Psychological Science
2018, Vol. 29(12) 2039–2047
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797618794931
www.psychologicalscience.org/PS
SAGE

Still Suspicious: The Suspicious-Coincidence Effect Revisited

Molly L. Lewis[1,2] and Michael C. Frank[3]
[1]Computation Institute, University of Chicago; [2]Department of Psychology, University of Wisconsin–Madison; and [3]Department of Psychology, Stanford University

Abstract

Imagine hearing someone call a particular dalmatian a "dax." The meaning of the novel noun *dax* is ambiguous between the subordinate meaning (dalmatian) and the basic-level meaning (dog). Yet both children and adults successfully learn noun meanings at the intended level of abstraction from similar evidence. Xu and Tenenbaum (2007a) provided an explanation for this apparent puzzle: Learners assume that examples are sampled from the true underlying category (strong sampling), making cases in which there are more observed exemplars more consistent with a subordinate meaning than cases in which there are fewer exemplars (the suspicious-coincidence effect). Authors of more recent work (Spencer, Perone, Smith, & Samuelson, 2011) have questioned the relevance of this finding, however, arguing that the effect occurs only when the examples are presented to the learner simultaneously. Across a series of 12 experiments (N = 600), we systematically manipulated several experimental parameters that varied across previous studies, and we successfully replicated the findings of both sets of authors. Taken together, our data suggest that the suspicious-coincidence effect in fact is robust to presentation timing of examples but is sensitive to another factor that varied in the Spencer et al. (2011) experiments, namely, trial order. Our work highlights the influence of pragmatics on behavior in experimental tasks.

Keywords

word learning, Bayesian inference, meta-analysis, concepts, open data, open materials, preregistered

Received 9/29/17; Revision accepted 6/7/18

Suppose you are learning a new language and someone tells you that a particular kind of chili pepper is called a "cabai." Does cabai mean chili pepper, pepper, or vegetable? The same object can be referred to by many different labels depending on the level of abstraction—subordinate (chili), basic level (pepper), or superordinate (vegetable)—that the speaker wishes to convey. In principle, this ambiguity could pose a challenge for language learners: Even though cabai means chili, in nearly every individual case in which *chili* can be used, the speaker could also have been saying *pepper*. Yet despite the apparent difficulty of the learning problem, children are able to quickly and successfully learn the meanings of words at multiple levels of abstraction (Markman, 1990; Waxman & Hatch, 1992; Waxman, Shipley, & Shepperson, 1991).

Like adults, young children have a bias to both interpret and use words at the basic level of abstraction (e.g., Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Waxman, 1990). A body of experimental work has examined how children might overcome this basic-level bias to learn words at different levels of the conceptual hierarchy (Waxman, 1990; Waxman & Hatch, 1992; Waxman et al., 1991). For example, Waxman (1990) presented children with three category exemplars from the same level of abstraction (e.g., a collie, a terrier, and a setter) and asked children to generalize to new exemplars of that category. The results suggest that labeling the category with a novel word helped children to correctly generalize to new category members, but only when the exemplars were superordinate or basic-level matches; when the exemplars were subordinate matches, the presence of a novel label decreased

Corresponding Author:
Molly L. Lewis, University of Chicago, Computation Institute, 5735 S. Ellis Ave., Chicago, IL 60637
E-mail: mollyllewis@gmail.com

accuracy in generalizations, suggesting that subordinate generalizations are particularly difficult for children to learn.

Xu and Tenenbaum (2007a) provided an account of how learners might make appropriate generalizations in word learning, particularly at the subordinate level. They observed that if *cabai* meant pepper, it would be quite odd for a learner to see several independent examples of a cabai that all happened to be chili peppers. Why not a bell pepper? This suspicious coincidence might provide evidence that the meaning of cabai instead was the narrower subordinate meaning, chili. Formally, this observation emerges from *strong sampling* (Tenenbaum & Griffiths, 2001), the idea that examples of cabai are sampled from within the extension of the corresponding concept. So if the word means pepper, the likelihood of observing a chili pepper three times in a row is low, whereas if the word means chili, the corresponding likelihood is higher.

One prediction of this model of generalization is that observing more word–object pairs should make a learner more likely to generalize to the subordinate level, as opposed to the basic level. Using a paradigm similar to that used by Waxman (1990), Xu and Tenenbaum directly tested this prediction by providing adults and children with novel words paired with exemplars at the subordinate level and found that both groups' generalizations narrowed when they observed three exemplars compared with when they observed only one. This finding was supported by another set of experiments that suggested that such narrowing was observed only when examples were chosen by an informative teacher (Lewis & Frank, 2016; Xu & Tenenbaum, 2007b).

