Still suspicious: The suspicious coincidence effect revisited

Molly L. Lewis[1,2] & Michael C. Frank[3]

[1] Computation Institute, University of Chicago

[2] Department of Psychology, University of Wisconsin, Madison

[3] Department of Psychology, Stanford University

*To whom correspondence should be addressed. Email: mollylewis@uchicago.edu

Author Note

Correspondence concerning this article should be addressed to Molly L. Lewis. E-mail: mollylewis@uchicago.edu

Abstract

In previous work, Xu and Tenenbaum (2007a) provide evidence that learners are able to infer the subordinate meaning of a word from only positive examples of the category. They argue that learners make this inference by assuming that examples are sampled from the true underlying category ("strong sampling"), entailing that certain data patterns are more consistent with a subordinate meaning than others (the "suspicious coincidence" effect). More recent work (Spencer, Perone, Smith & Samuelson, 2011) questions the relevance of this finding by arguing that the effect only occurs when the examples are presented simultaneously. Across a series of 12 studies, we systematically manipulate several experimental parameters that vary between the previous studies, and sucessfully replicate 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 a confound in previous experiments, trial order.

Keywords: word learning, Bayesian inference, meta-analysis, concepts

Word count: 1649

Still suspicious: The suspicious coincidence effect revisited

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 refered 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 where "chili" can be used, the speaker could also have been saying "pepper." Furthermore, children rarely receive the kind of negative evidence ("this is *not* a cabai") that would help rule out broader interpretations. Yet, despite the apparent difficulty of the learning problem, children quickly and successfully learn words at multiple levels of abstraction (Markman, 1990).

Xu and Tenenbaum (2007a; henceforth XT) provide an account of how children might make appropriate generalizations about word meaning without relying on negative evidence. They observe that, if "cabai" meant pepper, it would be quite odd for a learner to see a number of 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* (J. 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 consequence of this model of generalization is that learners should be sensitive to the number of word-objects pairs they observe when determining a word's meaning. In particular, a learner should be more likely to generalize narrowly to the subordinate level when they observe more word-object pairs. XT tested this prediction by providing adults and children with examples of novel words paired with objects and found that both groups' generalizations narrowed when they observed three examples compared with when they

observed only one. This finding was supported by another concurrent set of experiments with adults and children that suggested that such narrowing was only observed when examples were chosen by an informative teacher (Lewis & Frank, 2016; Xu & Tenenbaum, 2007a).

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

The empirical support for the role of strong sampling in XT's paradigm has been questioned, however. In a follow-up study to XT, Spencer, Perone, Smith, and Samuelson (2011; henceforth SPSS) offered an alternative explanation for the suspicious coincidence effect. They argued that the effect can be accounted for by basic memory processes in which the co-occurence of objects in time and space highlights differences across exemplars, thus leading to increased conceptual discrimination. They predicted that this increased conceptual discrimination should make it more likely for participants to generalize to the subordinate level when more subordinate category exemplars are observed – precisely the suspicious coincidence pattern observed by XT.

SPSS tested this possibility by replicating the original XT experiments with slightly different design parameters. Motivated by their theoretical claim, they presented the learning exemplars sequentially, rather than simultaneously, such that only one learning exemplar was visible at a time. The sequential presentation of objects, the 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, SPSS replicated XT's main finding – more basic level generalization with one exemplar than with three exemplars – with simultaneous presentation, but failed to replicate with sequential presentation. In fact, they observed a reversal of the effect under sequential presentation conditions, such that participants were more likely to generalize to the basic level when three subordinate exemplars were presented.

SPSS's findings are important because they call into question one major piece of evidence for the idea that children are sensitive to sampling processes, an idea that underpins a wide variety of recent research. At the same time, they are also surprising, in part because SPSS's account of how basic memory mechanisms lead to broader generalization is at odds with some other work. While SPSS argue that simultaneous presentation highlights differences across exemplars, others have suggested that this method highlights their commonalities and increases memory consolidation (Lawson, 2014, 2017), thus predicting *broader* generalization in the sequential condition. In addition, a closer examination of SPSS's design reveals a number of procedural differences from XT, which – while seemingly minor – might have led to the distinct pattern of findings reported by SPSS and XT.

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 SPSS *or* XT, 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. By exploring the space of possible procedures more fully we are then able to make strong inferences about the procedural factors responsible for the magnitude of the suspicious cooincidence effect.

