Understanding the effect of social context on learning: A replication of Xu and Tenenbaum (2007b)

Molly L. Lewis

Department of Psychology, Stanford University

Michael C. Frank

Department of Psychology, Stanford University

Address all correspondence to Molly L. Lewis, Stanford University, Department of Psychology, Jordan Hall, 450 Serra Mall (Bldg. 420), Stanford, CA, 94305. Phone: 650-721-9270. E-mail: mll@stanford.edu

Abstract

Keywords: word learning, induction, Bayesian inference

Introduction

Imagine you were on a hike and saw a rock positioned oddly at an ambiguous intersection of trails. If you thought you were on a trail which no one had traveled in years, you probably wouldn't think much of that rock. But, if you knew your campmate had traveled that same trail earlier that day, you might interpret the rock differently: You might interpret the rock as a sign intended to point you to the correct path. This intuition—that the source of a piece of information influences the strength of the inference to be drawn—suggests that social information may have a privileged role in human learning.

Xu and Tenenbaum (2007a) examine this phenomenon in the context of a particularly difficult inductive problem: Concept generalization in word learning. Faced with a novel word and its referent, children must decide between an infinite number of hypotheses about the concept extension of that word. For example, consider a child who hears the word "banana" in the context of a single banana on a table. While the referent of that object is clear in the moment of language use (i.e., the particular banana on the table), the broader concept is highly ambiguous: "banana" could refer to the category of bananas, the category of fruit, that particular species of banana (e.g., plantains), yellow things, or any number of other ad-hoc categories.

Xu and Tenenbaum (2007a) ask whether children make use of the information source of a new word to guide their inferences about how to generalize its meaning. In particular, they test a prediction that falls out of their word learning model. In their model, learners observe data as word-object pairs and make inferences about the concept associated with that word from a hypothesis space of all possible meanings. Critically, the model predicts that an ideal learner should generalize more broadly when the exemplar is sampled from the full hypothesis space of meanings (*weak sampling*), and should generalize more narrowly when the exemplar is sampled from only the true concept of the word within the full hypothesis space (*strong sampling*).

In their experiment, Xu and Tenenbaum (2007a) manipulate sampling "strength" through the presentation source of the data. The learner is either presented with three exemplars of the target word from a knowledgeable teacher or the learner makes (correct) guesses about the referent of the word. Critically, in both conditions, the data that the participants observes is the same: three exemplars from the same subordinate category. What differs is the strength of the sampling. Since the teacher knows the true concept, the data are sampled strongly from the true concept. But, in the learner-generated condition, the learner is naive about the true underlying concept and thus the data are sampled weakly from the full hypothesis space. The key prediction is that, given a hypothesis space with hierarchical concepts (basic, subordinate, superordinate; i.e., banana, plantain, fruit), participants in the teacher condition should be more likely to generalize broadly to the basic level, while participants in the learner condition should generalize conservatively to only the subordinate condition. Their data strongly support this prediction in both adults ($d = 1.49$ [0.02, 2.97]) and children ($d = 1.21$ [.19, 2.23]).

There are a number of reasons to conduct a replication of this study. First, replication attempts of other predictions of this model have challenged this framework. In a different paper, Xu and Tenenbaum (2007b) use the same model to make predictions about the effects of another factor in the learning context: the number of observed exemplars. In particular, the model predicts that a learner should generalize more narrowly when given three observations of a subordinate category, compared to just one (the "suspicious coincidence effect"). Across three experiments, they find both adults and children behave consistent with this prediction. Follow-up work has challenged this result, however (Jenkins, Samuelson, Smith, & Spencer, 2015; Spencer, Perone, Smith, & Samuelson, 2011).

There are additional theoretical reasons to attempt a replication of this result, in particular. For example, there is evidence that there may be a large degree of variability across participants in the strength of their sampling assumptions (Navarro, Dry, & Lee, 2012). More generally, evidence suggests that effects that rely on social manipulations are less likely to replicate than effects in more "cognitive" domains (Open Science Collaboration, 2015).

