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

Author Note

Correspondence concerning this article should be addressed to Molly L. Lewis, xxx.

E-mail: mollylewis@uchicago.edu

Abstract

Imagine hearing someone call a particular dalmatian "a dax." The meaning of the novel noun "dax" is ambiguous between the subordinate meaning (dalmation) 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 where there are more observed exemplars more consistent with a subordinate meaning than cases where there are fewer exemplars (the "suspicious coincidence" effect). More recent work (Spencer, Perone, Smith & Samuelson, 2011) questions the relevance of this finding, however, arguing that the effect only occurs when the examples are presented to the learner simultaneously. Across a series of 12 studies, we systematically manipulate several experimental parameters that vary across previous studies, and successfully 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 another factor that varied in the Spencer, Perone, Smith, & Samuelson (2011) experiments, namely, trial order. Our work highlights the influence of pragmatics on participants' behavior in experimental tasks.

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

Word count: 1500

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

Xu and Tenenbaum (2007a; henceforth XT) provide an account of how learners might make appropriate generalizations in word learning, particularly at the subordinate level. They observe 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 Waxman (1990), XT 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 concurrent set of experiments that suggested that such narrowing was only observed when examples were chosen by an informative teacher (Xu & Tenenbaum 2007b; Lewis & Frank, 2016).

These findings have been an important part of a re-evaluation 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 reference based on the idea that referential descriptions are produced via strong sampling (Frank & Goodman, 2014; Horowitz & Frank, 2016). Subsequent work has found that toddlers' non-linguistic 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 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 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 more narrow 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 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, 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, SPSS replicated XT's main finding – more basic level generalization 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, 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 and adults are sensitive to sampling processes. At the same time, they are also surprising because others have suggested that simultaneous presentation highlights exemplar commonalities and increases memory consolidation (Lawson, 2014, 2017). 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 diverging 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 (cf. Baribault et al., in press). 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 coincidence 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. We recover the suspicious coincidence effect with a large effect size in both sequential and simultaneous presentation conditions, except under a particular trial order: when the three-exemplar trials are presented *before* the one-exemplar trials. When the three-exemplar trials are presented first, we see 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 are presented first, participants are aware of the exemplars from the previous trial and consequently do not interpret the single exemplar as the *only* observed exemplar from the target category. 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 publicly 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 (see SI for details).

Stimuli

Our picture stimuli were gathered on the internet, and closely resembled 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 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 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 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

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

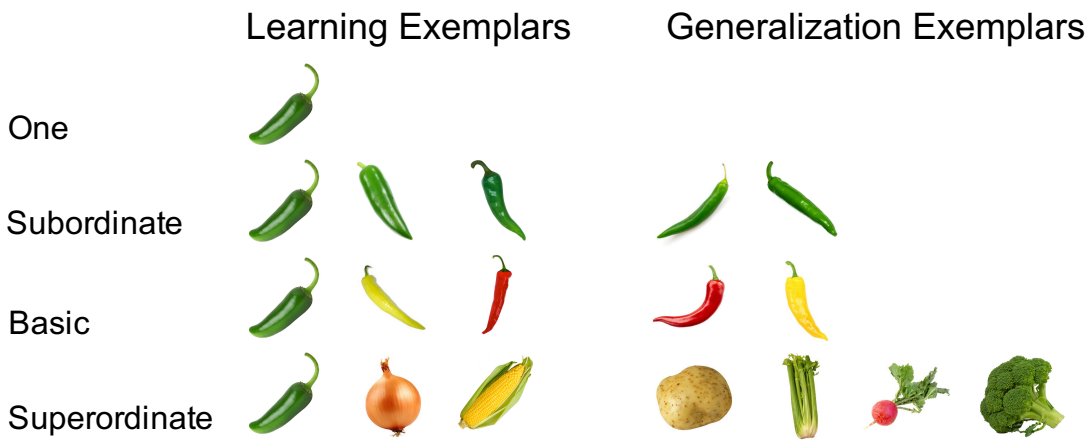

Figure 1. Sample learning and generalization sets. On a given trial, participants saw one or three exemplars of the same level from the learning set, followed by all exemplars from the generalization set (along with the generalization sets from the other categories).

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 generalization 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] Our set of experiments does not include all possible combinations of these design factors, but all levels are 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 XT (E1 and E2)[3] and SPSS (E2 and E3). In XT, the learning exemplars were presented statically and simultaneously, while in the key conditions from 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.[4] The generalization pictures did not appear in the sequential condition until after the training pictures had appeared for 6 seconds, but remained visible as participants selected generalization exemplars.