In the current paper, we report 12 experiments – 10 pre-registered – that varied four procedural elements: presentation timing (simultaneous vs. sequential), trial order, blocking of trials, and consistency of labels across trials. To preview our results, we recover the suspicious coincidence effect with a large effect size in both sequential and simultaneous presentation conditions. The effect only occurs, however, in experiments where the trial with

one exemplar is presented *before* the key trial with three subordinate-consistent exemplars (the "suspicious coincidence"). We attribute this difference to participants' awareness of the possibility of subordinate generalizations following the three-exemplar trial; in these conditions, we see a high level of subordinate generalizations even for the one-exemplar trial (leading to the absence of a difference between conditions). In sum, although we replicate SPSS exactly, our full set of studies leads us to a different interpretation of the data. We conclude 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.

Methods

We report how we determined our sample size, all manipulations, and all measures in the study. All stimuli, experimental code, sample sizes, and analyses were pre-registered with the exception of Exps. 8 and 12, and all are publically available (https://osf.io/yekhj/).

Participants

Fifty participants were recruited on Amazon Mechanical Turk for each of our 12 experiments (N = 600), and paid 40-50 cents 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 holds when these participants are excluded (see SI).[1]

We determined our sample size on the basis of a pre-registered power calculation using a meta-analytic estimate of the effect size from studies conducted by XT and SPSS. The chosen sample size was approximately twice the estimated sample size necessary to obtain a power of .99.

[1]Supplemental information can be found at https://mlewis.shinyapps.io/xtmem_SI/.

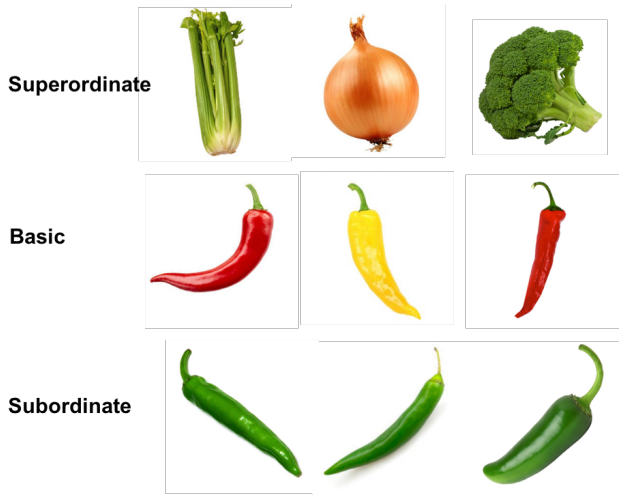

Figure 1. Sample stimuli. Three superordinate (top), basic (middle), and subordinate (bottom) exemplars from the vegetable category.

Stimuli

Our picture stimuli were gathered on the internet, and closely resembeled that of XT and SPSS. The linguistic stimuli were 12 one-syllable novel labels (e.g., "wug"), and the referent objects were three sets of 15 pictures from different basic level categories (vegetables, vehicles and animals). Within each category, five were subordinate exemplars (e.g., green peppers), four were basic level exemplars (e.g., peppers), and six were superordinate exemplars (e.g., vegetables; Fig. 1). The exemplars were divided into a learning and generalization set. For each category, the learning set consisted of 3 subordinate, 2 basic, and 2 superordinate pictures presented in different combinations on different trials (see Procedure). The generalization set for each category consisted of the remaining 8 pictures. The learning and generalization sets were the same for all participants.

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 experiment, which consisted of a series of 12 trials on separate screens.

Table 1
Summary of our 12 experiments.

		Manipulations					
Exp.	N	Timing	Order	Blocking	Label	Effect Size	Original Exp.
1	50	simult.	1-3	pseudo-random	same	1.32 [1.24, 1.4]	XT E1/E2
2	50	simult.	1-3	pseudo-random	same	1.14 [1.06, 1.22]	XT E1/E2
3	50	simult.	1-3	pseudo-random	diff.	1.16 [1.08, 1.24]	
4	50	simult.	3-1	blocked	diff.	0.02 [-0.06, 0.1]	SPSS ES1/ES2
5	50	simult.	3-1	blocked	diff.	-0.06 [-0.14, 0.02]	
6	50	simult.	3-1	blocked	same	-0.14 [-0.22, -0.06]	
7	50	seq.	1-3	pseudo-random	same	1.42 [1.32, 1.52]	
8	50	seq.	1-3	pseudo-random	diff.	1.26 [1.18, 1.34]	
9	50	seq.	1-3	blocked	diff.	1.31 [1.23, 1.39]	
10	50	seq.	3-1	blocked	diff.	-0.44 [-0.52, -0.36]	SPSS E2/E3
11	50	seq.	3-1	pseudo-random	same	-0.31 [-0.39, -0.23]	
12	50	seq.	3-1	blocked	same	-0.17 [-0.25, -0.09]	