Finally, this study is important to replicate because the broad theoretical question—how the source of information influences learning—has far-reaching implications for our understanding of human learning. Every piece of data observed in the world, including the experimental context, is observed in some social context. While the degree of this social pressure may vary (consider yourself observing flowers alone in a forest versus a case observing flowers received from your partner on Valentine's day), humans are always part of a social system. Thus, in our effort to understand how learners make inferences on the basis of observations in the world, it is important to understand what factors influence this inference, and the source of these observations is likely an important factor.

The social context of human learning also has practical consequences for the interpretation of data collected in psychological experiments. This is because experimental data are often consistent with at least two accounts—an account that relies on reasoning about the intention of the experimenter, and an account that relies on context-independent reasoning. Consider two examples. A well-known phenomenon in word learning is that children are biased to select a novel object for a novel word, given the presence of both a familiar and novel object (often referred to as *mutual exclusivity* in the literature, Markman & Wachtel, 1988). This pattern is difficult to account for psychologically, however, because there are at least two accounts of this behavior. On the one hand, this result could be due to a context-independent bias to assume that lexicons are structured with one word mapping to one concept, and one concept mapping to one word. Another possibility relies on reasoning about the intentions of the experimenter (Why would the experimenter use a strange word to refer to the familiar object if she meant the familiar one?; Clark, 1987, 1988). Both of these accounts make similar predictions, and are therefore difficult to disentangle empirically.

Another example of this interpretative ambiguity is the Heider and Simmel (1944) study. In this task, participants viewed a short movie showing several geometric shapes moving in a way that appeared to be contingent. Nearly all participants spontaneously interpreted the video as depicting animate beings, rather than as simple shapes moving around. Like in the case of mutual exclusivity, there are at least two ways to interpret this result. One possibility is that participants rely on low level features of the scene to infer animacy (e.g., contingency), but another possibility is that participants infer the intention of the experimenter who created the videos and assume an animate intention.

These two cases—mutual exclusivity and animacy projection—represent a sample of a pervasive theoretical issue in experimental psychology: Data consistent with both pragmatically rich and context-independent accounts. There is not a simple solution to this empirical challenge because it is impossible to fully eliminate a social context from experimental paradigms. Our best bet, therefore, is to try to understand the influence of the social context on learning, and Xu and Tenenbaum (2007a) represents a insightful attempt to systematically shed light on this practical issue.

Thus, given the challenges to the ? (?) result, as well as its theoretical and practical significance, we sought to replicate it. We conducted four replications of Xu and Tenenbaum (2007a), three online (Exp. 1, 2, and 4) and one in-person (Exp. 3). Our data replicate the original effect, but suggest an effect much smaller in magnitude than the original report. In the General Discussion, we suggest a number of factors that may influence the magnitude of this effect that should be explored in future work.

Experiment 1

In Experiment 1, we attempted to replicate the original design in an online paradigm.

Methods

Participants. The original sample with adults included 14 participants. We recruited 294 [not sure how to justify this number] adult participants online through Amazon Mechanical Turk. Participants were paid US $0.25.

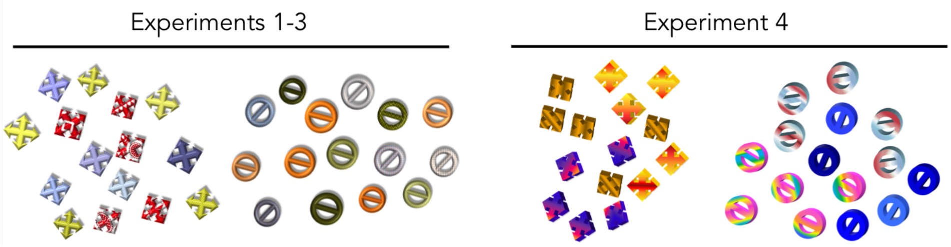


Figure 1. Sample stimuli used in Experiments 1-3 (left) and Experiment 4 (right). The Experiment 4 stimuli are intended to be maximally similar to the original Xu and Tenenbaum (2007b) stimuli, with lower subordinate-level variability than the stimuli used in Experiments 1-3. The full set of stimuli can be accessed here: xx.