Trial order. In XT E1, the three one-subordinate trials occurred first followed by all other trial types ("1-3").[5] In contrast, in the main experiments in SPSS (E2 and E3), the three-subordinate trials occurred first ("3-1"). SPSS's replication of XT's simultaneous design (SPSS E1) showed a single block of either one-subordinate or three-subordinate first (randomized). In supplemental experiments (ES1 and ES2), SPSS directly explored whether trial order influenced the effect size by replicating SPSS E1 with three subordinate trials followed by the single subordinate trials.

Labels. XT used the same label for each category for the three-subordinate and one-subordinate trials (e.g., both the single pepper and the three-pepper trials would be

[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.

[4]Our implementation of the sequential design differed slightly from the SPSS design, which did not include a 1s interval between exemplar presentations.

[5]XT E2 used a between-subject design.

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	1.27 [0.84, 1.71]	XT E1/E2
2	50	simult.	1-3	pseudo-random	same	1.2 [0.77, 1.64]	XT E1/E2
3	50	simult.	1-3	pseudo-random	diff.	1.1 [0.67, 1.52]	
4	50	simult.	3-1	blocked	diff.	0.02 [-0.37, 0.42]	SPSS ES1/ES2
5	50	simult.	3-1	blocked	diff.	-0.02 [-0.42, 0.37]	
6	50	simult.	3-1	blocked	same	-0.04 [-0.43, 0.36]	
7	50	seq.	1-3	pseudo-random	same	1.43 [0.99, 1.87]	
8	50	seq.	1-3	pseudo-random	diff.	1.24 [0.81, 1.67]	
9	50	seq.	1-3	blocked	diff.	1.27 [0.84, 1.71]	
10	50	seq.	3-1	blocked	diff.	-0.43 [-0.83, -0.02]	SPSS E2/E3
11	50	seq.	3-1	pseudo-random	same	-0.3 [-0.7, 0.1]	
12	50	seq.	3-1	blocked	same	-0.18 [-0.58, 0.21]	

[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).

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 referred 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 were randomized ("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.

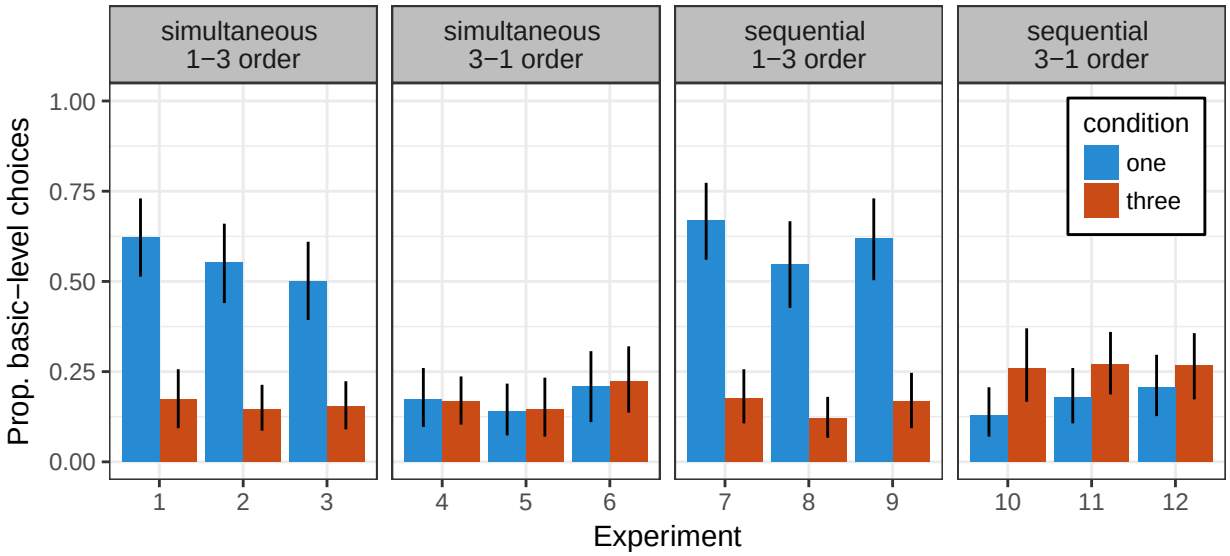

Figure 2. (#fig:bar_plot)Mean proportion generalizations to basic level exemplars in the one (blue) and three (red) subordinate exemplar conditions for all 12 of our experiments. Each facet corresponds to a pairing of presentation timing (simultaneous vs. sequential) and trial order (1-3 vs. 3-1). Ranges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

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 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. The model included both the present set of experiments as well as prior experiments by XT and SPSS. We used the metafor package (Viechtbauer, 2010) in R to fit our meta-analytic models.

