrow interpretation becomes exponentially stronger, and any other interpretation is de-emphasized (see the "size principle," introduced on page 252 of Xu & Tenenbaum, 2007).

Xu and Tenenbaum (2007) experimentally tested children's ability to infer narrow meanings in "suspicious coincidence" situations. Participants (42-60 months of age) who were shown one stuffed toy Dalmatian labeled "fep" later generalized the label to a variety of toys at different levels of hierarchy: other Dalmatians, different breeds of dogs, and even a few other species of animals, like seals. However, when participants were shown three separate Dalmatians, all labeled "fep," they almost never generalized the label to anything but other Dalmatians. In this experiment, separate toys were all presented simultaneously, so there was no ambiguity about whether the toys were unique or whether they were the same toy. Xu and Tenenbaum explain this "suspicious coincidence" behavior as a natural extension of a Bayesian model of word learning.

The suspicious coincidence is a conceptually important phenomenon because it is not predicted by other models of word learning. Nevertheless, relatively little is known about how this phenomenon is related to the dramatic changes in word learning that take place in early development. Thus, the goal of the present study was to examine how this phenomenon is related to children's emerging category knowledge. Prior knowledge and the similarity structure

children bring to the task play a central role in Bayesian accounts, yet these aspects of knowledge are rarely measured directly (Jones & Love, in press). It is possible that the exact details of category knowledge play a critical role in the suspicious coincidence. Individual children may have highly variable and difficult to predict histories of exposure that could influence their behavior under a Bayesian perspective.

Thus, in Experiment 1, we measured children's prior category knowledge and its relation to their noun generalization behavior in Xu and Tenenbaum's task.

Experiment 1 was an exact replication of Xu and Tenenbaum's (2007) Experiment 3, investigating the effects of the "suspicious coincidence" on children's novel noun generalization. Our only modification was that parents of participants completed an additional vocabulary questionnaire during the experiment. We compared the strength of children's "suspicious coincidence" effect with their prior knowledge of the categories used in the task. The Bayesian account predicts that children who know more about these categories should have shown a stronger "suspicious coincidence."

Experiment 1

Methods

Participants were 54 monolingual children from a Midwestern town, between the ages of 42 and 60 months. 13 participants were excluded from analysis: 7 for choosing distractor items, 1 for fussiness, 2 due to experimenter error, and 3 for extreme, reversed generalization behavior from adults. Therefore, 41 participants were included in analyses (mean age = 4 years, 3 months; range from 3 years, 6 months to 5 years, 0 months). Parents of participants were contacted via mail and a follow-up phone call. Parents provided informed consent prior to the study. Each participant received a small toy for participation.

The stimuli were chosen based on the set used by Xu and Tenenbaum (2007). Most toys used were of the exact same adult subordinate categories as those in Xu and Tenenbaum's experiments, and all toys conformed to the same basic level categories. The set of 45 toys was divided into three superordinate categories (referred to as "sets"): 15 animals, 15 vehicles, and 15 vegetables. Each set was further divided into basic level categories: 6 toys from different basic level categories and 9 toys from the same basic level category. Each basic level category of 9 toys was further divided into subordinate level categories: 4 toys from the same subordinate level category and 5 toys from different subordinate level categories. Using the animal set as an example, the category structure of the 15 animals was as follows: 6 "superordinate matches" from unique basic level categories (penguin, pig, cat, bear, seal, bee) and 9 "basic matches" from the same basic level category, 5 of which were also "subordinate matches" (Husky, Sheepdog, Pug, Terrier, and 5 examples of Black Labradors). Seven toys from each set were reserved as possible exemplars, and 8 were placed in a test array as generalization targets. Figure 1 shows an example of each possible combination of familiarization exemplars a child could be shown (only the animal set possibilities are depicted). Figure 2 shows the array of test objects that *every* child saw.

A vocabulary survey was developed to examine children's knowledge of the labels for the stimuli at each level in the taxonomic hierarchy. The survey was filled out by parents during the study and included an entry for each unique stimulus item used in the experiment. Each entry of the survey included a photograph of a toy, a line on which parents were instructed to write down what the child would spontaneously call the toy, and check boxes with different category labels for that toy that children might recognize (but not necessarily produce).