These findings have been an important part of a reevaluation of children's ability to make complex inferences from sparse data, provided the data are produced by an informative sampling process (e.g., strong sampling; Shafto, Goodman, & Frank, 2012). Children make inferences about ambiguous references on the basis of the idea that referential descriptions are produced via strong sampling (Frank & Goodman, 2014; Horowitz & Frank, 2016). Subsequent work has found that toddlers' nonlinguistic generalization is also consistent with sensitivity to sampling (Gweon, Tenenbaum, & Schulz, 2010; Xu & Denison, 2009). And strong sampling has been used to justify the narrowed generalizations made by preschoolers in pedagogical contexts (Bonawitz et al., 2011).

The empirical support for the role of strong sampling in Xu and Tenenbaum's paradigm has been questioned, however. In a follow-up to Xu and Tenenbaum's study, Spencer, Perone, Smith, and Samuelson (2011) offered an alternate explanation for the suspicious-coincidence effect. They argued that the effect can be accounted for by basic memory and perceptual processes in which the co-occurrence of objects in time and space leads to direct comparison, which highlights similarities and differences across exemplars (see, e.g., Gentner & Namy, 2006). This highlighting in turn should lead to better memory for the specific shared features of the target category and to a narrower generalization at test. Specifically, they predicted that better memory for specific shared features should make it more likely for participants to generalize to the subordinate level when multiple subordinate category exemplars are presented simultaneously—precisely the suspicious-coincidence pattern observed by Xu and Tenenbaum.

Spencer et al. tested this possibility by replicating the original Xu and Tenenbaum experiments with slightly different design parameters. Motivated by their theoretical claim, they presented the learning exemplars sequentially, rather than simultaneously, so only one learning exemplar was visible at a time. The sequential presentation of objects, they argued, more closely reflects the experience of learners in the real world who encounter word–object pairings at distinct points in time and space. In a series of experiments, Spencer et al. replicated Xu and Tenenbaum's main finding—more basic-level generalizations with one exemplar than with three exemplars—with simultaneous presentation but failed to replicate it with sequential presentation. In fact, they observed a reversal under sequential-presentation conditions: Participants were more likely to generalize to the basic level when three subordinate exemplars were presented.

Spencer et al.'s findings are important because they call into question one major piece of evidence for the idea that children and adults are sensitive to sampling processes. At the same time, they are also surprising because other authors have suggested that simultaneous presentation highlights exemplar commonalities and increases memory consolidation (Lawson, 2014, 2017). In addition, a closer examination of Spencer et al.'s design reveals a number of procedural differences from Xu and Tenenbaum's study, which—although seemingly minor—might have led to the diverging findings reported by Spencer et al. and Xu and Tenenbaum.

In light of the importance of the suspicious-coincidence effect and the complexity of the empirical picture, our goal in the current work was to replicate the suspicious-coincidence effect. Rather than choosing to follow up exclusively on either Spencer et al. or Xu and Tenenbaum, we chose to explore the space of design decisions that connect them, effectively replicating both paradigms as well as a number of unexplored design variants (cf. Baribault et al., 2018). By exploring the space of possible procedures more fully, we were then able to make strong inferences about the procedural factors responsible for the magnitude of the suspicious-coincidence effect.

In this article, we report 12 experiments (10 preregistered) that varied four procedural elements: presentation timing (simultaneous vs. sequential), trial order, blocking of trials, and consistency of labels across trials. We reproduced the suspicious-coincidence effect with a large effect size in both the sequential- and simultaneous-presentation conditions, except under a particular trial order: when the three-exemplar trials were presented before the one-exemplar trials. When the three-exemplar trials were presented first, we saw a high level of subordinate generalizations even for the one-exemplar trial. We attribute this difference to the fact that when the three-exemplar trials were presented first, participants were aware of these previous exemplars when they observed the single exemplar and consequently did not interpret it as the only observed exemplar from the target category. In sum, although we replicated the Spencer et al. study exactly, our full set of experiments led us to a different interpretation of the data. We concluded that the suspicious-coincidence effect is robust to sequential presentation. The effect is sensitive to some features of the general experimental context, however, suggesting a potential interpretation in terms of the pragmatics of the task.

Method

We report how we determined our sample size, manipulations, and measures in the study. For all experiments, the experimental code, stimuli, and analysis code are publicly available via the Open Science Framework (OSF). The experimental code, analysis code, stimuli, and sample size were all preregistered, with the exception of those for Experiments 8 and 12, and can also be found at the OSF (osf.io/zcbp7).

Participants

Fifty participants were recruited on Amazon Mechanical Turk for each of our 12 experiments ($N = 600$) and paid 40 to 50 cents each for their participation. Across all 12 experiments, 13% of participants completed more than one experiment. We report data from all participants in the main text, but the pattern of reported findings held when these participants were excluded (see the Supplemental Material available online).