Stimuli. We created four sets of objects similar to the original stimuli (Figure 1, left).[1] There were four basic level categories that included 15 unique objects each. Within each basic level category, there were three subordinate categories, with 5 unique objects each. The same novel words were used as the original: “wug”, “tupa,” “blicket,” and “fep.”

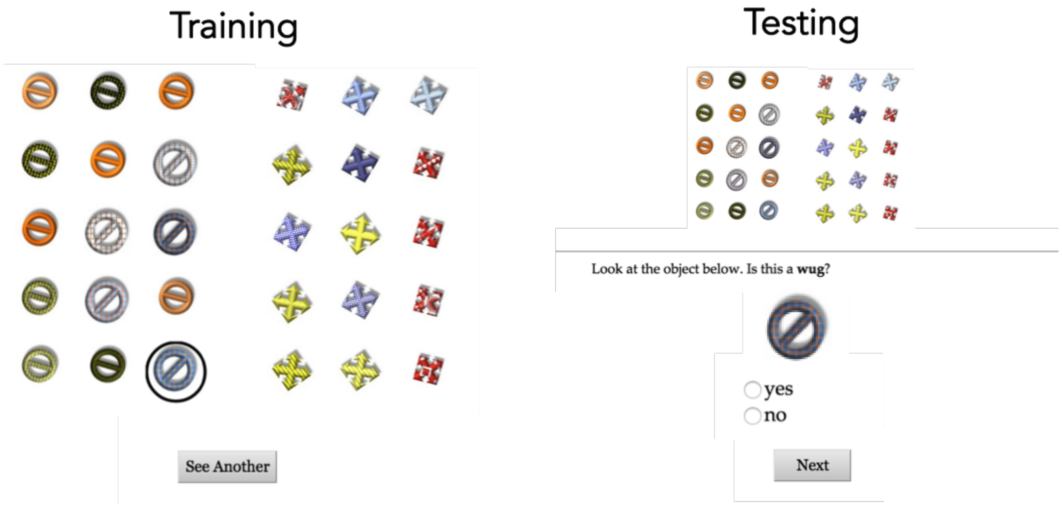
Procedure. Participants first viewed an instruction page that described the task. In the teacher condition, the instructions read:

> In the first part of the experiment, you will see pictures of objects. Some of the objects are called *wugs*. In order for you to learn which objects are *wugs*, I will circle three of them for you. After you learn about the *wugs*, you will be asked questions about them.

In the learner condition, the instructions were identical except the second sentence above was replaced with: “In order for you to learn which objects are *wugs* you will try to find two of them. I will tell you whether you are right or wrong.”

Participants then viewed a screen showing all the objects from two basic level categories. Within each basic-level category, the shapes were arranged in a 5 x 3 grid, and the two categories

[1] All stimuli, experiments, raw data and analysis code can be found at `https://github.com/mllewis/xtSamp/tree/master/`. The analyses can be viewed directly here: `http://rpubs.com/mll/xtSamp`



Look at the object below. Is this a **wug**?

○ yes
○ no

Figure 2. Screen shots of the training (left) and testing (right) phases in Experiment 1.

were spatially separated (Figure 2). One of the objects in the category on the left was circled. In the teacher condition, the instructions read: “Find the object that is circled below. That object is a *wug*. When you click on the ‘See Another’ button, I will show you another *wug*.” The participant was then asked to press a button which caused a circle to appear around one of the exemplars from the same subordinate category as the initially circled object. The participant clicked the button twice in total.

In the learner condition, the instructions were identical, except the last sentence was replaced with: “The object circled below is a *wug*. Click on two more *wugs*.” Participants were then asked to click on two more objects. After each click, a pop-up window appeared with the text “You’re correct! That’s a wug.” This text appeared regardless of the object the participant clicked on. The display then showed the object with a circle around it to indicate that it had been selected. Critically, in both the teacher and the learner conditions, the final display was identical: Thirty objects from two basic level categories, with three circled exemplars.