[1] N = sample size; Timing = presentation timing (sequential or simultaneous); Order = relative ordering of 1 and 3 subordinate trials; Blocking = trials blocked by category or pseudo-random; Label = same or different label in 1 and 3 trials; Effect size = Cohen's d [95% CI]; Original Exp. = corresponding experiment from prior literature (XT = Xu & Tenenbaum (2007a); SPSS = Spencer, et al. (2011); E = Main Experiment; ES = Supplemental Experiment).

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 3 categories) were displayed in a $4x6$ grid. The order of generalizaiton pictures was randomized 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. Once 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 semantic 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).

Across 12 experiments, we manipulated four aspects of the trial design that differed between XT and SPSS (summarized in Table 1): Presentation timing (simultaneous vs. sequential), trial order (1-3 vs. 3-1), label (same vs. different), and blocking (blocked vs. pseudo-random).[2] 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 XT (E1 and E2)[3] and SPSS (E2 and E3). In XT, the learning exemplars were presented statically and simultaneously, while in SPSS, participants saw a sequence of individual exemplars with each exemplar visible only for 1s 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 6s.

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

Trial order. In XT, the three one-subordinate trials occured first followed by all other trial types ("1-3"). In contrast, in SPSS (E2 and E3), the three-subordinate trials occured first ("3-1"). SPSS's replication of XT's simultanous design (SPSS E1) showed a single block of either one-subordinate or three-subordinate first (randomized).

[2] All experiments can be viewed directly in the SI.

[3] XT E1 and E2 differed in the age of participants (adults vs. children), but we collapse across this difference for the present analyses.

Labels. XT used the same label for each category for the three-subordinate and one-suborinate trials (e.g., both the single pepper and the three-pepper trials would be called "wug"; "same"). In contrast, SPSS used a different novel label on each of the 12 trials, such that the three-subordinate and one-subordinate trials were refered to with distinct labels ("different"). We reproduced these two design choices, and also randomized the mapping of labels to categories across trials.

Blocking. The studies also differed in whether the trials were blocked by trial type: In XT, the first three trials were a block of one-subordinate trials and the remaining 9 ("pseudo-random"), whereas SPSS blocked all four trial types in all experiments ("blocked"). We also reproduced these two design variants, while randomizing trial order within each block for the blocked design.

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 generalizations to subordinate exemplars within the same category (out of 2) and basic exemplars within the same category (out of 2), and averaged across categories for each participant. We estimated the difference between the one-subordinate and three-subordinate conditions by calculating an effect size (Cohen's d) for each experiment. We then estimated the influence of each our design manipulations on the overall effect size by fitting a random-effect meta-analytic model with each of our four manipulations as fixed effects. We used the metafor package (Viechtbauer, 2010) in R to fit our meta-analytic models.

Results

Figure 1 shows the mean proportion generalizations to the basic level in the one- and three-subordinate trials for all 12 experiments,[4] and Figure 2 shows the corresponding effect

[4]See SI for means across all measures and conditions.