We determined our sample size on the basis of a preregistered power calculation using a meta-analytic estimate of the effect size from the studies conducted by Xu and Tenenbaum and Spencer et al. The chosen sample size was approximately twice the estimated sample size necessary to obtain a power of .99 (for details, see the Supplemental Material).

Stimuli

Our picture stimuli were gathered on the Internet and closely resembled that of Xu and Tenenbaum and Spencer et al. The referent objects were three sets of 15 pictures from different basic-level categories (vegetables, vehicles, and animals). Within each category, 5 pictures were subordinate exemplars (e.g., green peppers), 4 were basic-level exemplars (e.g., peppers), and 6 were superordinate exemplars (e.g., vegetables; see Fig. 1). The exemplars were divided into learning and generalization sets. For each category, the learning set consisted of 3 subordinate, 2 basic, and 2 superordinate pictures presented in different combinations on different trials (see the Procedure section). The generalization set for each category consisted of the remaining 8 pictures. The learning and generalization sets were the same for all participants. The linguistic stimuli were 12 one-syllable novel labels (e.g., "wug").

Procedure

Participants were first introduced to a picture of a character ("Mr. Frog") and instructions describing the task. They were told that the character speaks a different language, and their job was to help the character find the toys he wants. Participants then advanced to the main task, which consisted of a series of 12 trials on separate screens. On each trial, one or three learning exemplars from one of the three stimulus categories appeared at the top of the screen, along with the following instructions: "Here [is a wug/are three wugs]. Can you give Mr. Frog all of the other wugs?" Below the learning exemplars, 24 generalization exemplars (8 from each of the three categories) were displayed in a 4 × 6 grid. The generalization pictures were displayed in random order across trials. Participants were instructed to select the target category members ("To give a wug, click on it below. When you have given all the wugs, click the Next button."). When an exemplar was selected, a red box appeared around the picture, and participants were allowed to change their selections by clicking on the picture a second time. The learning exemplars remained visible at the top of the screen during the generalization task. After they had made their selections, participants advanced to the next trial by clicking the "Next" button.

There were four trial types, distinguished by the number and conceptual level of the learning exemplars: one subordinate exemplar, three subordinate exemplars, three basic exemplars, and three superordinate exemplars. Each participant completed each trial type for each of the three stimulus categories (vegetables, vehicles, and animals).

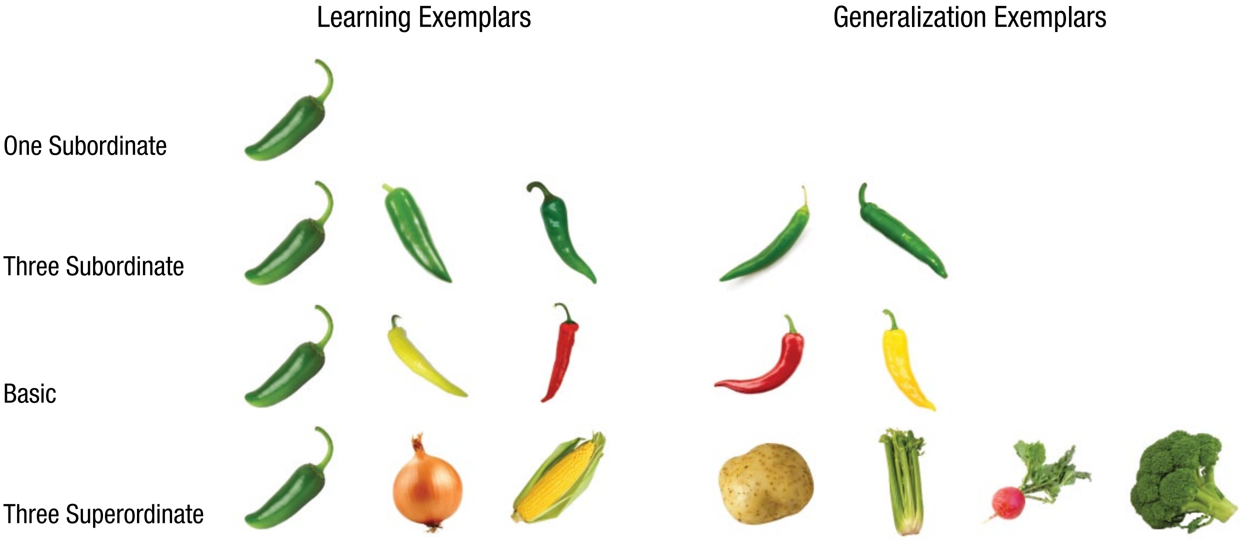

Fig. 1. Learning and generalization exemplars from the subordinate, basic, and superordinate conceptual levels of the vegetable category. On a given trial, participants saw one or three exemplars of the same level from the learning set, followed by all exemplars from the generalization set (along with the generalization sets from the other categories).