Participants then advanced to the test phase. On the test screen, the objects from the training phase were shown at the top of the page in an identical format to the training phase, but without the exemplars circled. Below these objects was a horizontal line, and a generalization question (Figure 2). There were five generalization questions presented sequentially in the same order as in the original report (subordinate match, basic non-match, basic match, subordinate match, basic match). In each test question, one of the exemplars was shown with the following text above: "Look at the object below. Is this a wug?" Participants responded by marking "yes" or "no" using radio buttons. [then we asked some other questions about a new category – is it necessary to describe this?] Finally, we asked an attention check question where participants had to select the label they previously learned from four alternatives.

Objects categories and word items were randomized across participants. The order of presentation of the individual objects on the screen was also randomized, as was the placement of the first circle. Sampling condition was manipulated between-participants. This and all subsequent online paradigms can be viewed directly here: `https://mllewis.github.io/projects/xtSamp/xtSampindex.html`.

Results and Discussion

We excluded participants from our analysis who responded "yes" to the basic non-match question ($N = 8$) or who did not select subordinate matches in the learning condition ($N = 21$), because this pattern of response indicated a misunderstanding of the task. No participant missed the attention check question. Our final sample included 274 participants ($N_{learner} = 128$; $N_{teacher} = 146$).

In all four experiments, we adopt two different criteria for categorizing participants' response pattern. Unlike in the original report, we found that not all participants responded consistently across questions of the same type within a trial (for example, a participant might respond "yes" to one subordinate match and "no" to another). Thus, we adopted a liberal criteria which categorized a participant as a basic-level generalizer if they responded "yes" to both the

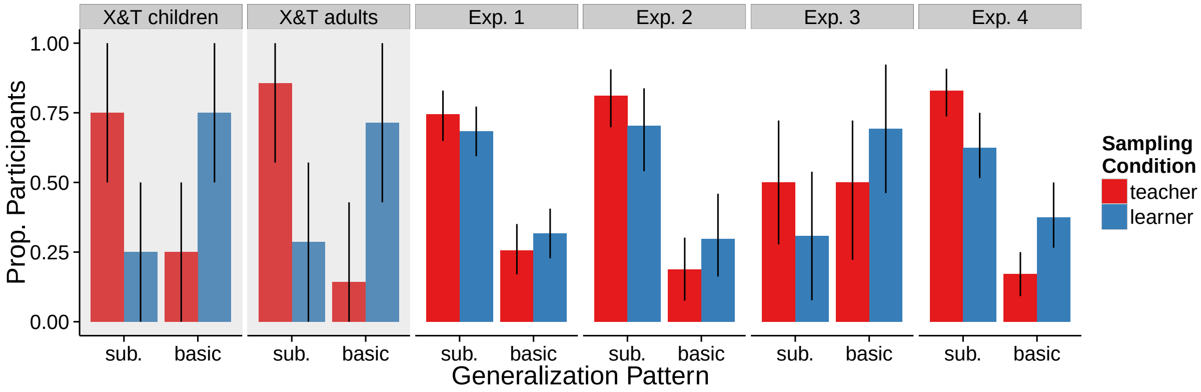


Figure 3. Proportion participants generalizing to the subordinate (sub.) or basic level category in the original child ($N = 24$) and adult experiment ($N = 14$), and in our four replication attempts ($N_1 = 294$; $N_2 = 150$; $N_3 = 41$; $N_4 = 200$). Experiments 1, 2, and 4 were conducted online, and Experiment 3 was conducted in-person. Error bars reflect 95% confidence intervals, calculated via non-parametric bootstrapping. Note that we report the Xu and Tenenbaum data by aggregating across participants, rather than trials, as in the original report.

subordinate matches and *at least one* basic-level match. We also analyzed our data using a strict criteria, where a participant was categorized as a basic-level generalizer if they responded “yes” to both the subordinate matches and *both* basic-level matches. Under both criteria, a participant was categorized as a subordinate-level generalizer if they responded “yes” to only the subordinate level matches. We excluded participants from our analysis who could not be categorized under the criteria. We report the results using the liberal criteria in the Main Text and describe the results using the strict criteria in Appendix A. The significance of none of the results depends on the criteria used.