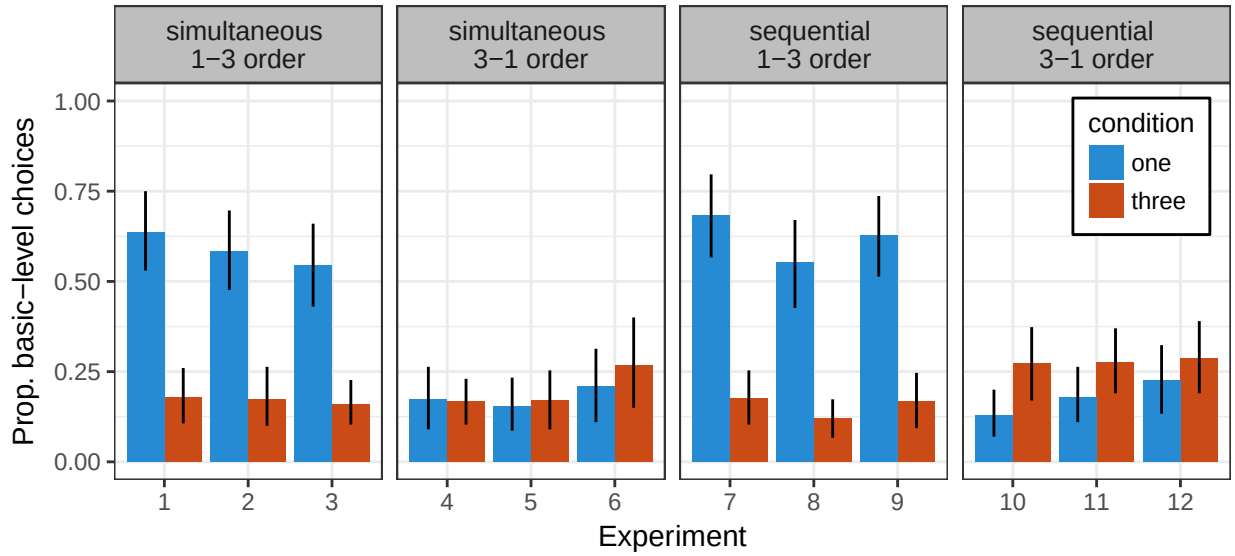

Figure 2. Mean proportion generalizations to basic level exemplars in the one (blue) and three (red) subordinate exemplar conditions for all 12 of our experiments. Each facet corresponds to a pairing of presentation timing (simultaneous vs. sequential) and trial order (1-3 vs. 3-1). Ranges are bootstrapped 95% confidence intervals.

sizes (with XT and SPSS experiments included for reference).

In two exact replications of the XT method, we replicate the suspicious coincidence effect (Exp. 1: d = 1.32 [1.24, 1.4]; Exp. 2: d = 1.14 [1.06, 1.22]), with a magnitude comparable to the original XT experiments (XT E1: d = 2 [1.73, 2.27]; XT E2: d = 1.01 [0.89, 1.13]). We also replicate the reversal in the suspicious coincidence effect observed by SPSS in an exact replication of their method (Exp. 10; d = -0.44 [-0.52, -0.36]), and with a magnitude comparable to the original experiments (SPSS E2: d = -0.61 [-0.81, -0.41]; SPSS E3: d = -0.3 [-0.52, -0.08]).

Critically, however, the meta-analyic model across all 12 experiments reveals that only trial order is a reliable predictor of effect size (β = -1.48; Z = -9.9; p <.0001), while timing (β = -0.13; Z = -1.18; p = 0.24), blocking (β = -0.09; Z = -0.52; p = 0.6), and label are not (β = 0.03; Z = 0.2; p = 0.84; Table 2). These data thus reveal that the suspicious coincidence is robust to spatio-temporal aspects of the presentation learning exemplars, in contrast to the conclusion drawn by SPSS. In the General Discussion, we consider why trial order might influence the suspicious coincidence effect.