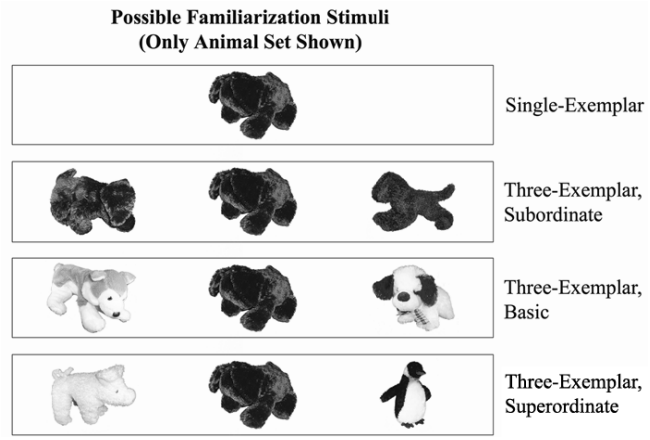

Figure 1: Possible familiarization conditions. Each child participated in only one condition over three trials.

Aside from parents filling out the vocabulary survey described above, the design of our first experiment was exactly the same as that of Experiment 3 of Xu & Tenenbaum (2007). It was divided into three trials, each with a familiarization phase and a test (generalization) phase. Every child was randomly assigned to either a "one-exemplar" condition or a "three-exemplar" condition.

The familiarization phase was performed first. In the one-exemplar condition, the experimenter pulled out one subordinate match toy, placed it on the exemplar mat, and labeled it three times in a row. In one trial of the three-exemplar condition, the experimenter pulled out one subordinate match toy and two other identical toys, and each of the three toys was labeled once (trials with three different basic level toys and three different superordinate level toys were also run, but are less direct as a test of the "suspicious coincidence"). Novel labels were used in both conditions, such as "fep." Participants in both conditions heard the same number of object-word pairings overall, either one object three times or three objects once each.

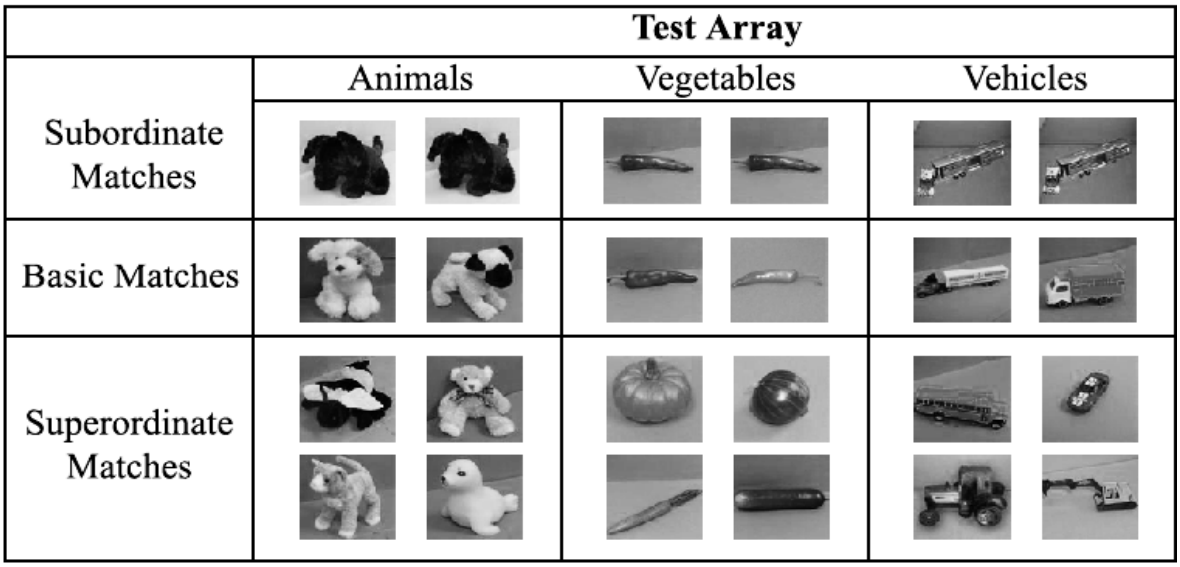

Figure 2: The toys in the test array, by category. These were placed on the floor in a random order.

Next, the test phase of the experiment began. The test phase involved a series of Yes/No generalization questions (e.g., "is this a fep?") about a subset of the toys in the test array. The stimuli presented in the test array are pictured in Figure 2. Note, however, that the test objects were randomly positioned for the experiment, rather than grouped as in the figure. Of the 24 objects in the test array, ten were used as the test set for a given trial. The test set included the two available subordinate matches, the two basic matches, the four superordinate matches, and any two distractors from the other two sets.