Across 12 experiments, we manipulated four aspects of the trial design that differed between the designs of Xu and Tenenbaum and Spencer et al. (summarized in Table 1; all experiments can be viewed in the Supplemental Material): presentation timing (*simultaneous* vs. *sequential*), trial order (*one–three* vs. *three–one*), label (*same* vs. *different*), and blocking (*blocked* vs. *pseudorandom*). Our set of experiments does not include all possible combinations of these design factors, but all levels were tested in at least one experiment. We describe each of these factors in more detail below.

Presentation timing. Presentation timing was the key, theoretically motivated experimental design difference between experiments by Xu and Tenenbaum (Experiments 1 and 2)[1] and Spencer et al. (Experiments 2 and 3). In the Xu and Tenenbaum study, the learning exemplars were presented statically and simultaneously, whereas in the key conditions from Spencer et al., participants saw a sequence of individual exemplars, with each exemplar visible for only 1 s at a time. In the sequential design, three-exemplar learning trials displayed pictures at three different locations (left, middle, and right) in a sequence that repeated twice, for a total of 6 s.

We reproduced these design aspects in the simultaneous and sequential versions of our experiments. In the one-exemplar sequential trials, the exemplar appeared (1 s) and disappeared (1 s) for three repetitions.[2] The generalization pictures did not appear in the sequential condition until after the training pictures had appeared for 6 s, but they remained visible as participants selected generalization exemplars.

Trial order. In Xu and Tenenbaum's Experiment 1, the three one-subordinate trials occurred first, followed by all other trial types (one–three; Xu and Tenenbaum's Experiment 2 used a between-subjects design). In contrast, in the main experiments in Spencer et al. (Experiments 2 and 3), the three-subordinate trials occurred first (three–one). Spencer et al.'s replication of Xu and Tenenbaum's simultaneous design (Spencer et al.'s Experiment 1) showed a single block of either one-subordinate or three-subordinate trials first (in random order). In Supplemental Experiments 1 and 2, Spencer et al. directly explored whether trial order influenced the effect size by replicating their Experiment 1 with three-subordinate trials followed by one-subordinate trials.

Labels. Xu and Tenenbaum used the same label for each category for the three-subordinate and one-subordinate trials (e.g., both the one-pepper and the three-pepper trials would be called *wugs*; same). In contrast, Spencer et al. used a different novel label on each of the 12 trials, such that the three-subordinate and one-subordinate trials were referred to with distinct labels (different). We reproduced these two design choices and also randomly mapped labels to categories across trials.

Blocking. The studies also differed in whether the trials were blocked by trial type: In the Xu and Tenenbaum study, the first three trials were a block of one-subordinate trials and the remaining nine were at random (*pseudorandom*), whereas Spencer et al. blocked all four trial types in all experiments (*blocked*). We also reproduced these two design variants, randomizing the order of the trials within each block for the blocked design.

Table 1. Manipulated Variables and Effect Sizes for Our 12 Experiments

		Manipulation					
Experiment	*N*	Presentation timing	Order	Blocking	Label	Cohen's *d*	Original experiment
1	50	Simultaneous	One–three	Pseudorandom	Same	1.27 [0.84, 1.71]	Xu & Tenenbaum (2007a) Experiments 1 and 2
2	50	Simultaneous	One–three	Pseudorandom	Same	1.2 [0.77, 1.64]	Xu & Tenenbaum (2007a) Experiments 1 and 2
3	50	Simultaneous	One–three	Pseudorandom	Different	1.1 [0.67, 1.52]	
4	50	Simultaneous	Three–one	Blocked	Different	0.02 [−0.37, 0.42]	Spencer, Perone, Smith, & Samuelson (2011) Supplemental Experiments 1 and 2
5	50	Simultaneous	Three–one	Blocked	Different	−0.02 [−0.42, 0.37]	
6	50	Simultaneous	Three–one	Blocked	Same	−0.04 [−0.43, 0.36]	
7	50	Sequential	One–three	Pseudorandom	Same	1.43 [0.99, 1.87]	
8	50	Sequential	One–three	Pseudorandom	Different	1.24 [0.81, 1.67]	
9	50	Sequential	One–three	Blocked	Different	1.27 [0.84, 1.71]	
10	50	Sequential	Three–one	Blocked	Different	−0.43 [−0.83, −0.02]	Spencer, Perone, Smith, & Samuelson (2011) Experiments 2 and 3
11	50	Sequential	Three–one	Pseudorandom	Same	−0.3 [−0.7, 0.1]	
12	50	Sequential	Three–one	Blocked	Same	−0.18 [−0.58, 0.21]	

Note: Order refers to the relative ordering of one- and three-subordinate trials. Blocking refers to whether trials were blocked by category or pseudorandomly. Label indicates whether the labels in one- and three-subordinate trials were the same or different. For Cohen's *d*s, 95% confidence intervals are given in brackets.