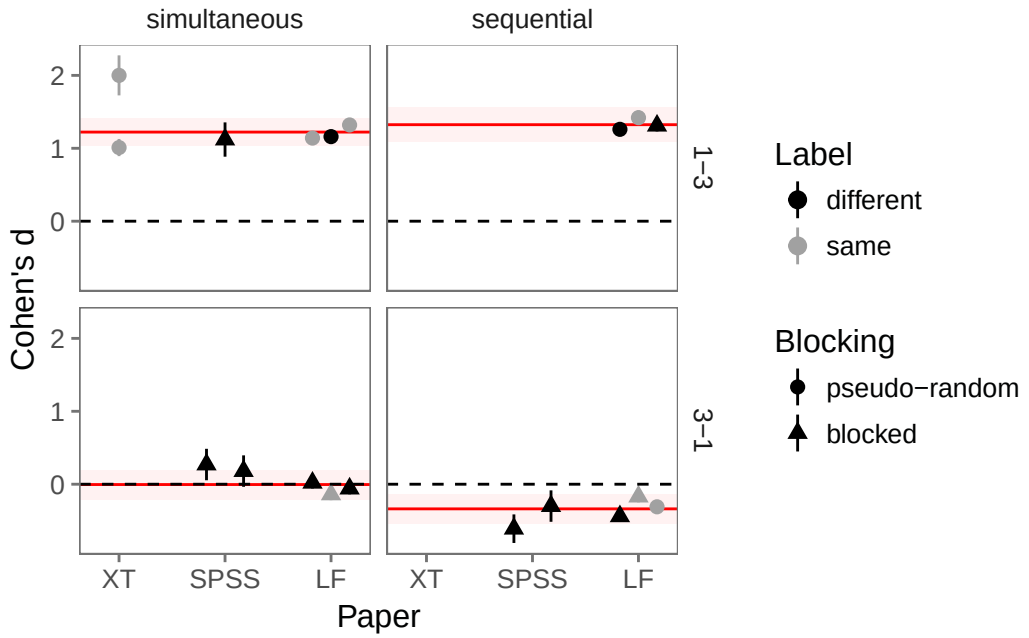

Figure 3. Effect sizes for all 19 studies conducted on the suspicious coincidence effect by XT (Xu & Tenenbaum, 2007a), SPSS (Spencer, et al, 2011), and the current authors. The top-bottom facets indicate whether the single exemplar trial occurred first (1-3) or second (3-1). The left-right facets indicate whether the exemplars were presented simulateously as in XT or sequentially as in SPSS. Point color indicates whether the single exemplar and three subordinate exemplars received the same (grey) or different (black) label. Point shape indicates whether trials were blocked by category (circle) or pseudo-random (triangle). Points are jittered along the x-axis for visibility. The red line reflects the meta-analytic estimate of the effect size (for the XT experiments, standard deviations on effect sizes are estimated from the SPSS replication). All ranges are 95% confidence intervals.

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, Perone, Smith, & Samuelson, 2011), however, suggests that the effect may occur only under ecologically invalid learning conditions—namely, when the training exemplars are presented simultaneously to the learner. Across 12 studies, we explore the experimental parameter space of the suspicious coincidence paradigm and successfully replicate the findings from both sets of authors. Taken together, though, our studies lead us to a different conclusion than SPSS: We find that the suspicious coincidence effect is robust to the presentation timing of exemplars, but is sensitive to order

Table 2
Meta-analytic model with manipulations as fixed effects.

Fixed effect	beta	z-value	p-value
Intercept	1.37 [1.09, 1.65]	9.48	<.0001
Simultaneous vs. sequential timing	-0.13 [-0.33, 0.08]	-1.18	0.24
1-3 vs. 3-1 trial order	-1.48 [-1.77, -1.18]	-9.90	<.0001
Different vs. same label	0.03 [-0.21, 0.26]	0.20	0.84
Blocked vs. pseudo-random trial structure	-0.09 [-0.41, 0.24]	-0.52	0.6

effects. Specifically, we only observe the suspicious coincidence effect when the single exemplar trials are presented before the three-exemplar trials.

The critical difference between the 1-3 and 3-1 ordering was the rate of generalization to the basic level in the one exemplar trial: When the single exemplar trial occurred second, participants generalized to the basic level at a much lower rate than when the single exemplar trial occurred first. Why might this ordering matter? While our data are not able to directly speak to this question, we speculate that this difference may be due to the possibility that the pragmatics of the task lead learners to track exemplar frequency across trials. In other words, when the single exemplar trial occurs second, learners may have interpretted this as the *fourth* exemplar from the same subordinate category in a row, rather than a single exemplar. This hypothesis predicts that participants should be less likely to generalize to the basic level on "single" exemplar trials compared to three-subordinate trials under the 3-1 ordering, thus leading to a reversal of the suspicious coincidence effect. We find some evidence to suggest such a reversal in an analysis of generalizations to the basic level across all experiments with the 3-1 ordering ($t(581)$ = -2.23; p = 0.03; d = -0.18 [-0.2, -0.16]).

Unfortunately, the trial order effect we observe is not theoretically relevant to our understanding of the suspicious coincidence effect, and sampling effects more broadly. This finding does, however, highlight the influence of seemingly minor experimental design parameters on the observed pattern of data. In the present set of experiments, experiments with the 1-3 versus 3-1 ordering differed by an effect size of 1.45—a sizeable difference that is

likely to invite an unwarrented theoretical explanation. Our results demonstrate the importance of verifying that an effect is robust to theoretically-irrelevant design decisions before positing theoretically-rich explanations of an observed effect.

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 decisions 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.

Our studies demonstrate that the suspicious coincidence effect is robust to a range of experimental parameters, and adds to a growing body of work suggesting that sampling plays a critical role in learners' ability to make efficient inferences on the basis of sparse data.

References

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*(3), 322–330.

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

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*, *107*(20), 9066–9071.

Horowitz, A. C., & Frank, M. C. (2016). Children's pragmatic inferences as a route for learning about the world. *Child Development*, *87*(3), 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 impact the strength and scope of property projections. *Journal of Cognition and Development*.

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*(1), 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*(9), 1353–1367.

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*(4), 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*(8), 1049–1057.

Tenenbaum, J., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*(1), 97–104.

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

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