Under the liberal criteria, 79 participants could not be categorized as either basic or subordinate-level generalizers, and thus were excluded. While the remaining responses followed the same qualitative pattern as the original, the difference between sampling conditions was not reliable ($\chi^2(1) = 0.63$, $p = .42$; Figure 3). The effect size was also much smaller than the original finding ($d = 0.17$ $[-0.17,\ 0.51]$; Figure 4).

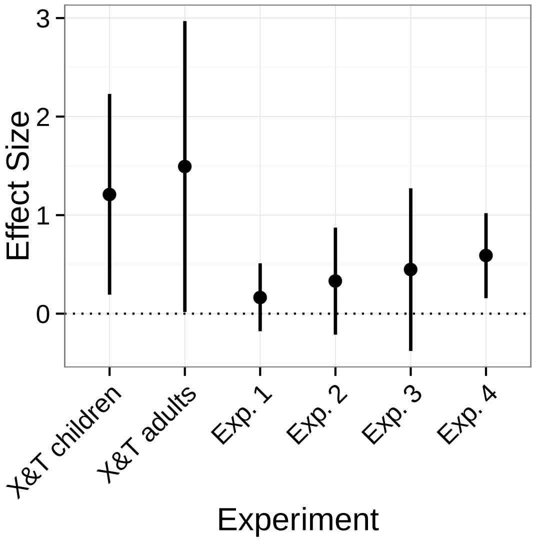


Figure 4. Effect size for the original experiment with children and adults, and our four replication attempts. Effect sizes were calculated using the log odds ratio (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003). Error bars are 95% confidence intervals.

Experiment 2

Given that we did not replicate the original effect, we sought to alter our paradigm in Experiment 2 to more closely replicate the original in-person experiments. One critical difference between the in-person and online paradigm is the salience of the experimenter: In the in-person version, the teacher was an actual person interacting with the participant, while in Experiment 1, the reality of the teacher had to be inferred based only on first person text (“I will show you...”). Thus, one possible reason we observed a smaller effect in Experiment 1 is that participants may not have assumed strong sampling in the teacher condition. In Experiment 2, we tried to strengthen this manipulation by introducing the teacher with a picture.

We made also made several other changes to our design. Because many participants in Experiment 1 did not generalize at all (answered “no” to all generalization questions), we added questions that queried the original three exemplars that participants learned about. These “proper name" trials allowed us to more directly understand these participants' generalization strategy. We also reduced memory demands between the training and testing phases by circling the training exemplars in the testing phase, as well as the training phase.

Methods

Participants. In Experiment 2, we recruited 150 participants from Amazon Mechanical Turk. Participants were paid US $0.25 for their participation.

Stimuli. The object and word stimuli were identical to Experiment 1.

Procedure. The procedure was identical to Experiment 1, with the exception of several key changes described below.

First, we increased the saliency of the experimenter by adding a clipart image of a woman. In the initial instructions, we introduced her by saying, "Hi, my name is Natalie." The image of the teacher was also present in the training and testing phases. These changes were made in both the teacher and learner conditions.

Second, we slightly altered the language of the experimenter to sound more natural. In the learner condition, the instructions in the training phase read: "Click on two more objects that you think might also be called wugs." We also changed the feedback in the training phase to be, "Yeah, that's a wug." In the testing phase, we added the following text above the training items: "You identified the objects below as wugs" (the training exemplars were circled, see below). In the teacher condition, the instructions in the training phase read: "Find the object that is circled below. That object could be called a wug. When you click on the "See Another" button, I will show you another wug." In the testing phase, we added "I showed you these objects were wugs" above the training items. We also changed the critical question to be "Could this be called a wug?," instead of "Is this a wug?" in both conditions. This was done to increase the number of basic-level interpretations of the generalization question.