Data analysis

The key prediction of the suspicious-coincidence effect is that participants should generalize to the basic level more often in one-subordinate trials relative to three-subordinate trials. To measure this effect, for each trial, we calculated the proportion of generalizations to basic exemplars within the same category (out of two) and averaged across categories for each participant. We estimated the difference between the one-subordinate and three-subordinate conditions by calculating an effect size for each experiment (Cohen's *d*; for details, see the Supplemental Material). We then estimated the influence of each of our design manipulations on the overall effect size by fitting a random-effects meta-analytic model with each of our four manipulations as fixed effects. The model included both the present set of experiments and prior experiments by Xu and Tenenbaum and Spencer et al. We used the *metafor* package (Viechtbauer, 2010) in the R programming environment (R Core Team, 2008) to fit our meta-analytic models.

Results

Figure 2 shows the mean proportion of generalizations to the basic level in the one- and three-subordinate trials for all 12 experiments (for means across all measures and conditions, see the Supplemental Material), and Figure 3 shows the corresponding effect sizes (with experiments by Xu and Tenenbaum and Spencer et al. included for reference).

We replicated the suspicious-coincidence effect in two exact replications of Xu and Tenenbaum's method (Experiment 1: d = 1.27, 95% confidence interval, or CI = [0.84, 1.71]; Experiment 2: d = 1.2, 95% CI = [0.77, 1.64]), with a magnitude comparable with that found in Xu and Tenenbaum's original Experiment 1 (d = 2.0, 95% CI = [1.25, 2.74]) and Experiment 2 (d = 1.01, 95% CI = [0.51, 1.51]). We also replicated the reversal in the suspicious-coincidence effect observed by Spencer et al. in an exact replication of their method (Experiment 10: d = −0.43, 95% CI = [−0.83, −0.02]), and with a magnitude comparable with that found in their original experiments (Experiment 2: d = −0.61, 95% CI = [−1.27, 0.04]; Experiment 3: d = −0.3, 95% CI = [−0.96, 0.36]).

Critically, however, the meta-analytic model across all 12 experiments revealed that only trial order was a reliable predictor of effect size ($\beta = -1.44$, $z = -9.27$, $p < .0001$), whereas timing ($\beta = -0.16$, $z = -1.45$, p = .15), blocking ($\beta = -0.1$, $z = -0.56$, p = .58), and label ($\beta = 0.06$, $z = 0.51$, p = .61; see Table 2) were not. These data thus suggest that the suspicious coincidence is robust to spatiotemporal aspects of the presentation

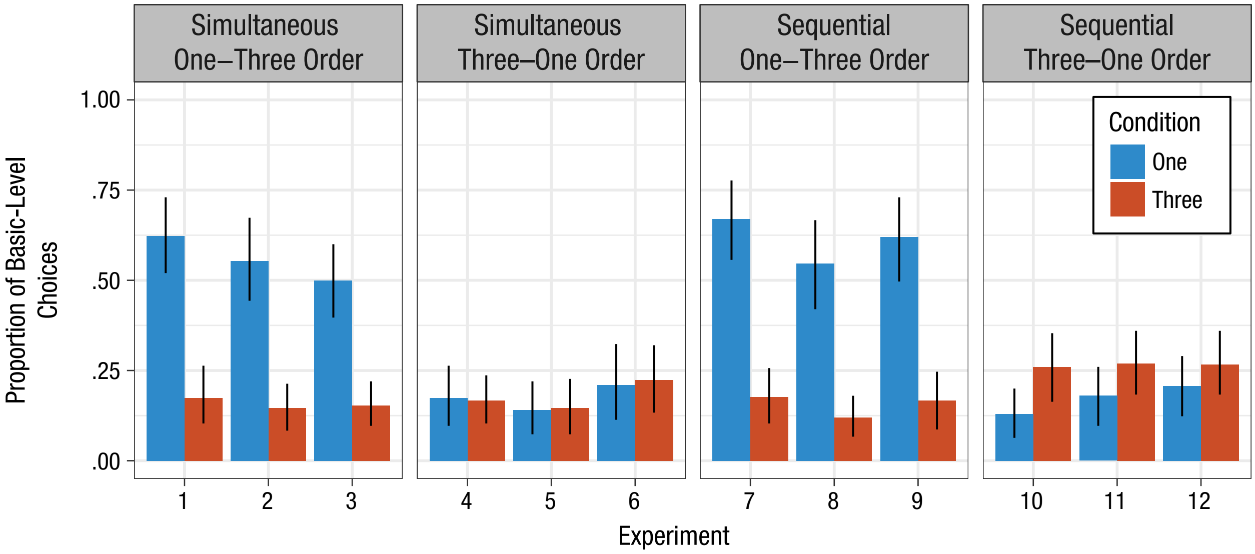

Fig. 2. Mean proportion of generalizations to basic-level exemplars in the one-subordinate-exemplar and three-subordinate-exemplar conditions for all 12 of our experiments. Results are shown separately for each pairing of presentation timing (simultaneous vs. sequential) and trial order (one–three vs. three–one). Error bars indicate bootstrapped 95% confidence intervals.