Results

Replication As Experiment 1 was procedurally a replication of Xu and Tenenbaum's third experiment, we first examined whether the "suspicious coincidence" effect was, in fact, replicated in ours. Figure 3 summarizes Xu and Tenenbaum's results from the relevant experiment, compared with our replication results. The data in all of our graphs are based on the proportion of "Yes" responses to different hierarchical levels of matching in the test phase.

The most direct test of Xu and Tenenbaum's "suspicious coincidence" is the difference between the basic level (middle) bars in the one-exemplar versus 3-exemplar-subordinate conditions. The "suspicious coincidence" is seen as a drop in the proportion of basic match generalizations (e.g., "Yes" responses to the Sheep dog and Pug) from when one exemplar (Black Labrador) is shown to when three of that same exemplar are shown. Our replication of the effect was a success. Basic match generalization decreased significantly when three subordinate match exemplars were used, both in Xu and Tenenbaum's study and in ours (From 25% to 13% in our results, $p < 0.05$; and from 40% to 6% in Xu & Tenenbaum's results in their 2007 study).

Results by vocabulary knowledge The main motivation for Experiment 1 was to look more carefully at the effect of prior category knowledge on word generalization, specifically the "suspicious coincidence" phenomenon. We will focus only on subordinate level vocabulary knowledge, because labels at this level are most relevant for evaluating the "suspicious coincidence" effect.

Vocabulary knowledge for the most specific category for each toy was summed over the entire survey and recorded as the child's "specific vocabulary score," or SVS. The maximum possible SVS was 30, but no child received this (mean 16.83, median 16, range 12-23). A median split was performed on the scores in order to divide children into "low-SVS" and "high-SVS" groups (with additional subjects recruited as necessary for statistical power, included in the methods sections of this paper). We then analyzed the "suspicious coincidence" for each group. A dramatic difference was found: low-SVS children showed a 25% drop in basic match generalization from the one-exemplar to the three-exemplar condition (the "suspicious coincidence" effect. $p < 0.0001$), while high-SVS children showed a 2% *rise* in basic match generalization (the opposite trend of the "suspicious coincidence"). This data is

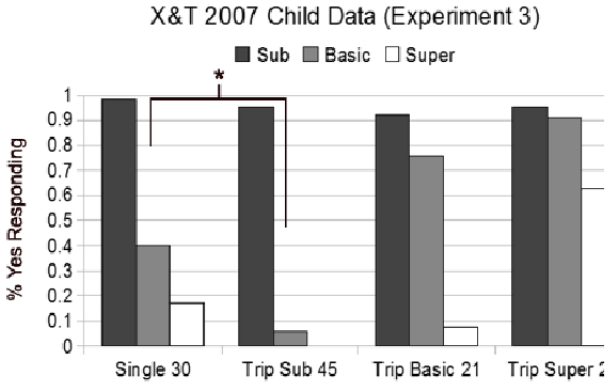

Figure 3a: Xu & Tenenbaum's results

presented in Figure 4. The results of our vocabulary survey are not correlated with age ($r < 0.1$, $p > 0.25$).

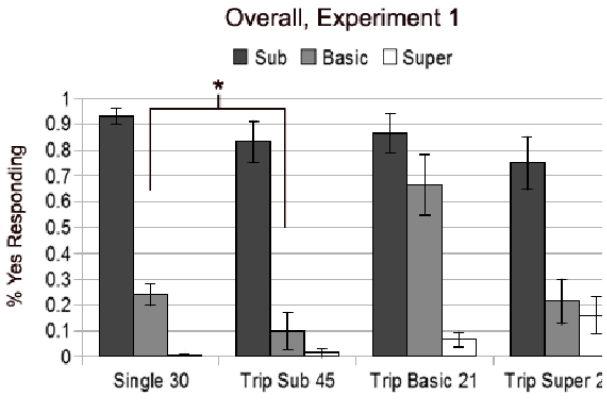

Figure 3b: Overall results of Experiment 1. we succeeded in replicating Xu & Tenenbaum (2007).