Third, we added more generalization questions and randomized their order. Each participant was queried about 10 objects in total: The three training exemplars, two objects from the same subordinate level category as the training exemplars, three basic matches, and two basic non-matches.

Fourth, we reduced memory demands between the training and testing phases by showing the full set of training items during testing, as in Experiment 1, but also leaving the selected exemplars circled. This ensured that participants remembered which objects had been identified as examples of the target category.

Finally, we added three additional check questions. We showed participants an object from the target category, as well as two objects from never-seen categories. For each of these objects, we asked "Did you learn about this kind of object?." Participants responded by indicating "yes" or "no" on a radio button.

Results and Discussion

As in Experiment 1, we excluded participants who responded "yes" to the basic non-match question ($N = 15$) or who did not select subordinate matches in the learner condition ($N = 22$). We also excluded participants who missed any of the attention check questions ($N = 9$). Our final sample included 118 participants ($N_{learner} = 43$; $N_{teacher} = 75$).

The criteria for categorizing a participant's generalization strategy was identical to Experiment 1, except for the inclusion of the additional questions. To be categorized as either a subordinate or basic-level generalizer, a participant had to respond "yes" to all training items. To be categorized as a basic-level generalizer, a participant had to respond "yes" to one of the three basic-level questions under the liberal criteria, and "yes" to all three under the strict criteria.

An additional twenty-eight participants were excluded because they could not be categorized as a subordinate or basic-level generalizer. Of the remaining participants, there was not a reliable effect of sampling on generalization ($\chi^2(1) = 0.89$, $p = .34$; $d = 0.33$ $[-0.21,\ 0.87]$).

Experiment 3

Despite increasing the saliency of the teacher, we did not replicate the original effect in Experiment 2. However, we remained concerned that the teacher was less salient in our version, relative to the original. To address this possibility, we next conducted an exact replication of the original study in the laboratory with a real experimenter.

Methods

Participants. In Experiment 3, we recruited 41 undergraduate participants. Participants received either course credit or payment (US $5.00) for their participation.

Stimuli. The object and word stimuli were identical to Experiments 1 and 2.

Procedure. The procedure in Experiment 3 closely followed the original. The experimenter presented 15 objects from two basic-level categories on two pieces of paper. On each paper, the objects were spatially unstructured. As in the original, the experimenter asked five generalization questions (see Experiment 1) and each participant completed two trials such that they were trained and tested on two different categories. Participants were incentivized in the training phase with a sticker. The exact script used by the experimenter is in Appendix B.

Results and Discussion

We excluded one participant who did not select subordinate-level matches in the training phase. We also excluded nine participants because they could not be categorized as a subordinate or basic-level generalizer. Of the remaining 31 participants ($N_{learner} = 13$; $N_{teacher} = 18$), there was not a reliable effect of sampling condition on generalization ($\chi^2(1) = 0.49$, $p = .48$; $d = 0.45$ $[-0.38,\ 1.27]$). One notable difference in this experiment, however, was the rate of generalizations to the basic level: In Experiment 4, there were overall more basic level generalizations compared to the online versions. We return to this difference in the General Discussion.

Experiment 4

Experiment 3 suggests that the decreased effect sizes in the online paradigm were not due the decreased saliency of the experimenter. In this final replication, we explored another possible difference between our experiments and the original: the object stimuli. While similar to the original, our objects had slightly more variability within each subordinate level than the original. This increased variability may lead participants to be less likely to generalize to the basic level. Thus, in Experiment 4, we conducted an online replication using the same procedure as Experiment 2, but with less variable objects at the subordinate-level.

Methods

Participants. We recruited 200 participants from Amazon Mechanical Turk. Participants were paid US $0.30 for their participation.

Stimuli. The objects contained less subordinate-level variability than that used in Experiment 1-3, and were highly similar to the original (Figure 1, right).