learning exemplars, in contrast to the conclusion drawn by Spencer et al.

Our data also suggest that the three–one ordering interacts with presentation timing: In experiments with the three–one ordering and sequential presentation (Experiments 10–12), we saw a reversal of the suspicious-coincidence effect, as observed by Spencer et al. To examine this pattern, we fit a second meta-analytic model that included presentation timing and trial order as additive effects and a third term for their interaction. As in the initial model, there was a main effect of trial order ($\beta = -1.18$, $z = -8.04$, $p < .0001$) but not presentation timing ($\beta = 0.1$, $z = 0.6$, $p = .55$). However, there was also a significant interaction between the effects of two design parameters ($\beta = -0.47$, $z = -2.12$, $p = .03$). This interaction effect was due to increased generalizations to the basic level when the three-subordinate trials were presented sequentially (Experiments 10–12) compared with simultaneously (Experiments 4–6). In the General Discussion section, we consider why trial order might influence the suspicious-coincidence effect as well as possible reasons for the interaction with presentation timing.

General Discussion

The suspicious-coincidence effect (Xu & Tenenbaum, 2007a) suggests a powerful mechanism by which learners might overcome the inherent ambiguity associated with learning subordinate word meanings. Other evidence (Spencer et al., 2011), however, suggests that the effect may occur only under particular learning conditions, namely, when the training exemplars are presented simultaneously to the learner. Across 12 studies, we explored the experimental parameter space of the suspicious-coincidence paradigm and successfully replicated the findings from both sets of authors. Taken together, our studies led us to a different conclusion from that reached by Spencer et al.: The suspicious-coincidence effect is robust to the presentation timing of exemplars but is sensitive to order effects. These order effects (where three-exemplar trials are presented before one-exemplar trials) were not predicted by Xu and Tenenbaum. Below, we offer an account of these results based on recent generalizations of strong sampling models to describe pragmatic inferences.

The critical difference between the one–three and three–one ordering was the rate of generalization to the basic level in the one-exemplar trial: When the one-exemplar trial occurred second, participants were less likely to generalize to the basic level compared with when the one-exemplar trial was presented first. Why might this ordering matter? Consider a scenario in which first the learner observes a trial with three subordinate peppers followed by a second trial with only a single pepper. Although the two trials were intended to be interpreted as independent from each other, their co-occurrence in the task may have suggested to participants that they are pragmatically related, leading participants to track their frequency across trials. If true, when the learner observes the single pepper on the second trial, it is effectively the fourth subordinate exemplar from the same category (identical to an exemplar from the three subordinate trials). This account