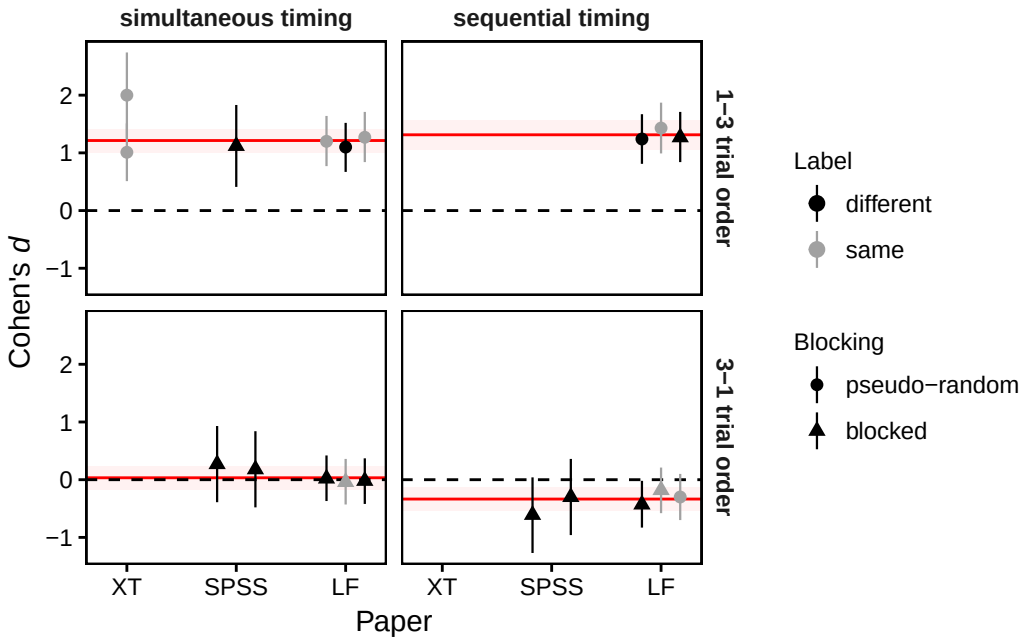

Figure 3. (#fig:es_plot)Effect sizes for all 19 studies conducted on the suspicious coincidence effect by XT (Xu & Tenenbaum, 2007a), SPSS (Spencer, et al, 2011), and the current authors. Each point corresponds to a study, with study authors on the x-axis and effect size on the y-axis. Facets correspond to different design parameters: Top facets show experiments with single exemplar trial first (1-3); Bottom facets show experiments with single exemplar trial second (3-1); Left facets show experiments with simultaneous presentation of exemplars, as in XT; Right facets show experiments with sequential presentation of exemplars, 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). 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. Points are jittered along the x-axis for visibility.

Results

Figure 1 shows the mean proportion generalizations to the basic level in the one- and three-subordinate trials for all 12 experiments,[6] and Figure 2 shows the corresponding effect 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: $d = 1.27$ [0.84, 1.71]; Exp. 2: $d = 1.2$ [0.77, 1.64]), with a magnitude comparable to the original XT experiments (XT E1: $d = 2$ [1.25, 2.74]; XT E2: $d = 1.01$

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

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

Fixed effect	beta	z-value	p-value
Intercept	1.36 [1.06, 1.65]	9.02	<.0001
Simultaneous vs. sequential timing	-0.16 [-0.37, 0.06]	-1.45	0.15
1-3 vs. 3-1 trial order	-1.44 [-1.75, -1.14]	-9.27	<.0001
Different vs. same label	0.06 [-0.19, 0.31]	0.51	0.61
Blocked vs. pseudo-random trial structure	-0.1 [-0.44, 0.24]	-0.56	0.58

[0.51, 1.51]). We also replicate the reversal in the suspicious coincidence effect observed by SPSS in an exact replication of their method (Exp. 10; d = -0.43 [-0.83, -0.02]), and with a magnitude comparable to the original experiments (SPSS E2: d = -0.61 [-1.27, 0.04]; SPSS E3: d = -0.3 [-0.96, 0.36]).