Procedure. The procedure was identical to Experiment 2.

Results and Discussion

As in the previous experiments, we excluded participants who responded “yes” to the basic non-match question ($N = 17$) or who did not select subordinate matches in the learning condition ($N = 27$). We also excluded participants who missed an attention check question ($N = 10$). Our final sample included 161 participants ($N_{learner} = 69$; $N_{teacher} = 92$).

An additional twenty-one participants were excluded from our analyses because they could not be categorized as a subordinate or basic-level generalizer. Of the remaining sample, there was a reliable effect of sampling on generalization: Participants in the teacher condition were more likely to generalize to the subordinate category, while participants in the learner condition were more likely to generalize to the basic-level ($\chi^2(1) = 6.42$, $p = .01$; $d = 0.59$ $[.16,\ 1.02]$). This replicates the pattern seen in the original Xu and Tenenbaum (2007b) report, though with a much smaller effect size.

General Discussion

Across four replication attempts, we replicate the original result in Experiment 4. While we successfully replicate the effect, what is most notable about our results is that across all four replication attempts, the effect sizes is much less robust than in the original report.

- things that might matter: * variability visually * spatial structure * features of the teacher (reliable?) * cost of getting it wrong (”good enough processing” - in communication vs. turk)

- in the learner condition, only in lab shows same pattern as original in sub vs. basic - would be nice to have paradigm where sampling didn’t come from learner in learning context (other random way)

TO DO: - write GD - anonymize data - add links - write abstract

Appendix A

	N excluded	χ^2	*p*	*d*
Exp. 1	94	0.08	.78	0.09 [−0.30, 0.47]
Exp. 2	36	0.07	.79	0.19 [−0.47, 0.85]
Exp. 3	14	0.63	.43	0.53 [−.035, 1.42]
Exp. 4	31	4.13	.04	0.54 [0.06, 1.03]

Table 1
Results of all four experiments using the strict categorization criteria, described in Experiment 1. “N excluded” refers to the number of participants excluded from analyses because they could not be categorized as either basic or subordinate-level generalizers. All χ^2 tests have 1 degree of freedom.

Appendix B

Below is the script used by the experimenter in Experiment 3. “[word]” denotes a randomly selected novel label. Different labels were used in Trial 1 and Trial 2.

Thank you for participating in this study. We're going to play a game that was initially designed for preschoolers, so it may seem a little silly, but just play along. Are you ready to begin?

Training

Teacher Condition: *See this? It's a [word]. See this one? It's a [word]. See this one? It's a [word]. Thank you for paying attention. Would you like to choose a sticker?*

Learner Condition: *See this? It's a [word]. Can you point to two other [word]s? If you get both of them right you get a sticker! You're correct! You're correct! Would you like to choose a sticker?*

Testing

Alright, now I'm going to ask you some questions about [word]s. Are you ready?

⟨point to subordinate⟩ *Is this a [word]?*

⟨point to basic non-match ⟩ *Is this a [word]?*

⟨point to basic match⟩ *Is this a [word]?*

⟨point to subordinate⟩ *Is this a [word]?*

⟨point to basic match⟩ *Is this a [word]?*

Alright, now I'm going to show you some new shapes. Are you ready?

⟨repeat Training and Testing for Trial 2⟩

All done! Thank you for participating!

References

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition. Hillsdale, NJ: Erlbaum*.

Clark, E. (1988). On the logic of contrast. *Journal of Child Language*, *15*(02), 317–335.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243–259.

Jenkins, G. W., Samuelson, L. K., Smith, J. R., & Spencer, J. P. (2015). Non-Bayesian noun generalization in 3-to 5-year-old children: Probing the role of prior knowledge in the suspicious coincidence effect. *Cognitive Science*, *39*(2), 268–306.

Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448.

Spencer, J. P., Perone, S., Smith, L., & Samuelson, L. K. (2011). Learning words in space and time: Probing the mechanisms behind the suspicious-coincidence effect. *Psychologial Science*, *22*(8), 1049-1057.

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

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288-297.