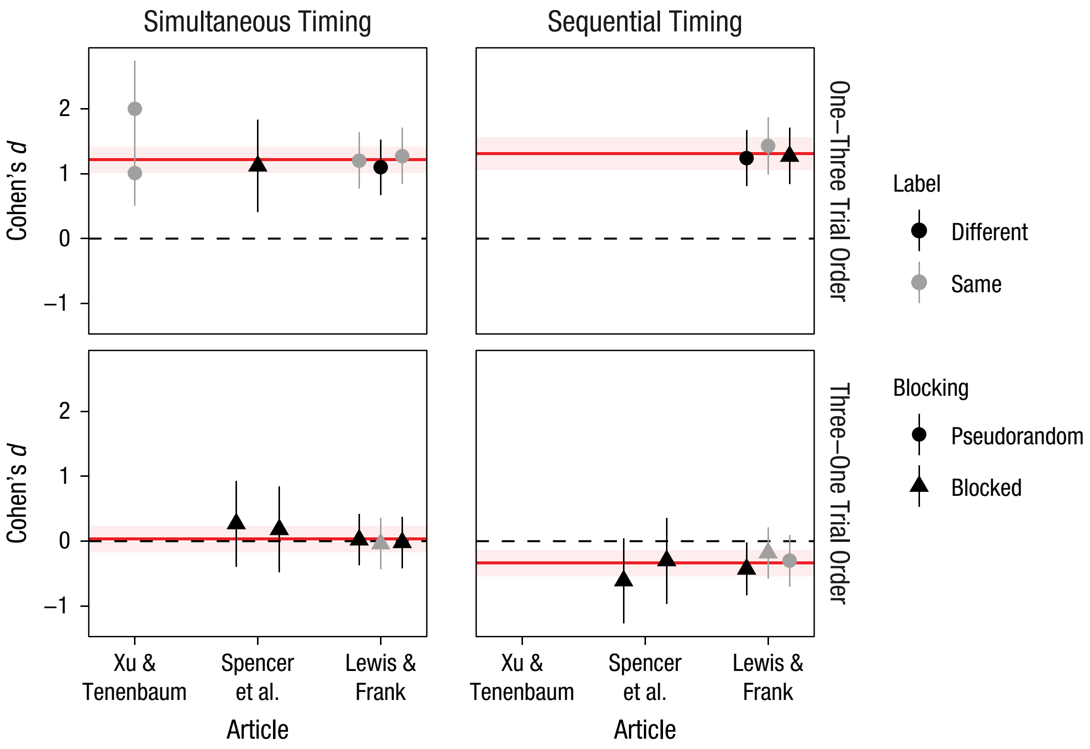

Fig. 3. Effect sizes for all 19 experiments conducted on the suspicious-coincidence effect by Xu and Tenenbaum (2007a), Spencer, Perone, Smith, and Samuelson (2011), and the current authors. Results are shown separately for each pairing of presentation timing (simultaneous vs. sequential) and trial order (one–three vs. three–one). The color of each data point indicates whether the single exemplar and three subordinate exemplars received the same label or a different label. The shape of each data point indicates whether trials were blocked by category or were pseudorandom. The red lines reflect the meta-analytic estimate of the effect size (for the Xu & Tenenbaum experiments, standard deviations of effect sizes were estimated from the Spencer et al. replication). The error bars on the data points, as well as the red band around the estimates of effect sizes, indicate 95% confidence intervals. Points are jittered along the *x*-axis for visibility.

predicts that learners should be less likely to generalize to the basic level when the single exemplar is presented second, consistent with our findings. It also makes a second prediction: In the case of the three–one ordering, learners should be slightly more likely to generalize to the basic level on the first trial (three exemplars) compared with the second trial (single exemplar, fourth observed exemplar), because seeing four exemplars is a bigger suspicious coincidence than three. We found some evidence consistent with this prediction from the meta-analytic model indicating a reversal of the effect under the sequential-timing, three–one ordering conditions.

Although Xu and Tenenbaum's model does not directly predict participants' behavior in the three–one ordering, there is a broader class of Bayesian models, of which Xu and Tenenbaum's model is an instance, that does. These models account for pragmatic reasoning by assuming that speakers reason about the intention of other people when making linguistic inferences (e.g., Frank & Goodman, 2012). In this case, reasoning about the speaker's intention may lead participants to

Table 2. Results of the Meta-Analytic Model With Manipulations as Fixed Effects

Fixed effect	β	*z*	*p*
Intercept	1.36 [1.06, 1.65]	9.02	< .0001
Presentation timing: simultaneous vs. sequential	–0.16 [–0.37, 0.06]	–1.45	.15
Trial order: one–three vs. three–one	–1.44 [–1.75, –1.14]	–9.27	< .0001
Label: different vs. same	0.06 [–0.19, 0.31]	0.51	.61
Blocking: Blocked vs. pseudorandom	–0.1 [–0.44, 0.24]	–0.56	.58

Note: Values in brackets are 95% confidence intervals.

assume discourse continuity across trials. Indeed, there is experimental evidence that children reason about the intention of the speaker to assume discourse continuity when inferring the meaning of a novel word (Horowitz & Frank, 2016). In future work, the pragmatic influence of discourse continuity in this task can be eliminated by using a between-subjects design, as in Xu and Tenenbaum's experiment with children (Experiment 2).

If indeed participants interpret the one-exemplar trial in three–one orders as a fourth exemplar, then it is somewhat surprising that the identity of the label between the two trials does not matter: We saw the same pattern when the labels were different (Experiment 10) as when they were the same (Experiments 11 and 12). Given evidence that children and adults tend to assume that different words have different meanings (Clark, 1987), we might expect that a different label on the one-exemplar trial would lead participants to treat the new exemplar as referring to a new category. However, there are a number of reasons that participants may not have carefully attended to labels across trials. First, participants were never tested on the meaning of labels, and the labels were not directly relevant to completing the generalization task. Second, the three- and one-exemplar trials for the same category rarely occurred adjacent to each other (because the one-exemplar trials were always blocked), and this delay might have made it more difficult for participants to remember the labels across the critical trials. Consistent with this pattern, we found that label identity did not mediate the suspicious-coincidence effect across experiments.

