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

Enter abstract here. Each new line herein must be indented, like this line.

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

Word count: X

Still suspicious: The suspicious coincidence effect revisited

Intro

What is the suspicious coincidence effect?

(Spencer, Perone, Smith, & Samuelson, 2011; F. Xu & Tenenbaum, 2007; Fei Xu & Tenenbaum, 2007)

Why is it important?

Spencer et al. paper

Methodological differences:

- simultaneous vs. sequential
- 3-1 vs. 1-3
- blocking
- same label vs. different label

other evidence relevant on this replicatione

Our current paper reports 12 experiments, 10 of which were pre-registered. 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, and contra SPSS, 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.

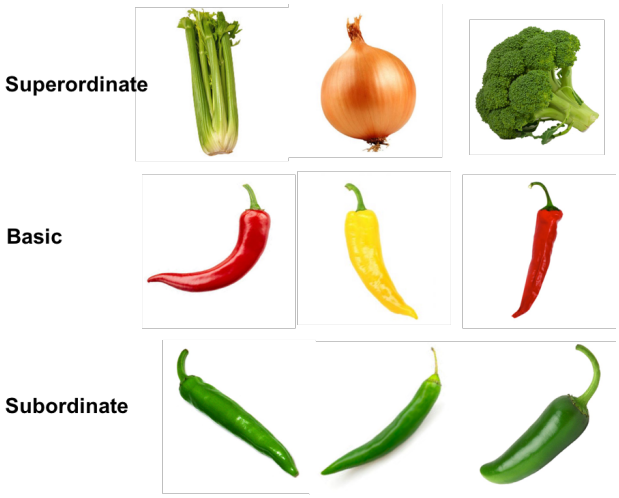

Figure 1. Sample stimuli. Three superodinate (top), basic (middle), and subordinate (bottom) exemplars from the vegetable category.

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 (https://osf.io/yekhj/), with the exception of Exps. 4 and 8.

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).

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 1.

Stimuli

Our stimuli closely replicated 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 pepper), 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. 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 other category memebers ("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

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	seq.	1-3	pseudo-random	same	1.42 [1.32, 1.52]	
5	50	seq.	1-3	pseudo-random	diff.	1.26 [1.18, 1.34]	
6	50	seq.	1-3	blocked	diff.	1.31 [1.23, 1.39]	
7	50	simult.	3-1	blocked	diff.	0.02 [-0.06, 0.1]	SPSS ES1/ES2
8	50	simult.	3-1	blocked	diff.	-0.06 [-0.14, 0.02]	
9	50	simult.	3-1	blocked	same	-0.14 [-0.22, -0.06]	
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.

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 superordate 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 the XT and SPSS studies (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)[1]. We describe each of these factors in more detail below.

[1]All experiments can be viewed directly at XXX.

Presentation Timing. Presentation timing was the key, theoretically motivated experimental design difference between the XT (E1 and E2[2]) 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 sequnetial condition until after the training pictures has appeared for 6 seconds, but remained visibile as participants selected generalization exemplars.

Trial order. In XT, the three one-subordinate trials occured first followed by all other trial types. In contrast, in SPSS (E2 and E3), the three-subordinate trials occured first. SPSS's replication of XT's simultanous design (SPSS E1) used the 1-3 ordering.[This isn't actually quite true: "The first block of trials always involved either one exemplar or three subordinate-level exemplars from each domain. The remaining blocks of trials were randomly ordered for each participant"... so 1 exemplar trials were first only half the time?].

Labels. XT used the same label for each category for the three-subordinate and one-suborinate trials. 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. Labels were randomized across trials.

Blocking. Finally, the studies by XT and SPSS differ in terms of 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 trials were randomized, whereas SPSS blocked all four trial types. Within each block, category order was randomized.

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

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, 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 2 shows the mean proportion generalizations to the basic level in the one- and three-subordinate trials for all 12 experiments, and Figure 3 shows the corresponding effect sizes (with XT and SPSS experiments included for reference).

In two exact replications of the XT method (XT E1 and X2), 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 (d = 2 [1.73, 2.27] and d = 1.01 [0.89, 1.13]). We also replicate the reversal in the suspicious coincidence effect observed by SPSS (SPSS E2 and E3) 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 suggests that only trial order is a reliable predictor of the magnitude of the susicpious coincidence effect. THE MODEL.