Discussion

The goal of Experiment 1 was to examine the effect of children's prior category knowledge on noun generalization at multiple hierarchical levels, specifically the "suspicious coincidence" effect. The vocabulary effect we discovered was clear: "suspicious coincidence" behavior was highly *inversely* correlated with children's vocabulary knowledge, as measured by our parental survey. Since vocabulary was not a manipulated variable, the direction of the relationship is unclear. However, *either* causal direction is difficult to explain from a Bayesian perspective. In the Bayesian account, children with the most detailed knowledge about subordinate-level categories like "Black Labradors" and other breeds of dogs should be able to most easily determine which sets of stimuli are statistically "suspicious" when given the same novel label. After all, if a child only knows about two breeds of dog, seeing three of one kind at once is not very "suspicious." Alternatively, if the "suspicious coincidence" drives vocabulary learning, then those most skilled at it should have acquired greater vocabulary knowledge. Our findings in Experiment 1 run counter to both of these possibilities, and thus suggest that children are not learning words according to Bayesian principles.

Before accepting this conclusion, however, we ran an additional experiment to rule out two alternate expanations for our vocabulary effect. Firstly, children who knew more of the correct English names for the subordinate categories tested may have had trouble learning a second set of labels for the exact same set of referents (even though mutual exclusivity is unhelpful for partially overlapping labels, it is useful for ruling out exactly overlapping labels). Secondly, even if the novel label was learned and accepted, if these children were temporarily distracted from the goal of the task, then they could have forgotten the new label and played the game based entirely on their English labels for the toys. Low-SVS children, on the other hand, were less likely to have their learning of novel labels blocked by existing knowledge and were less likely to default to English labels they didn't have.

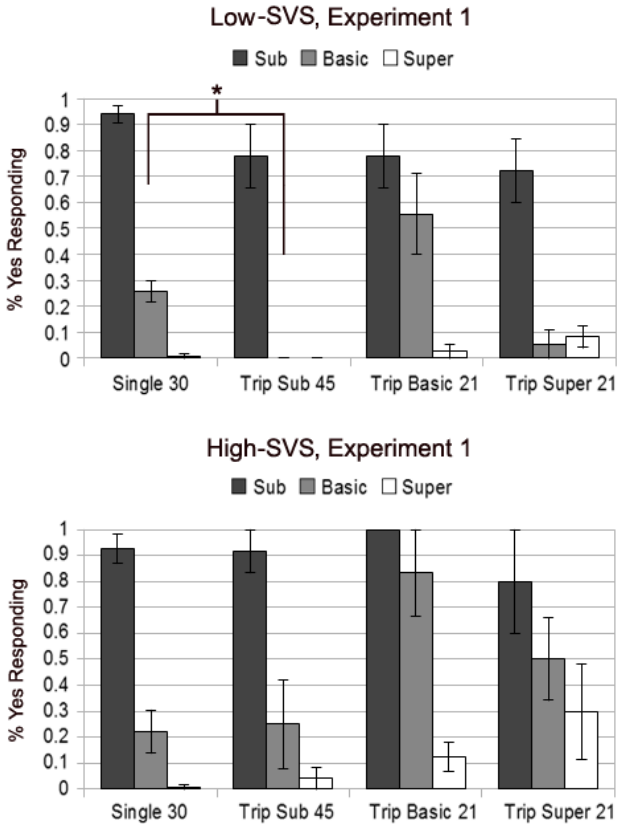

Figure 4: Results of Experiment 1 by prior vocabulary knowledge. Low vocabulary participants are driving the "suspicious coincidence" effect seen in the overall data.

Thus, in Experiment 2, we increased the number of times we recited Mr. Frog's labels to children. This should have helped overcome high-SVS children's strongly learned English names for the objects. We also spread this repeated reinforcement throughout the test phase of the task. This distribution of reminders was designed to prevent children from getting distracted. If high-SVS children in Experiment 1 were indeed distracted or not learning novel labels at all, then they should have begun to show a stronger "suspicious coincidence" in Experiment 2.

Importantly, our altered methodology should not affect (or if anything, should magnify) the predictions of Xu and Tenenbaum's formal Bayesian account of the "suspicious coincidence." The only evidence for a child that is relevant to their model is *new* evidence. No parameters exist in their (2007) equations that can model the effects of repeated exposure to old information.

Experiment 2

Methods

45 participants (again, aged 42-60 months) were recruited in the same manner as in Experiment 1 (a different number of participants was required due to not screening vocabulary knowledge before children entered the laboratory). 5 participants were excluded from analysis, all for choosing distractor items. The stimuli were also the same as in Experiment 1.