We also found that trial order interacted with presentation timing: When we replicated Spencer et al.'s experiments, sequential presentation in the three–one ordering led to a reversal of the suspicious-coincidence effect. Spencer et al.'s theory predicts the reversal under sequential-presentation conditions, but it does not predict the observed interaction with trial order. There is also not a straightforward explanation from Xu and Tenenbaum's model. We offer one highly speculative account: Sequential-presentation conditions may have appeared relatively more complex to participants compared with simultaneous-presentation conditions, resulting in higher overall uncertainty in the generalization judgment. This increased uncertainty may have led participants to be more likely to generalize conservatively—at the basic level—on the first trial when exemplars were presented sequentially as opposed to simultaneously. Future research could test this cognitive load explanation more directly.

Broadly, our findings highlight the influence of seemingly minor experimental design parameters on the observed pattern of data. In the present studies, experiments with the one–three versus three–one ordering differed by an effect size of 1.42—a sizable difference that is likely to invite an unwarranted theoretical explanation. Experimental-design parameters are especially important in the context of replication. When conducting a replication of an existing finding, small design parameters may influence the magnitude of the effect (Lewis & Frank, 2016) and even its presence (Phillips et al., 2015). This sensitivity requires that replicators reproduce the original design with as much fidelity as possible before concluding that an effect fails to replicate. Only then can the effect be explored and possible confounds and moderators identified.

The work by both Xu and Tenenbaum and Spencer et al. addresses an important puzzle in the psychological sciences: How do learners learn concepts at multiple levels of abstraction? Their work focuses on a simplified version of this puzzle in which the learner must determine the corresponding labels to known concepts. Our findings here support the idea that learners solved this puzzle via probabilistic inferences about the level of abstraction that is most likely given the observed data (the original suspicious-coincidence effect). Importantly, given the assumption that trials are nonindependent, our interpretation is consistent not only with Xu and Tenenbaum's original set of findings but also with the observed trial-order effects. Our data add to the growing body of work suggesting that suspicious-coincidence effects may arise during pragmatic reasoning in language comprehension (Frank & Goodman, 2014; Goodman & Frank, 2016) as well as through nonlinguistic reasoning (Shafto et al., 2012). Such probabilistic reasoning is likely to play a critical role in learners' ability to make efficient inferences on the basis of sparse linguistic data.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

Both authors designed the experiments. M. L. Lewis coded the experimental paradigm and collected and analyzed the data. Both authors interpreted the findings. M. L. Lewis drafted the manuscript, and M. C. Frank provided critical revisions. Both authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618794931

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at osf.io/zcbp7/. The design and analysis plans for all experiments except 8 and 12 were preregistered and can also be found at osf.io/zcbp7/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797618794931. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. The age of participants (adults vs. children) differed in Xu and Tenenbaum's Experiments 1 and 2, but we collapsed across this difference for the present analyses.
2. Our implementation of the sequential design differed slightly from the Spencer et al. design, which did not include a 1-s interval between exemplar presentations.

References

Baribault, B., Donkin, C., Little, D. R., Trueblood, J., Oravecz, Z., van Ravenzwaaij, D., . . . Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences, USA*, *115*, 2607–2612.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*, 322–330.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1–34). Hillsdale, NJ: Erlbaum.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, *15*, 297–301.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*, 818–829.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences, USA*, *107*, 9066–9071.

Horowitz, A. C., & Frank, M. C. (2016). Children's pragmatic inferences as a route for learning about the world. *Child Development*, *87*, 807–819.

Lawson, C. A. (2014). Three-year-olds obey the sample size principle of induction: The influence of evidence presentation and sample size disparity on young children's generalizations. *Journal of Experimental Child Psychology*, *123*, 147–154.

Lawson, C. A. (2017). The influence of task dynamics on inductive generalizations: How sequential and simultaneous presentation of evidence impacts the strength and scope of property projections. *Journal of Cognition and Development*, *18*, 493–513.

Lewis, M. L., & Frank, M. C. (2016). Understanding the effect of social context on learning: A replication of Xu and Tenenbaum (2007b). *Journal of Experimental Psychology: General*, *145*(9), e72–e80.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57–77.

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, *26*, 1353–1367.

R Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*, 341–351.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral & Brain Sciences*, *24*, 629–640.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3). doi:10.18637/jss.v036.i03

Waxman, S. (1990). Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development*, *5*, 123–150.

Waxman, S., & Hatch, T. (1992). Beyond the basics: Preschool children label objects flexibly at multiple hierarchical levels. *Journal of Child Language*, *19*, 153–166.

Waxman, S., Shipley, E. F., & Shepperson, B. (1991). Establishing new subcategories: The role of category labels and existing knowledge. *Child Development*, *62*, 127–138.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*, 97–104.

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

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