Critically, however, the meta-analytic model across all 12 experiments reveals that only trial order is a reliable predictor of effect size (β = -1.44; Z = -9.27; p <.0001), while timing (β = -0.16; Z = -1.45; p = 0.15), blocking (β = -0.1; Z = -0.56; p = 0.58), and label are not (β = 0.06; Z = 0.51; p = 0.61; 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.

Our data also suggest that the 3-1 ordering interacts with presentation timing: In experiments with the 3-1 ordering and sequential presentation (Exp. 10-12), we see a reversal of the suspicious coincident effect, as observed by SPSS. 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 (β = -1.18; Z = -8.04; p <.0001), but not presentation timing (β = 0.1; Z = 0.6; p = 0.55). However, there was also a significant interaction between the effects of two design parameters (β = -0.47; Z = -2.12; p = 0.03). This interaction effect is due to increased generalizations to the basic level when the three subordinate trials are presented sequentially (Exp. 10-12) compared to simultaneously (Exp. 4-6). In the General Discussion,

we consider why trial order might influence the suspicious coincidence effect, and 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, Perone, Smith, & Samuelson, 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 lead us to a different conclusion than SPSS: 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 the one exemplar trials) were not predicted by XT. 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 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 were less likely to generalize to the basic level compared to when the single 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 one of the exemplars on the three subordinate trials). This account 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 3-1 ordering, learners should be more likely to generalize to the basic level on the first trial (three exemplars) compared to the second trial (single exemplar, fourth observed exemplar), since seeing four exemplars is a bigger "suspicious coincidence" than three. We find some evidence consistent with this prediction from the meta-analytic model indicating a reversal of the effect under sequential timing, 3-1 ordering conditions.

Notably, while XT's model does not directly predict participants' behavior in the 3-1 ordering, there is a broader class of Bayesian models, of which XT's model is an instance, that does. These models account for pragmatic reasoning by assuming that speakers reason about the intention of others when making linguistic inferences (e.g., Frank, Goodman, & Tenenbaum, 2009). In this case, reasoning about the speaker's intention may lead participants to 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-subject design, as in XT's experiment with children (Xu & Tenenbaum, 2007a; E2).

If indeed participants interpret the one-exemplar trial in 3-1 orders as a fourth exemplar, then it is somewhat surprising that the identity of the label between the two trials does not matter: We see the same pattern when the labels are different (Exp. 10) as when they are the same (Exp. 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 a reasons why participants may not have carefully attended to labels across trials. First, participants are never tested on the meaning of labels, and the labels are 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 (since 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 find that label identity does not mediate the suspicious coincidence effect across experiments.

We also find that trial order interacts with presentation timing: Replicating SPSS, sequential presentation in the 3-1 ordering leads to a reversal of suspicious coincidence effect. SPSS's theory predicts the reversal under sequential presentation conditions, but it does not predict the observed interaction. There is also not a straight-forward explanation from XT's model. We offer one highly speculative account: Sequential presentation conditions may have appeared relatively more complex to participants compared to simultaneous presentation conditions, resulting in higher overall uncertainty in generalization judgement. This increased uncertainty may have lead 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 1-3 versus 3-1 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.

Both XT's and SPSS's work address an important puzzle in psychological sciences: How do learners learn concepts at multiple levels of abstraction? Their work focuses on a

simplified version of this puzzle where the learner must determine the corresponding labels to known concepts. Our findings here support the idea that they solve this puzzle via probabilistic inferences about the level of abstraction that is most likely given the observed data (the original "suspicious coincidence" effect). Importantly, by assuming that trials are non-independent, our interpretation is consistent not only with the original XT 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 non-linguistic 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.

References

Baribault, B., Donkin, C., Little, D. R., Trueblood, J., Oravecz, Z., Ravenzwaaij, D. van, ... Vandekerckhove, J. (in press). Meta-studies for robust tests of theory. *Proceedings of the National Academy of Sciences*.

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.

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

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

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, *15*(6), 297–301.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 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*, *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.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 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*(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/

Waxman, S. (1990). Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development*, *5*(2), 123–150.

Waxman, S., & Hatch, T. (1992). Beyond the basics: Preschool children label objects flexibly at multiple hierarchical levels. *Journal of Child Language*, *19*(1), 153–166.

Waxman, S., Shipley, E. F., & Shepperson, B. (1991). Establishing new subcategories: The

role of category labels and existing knowledge. *Child Development*, *62*(1), 127–138.

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.