The procedure was also identical to that of Experiment 1, except for one modification: instead of a familiarization phase and then a test phase, the familiarization phase was expanded and intermixed with test questions. After each of the first nine test questions, the labeling of exemplar(s) was repeated. Remember that in Experiment 1, there was a single round of labeling, followed by all test questions without any familiarization reminders. In Experiment 2, labeling was also repeated in between every single test question. Instead of three object-label pairings per trial as in Experiment 1, this new procedure resulted in thirty object-label pairings per trial, or 16.5 object-label pairings heard by the child prior to each test question *on average.*

Results

Results for Experiment 2 are presented in Figure 5, in the same format used previously. Experiment 2 was designed to manipulate "suspicious coincidence" behavior in high-SVS children. We expected that by reinforcing our novel vocabulary words repeatedly, we would either cause no change to "suspicious coincidence" behavior, or (as hoped) we would help to keep high-SVS children focused and open to the novel word generalization task and thus cause stronger "suspicious coincidence" behavior. What we found instead was that high-SVS children still show the reverse of the "suspicious coincidence," as in Experiment 1. Thus, no benefit was gained by providing extra novel label exposure. Surprisingly, however, low-SVS children now *also* fail to show any "suspicious coincidence." Rather than helping to guide children toward more "rational" word interpretations, our attempts to make the task more conducive to "suspicious coincidence" behavior actually had the very different effect of reducing this behavior.

Before going any further with theoretical conclusions, we performed two analyses to confirm that our results were not due to children becoming bored by the many extra labeling events. First, we reviewed our video recordings to measure the length of the familiarization + test phases for the two experiments. In a sample of 30 participants' videos, the length of Experiment 2 was not found to be significantly different from the length of Experiment 1 in either condition ($p > 0.5$ for each condition) Thus, the time spent playing the game could not have been a source of boredom in Experiment 2. Second, we compared children's generalization behavior for objects in the first half of each trial versus the last half of each trial, in Experiment 2 only. If repetitive labeling events are a source of behavior-changing boredom for children, then their responses in the last half of each trial should be different than in the first half, as boredom accumulates. We found no differences in "Yes" responding between the two halves of each trial in chi-square tests, either for all toys ($p > 0.25$) or for basic-level toys only ($p > 0.25$). Together, these results suggest that boredom does not appear to play a role in causing our Experiment 2 results.

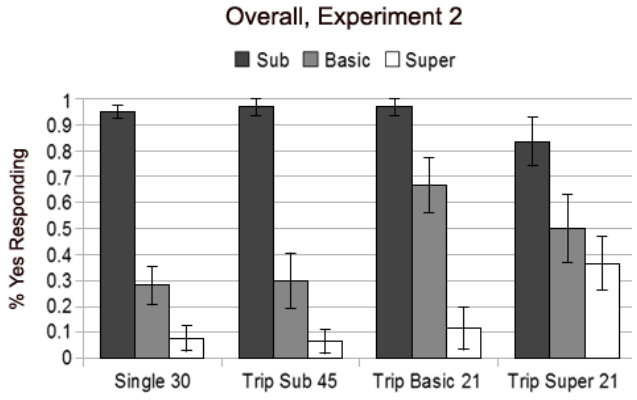

Figure 5: Results, Experiment 2. The suspicious coincidence effect has been eliminated, both overall and in the low-SVS group (not shown). High-SVS children continue to not show a suspicious coincidence effect (also not shown).

Discussion

Experiment 2 was designed to provide high-SVS children with an ideal environment for potentially demonstrating "suspicious coincidence" behavior. It was possible that high-SVS children were not learning or maintaining representations of our novel labels strongly enough, due to their more entrenched English labels for the same categories. Making the novel stimuli more salient/readily available, however, did not change the results of Experiment 1, as we expected from the above hypothesis. Instead, the results of Experiment 2 deviated even further from the predictions of Xu & Tenenbaum's Bayesian account. Not only did high-SVS children still fail to demonstrate a "suspicious coincidence", but the "suspicious coincidence" effect was eliminated and almost reversed ($p = 0.21$, trending in the opposite direction). Even low-SVS children alone failed to show the predicted effect. Overall, then, children in Experiment 2 failed to show a rational interpretation of our novel category labels.