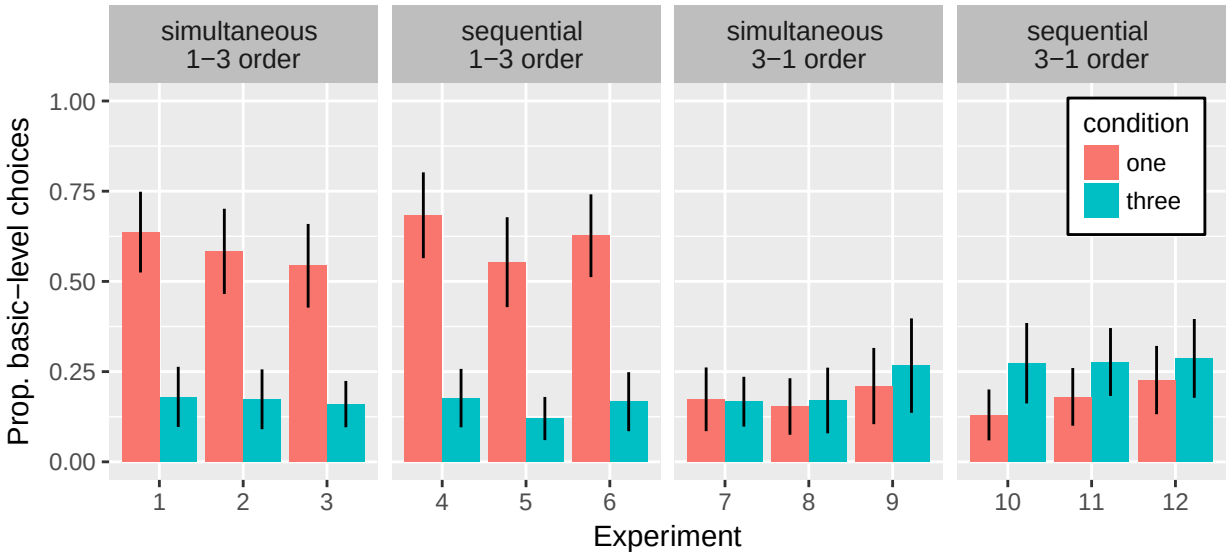

Figure 2. Mean proportion generalizations to basic level exemplars in the one (pink) and three (green) subordinate exemplar conditions for all 12 of our experiments. Each facet corresponds to a pairing of presentation timing (simultaneous vs. sequential) and trial order (1-3 vs. 3-1). Error bars are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

Discussion

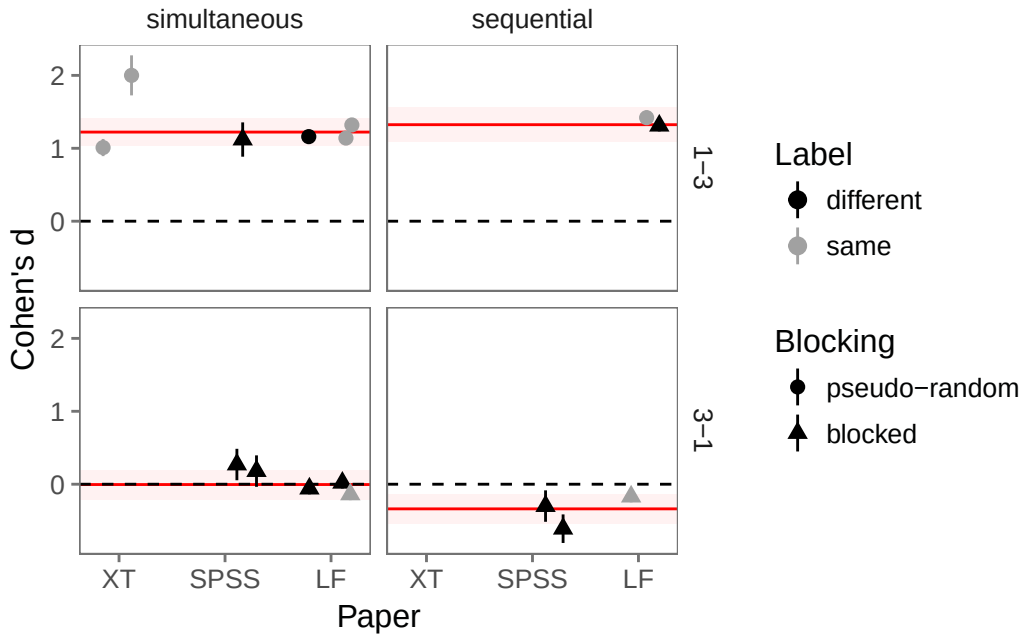

Figure 3. Effect sizes for all 19 studies conducted on the suspicious coincidence effect by XT (Xu & Tenenbaum, 2007a), SPSS (Spencer, et al, 2011), and the current authors. The top-bottoom facets indicate whether the single exemplar trial occurred first (1-3) or second (3-1). The left-right facets indicate 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 error bars are 95% confidence intervals.

References

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.

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

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

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 condition 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 order	-0.09 [-0.41, 0.24]	-0.52	0.6