One possible explanation of these results is that when children repeatedly glance back and forth between exemplars and the test array, there is less time to observe and process the entire test array or (in the case of three exemplars) the entire exemplar set. Instead, children in the more attentionally demanding Experiment 2 may be focusing more exclusively on only the current test item or the current exemplar being held up in front of them, compared to Experiment 1. This may cause all conditions to resemble the single exemplar condition to participants,

causing broader generalization. Easier structural alignment of highly similar objects—like in Experiment 1, where attention could linger long enough to draw comparisons amongst exemplar—has been shown to highlight differences (Gentner, 2001). In our case, this would lead to a stronger "suspicious coincidence" effect, as observed. Explanations like this remain to be tested, but one thing that is clear from the current results is that that any theory behind these results must take into account specific, task-related details, not high level accounts of rational inference alone.

General Discussion

In Experiments 1 and 2, children generalized novel labels in unexpected ways in response to two variables: increased prior vocabulary knowledge and an increased number of redundant labeling events. It is not clear how a Bayesian account alone can explain either of these behaviors, since children's behavior clearly departs from the mathematically optimal solution to this word learning task. Prior category knowledge as a *helpful* source of information for word learning is at the core of the Bayesian theory, and it is not clear how one would even implement redundant labeling in Xu and Tenenbaum's model. The closest approximation of such an implementation would also predict that redundant labeling should strengthen the "suspicious coincidence" effect, yet if anything, it weakens the effect.

Non-Bayesian behavior has been found in the adult version of Xu and Tenenbaum's task as well. In particular, Spencer et al. (in press) reported that adults failed to show a suspicious coincidence when items were presented sequentially (rather than simultaneously). This was the case even when six subordinate-level examples were presented.[1]

The experiments reported here and data from Spencer et al. suggest that word learning is a sensitive and interactive process that depends largely upon the specific environment and circumstances under which it occurs. The "suspicious coincidence" effect relies on details about prior vocabulary knowledge, the manner in which experimenters ask their generalization questions, and sequential versus simultaneous exemplar presentation. Ongoing work also suggests that children's perceptual similarity judgments change depending on the statistical structure of sets of toys they observe. When three Black Labradors are included in a set of toys to sort based on similarity, children treat them differently in relation to other toys than when only one Black Labrador is included in the set (Jenkins & Samuelson, 2011). This suggests that even for a given child, prior vocabulary, and stimulus set type, category-relevant feature information is likely task-dependent as well.

The Bayesian approach does not account for these idiosyncrasies, because it explicitly lives at a computational level and does not specify the processes by which task factors and category knowledge cohere in the moment to create word learning.

[1] Note that this sequential presentation effect could also be explained from a task-specific, structural alignment standpoint (Gentner, 2001), similar to that discussed for Experiment 2.

Instead, our current (and Spencer et al.'s) findings demand explanations of a mechanistic nature. Mechanistic explanations link behavior to more general processes of attention, memory, and perception, while allowing for the influence of particular kinds of tasks and contexts. The mechanistic perspective is powerful for understanding processes that change non-linearly based on different memories, changes to the task context, and so on. As shown in Experiments 1 and 2 and in Spencer et al. (in press), the "suspicious coincidence" effect is a variable phenomenon that comes and goes in different situations based on a host of factors. Thus, a mechanistic perspective is a natural choice for making sense of this emerging body of data.

Accordingly, future research must include a mechanistic-level model of children's word learning. This does not mean that a rational account is not also useful. After all, Xu and Tenenbaum's model predicted the existence of the "suspicious coincidence" effect and introduced this behavior to the field. Nevertheless, future work will need to move beyond the rational view to explain when, why, and *how* children show this finicky behavioral pattern and how the "suspicious coincidence" is linked to emerging category knowledge and word learning in early development.

Acknowledgments

This project was funded by grant number R01HD045713 to Larissa Samuelson.

References

Markman, E.M. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press.

Gentner, D. & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory and Cognition, 29(4),* 565-577.

Golinkoff, R.M., Hirsh-Pasek, K., Bailey, L., & Wenger, N. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, *28,* 99-108.

Huttenlocher, J., Newcombe, N., & Sandberg, E. (1994). The coding of spatial location in young children. *Cognitive Psychology, 27,* 115-147.

Jenkins, G.J. & Samuelson, L.K. (2011). Non-featural influences on child similarity judgments. Submitted for publication.

Jones, M. & Love, B.C. (in press). Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition. *Behavioral and Brain Sciences* (target article).

Spencer, J.P., Perone, S., Smith, L.B., & Samuelson, L.K. (in press). The process behind the 'Suspicious Coincidence:' Using space and time to learn words. *Psychological Science.*

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review, 114(2),* 245-272.