Dear Dr. Lindsay,

Thank you and the reviewers for the thoughtful comments on our manuscript, “Still suspicious: The suspicious coincidence effect revisited.” Please accept our resubmission. We have addressed your comments and the comments of the reviewers, and we believe that the manuscript is improved. Please find below a point-by-point response to the comments.

Please do not hesitate to contact us if you have any questions or concerns.

Sincerely,

Molly Lewis and Michael C. Frank

Reviewer: 1

> *I see three facts:*
>
> > *1. The basic SC effect is replicated in the 1-3 order regardless of the stimulus presentation details*
> > *2. The basic SC effect is not replicated in the 3-1 order with the effect primarily impacting the one exemplar trials.*
> > *3. The SC effect is reversed in the 3-1 ordering as a function of the presentation timing with broader generalization on the 3-sub trials with sequential presentation.*
>
> *These three facts are taken directly from the paper. Only points 1 and 2 are in the results section. Point 3 is raised in the general discussion. I would recommend that this be included in the results section as this analysis follows directly from the previous literature (i.e., it examines whether a key effect in the literature was replicated).*

Thank you for this suggestion. We have added a paragraph in the results section discussing Point 3 by including a meta-analytic model that reveals this reversal.

> *If we agree on the facts, the next question is what do these facts mean? I will be honest and say that my read of the paper suggests that the authors are happy with a Bayesian status quo interpretation—that the data are generally consistent with XT's account, and that the data en masse don't support the SPSS account. I suspect the authors will agree with my assessment. But is this interpretation warranted by the facts?*
>
> *I think not. My conclusion stems from the answer to a central question: what is the SC effect about? At the heart, it's about the size principle (see equation 5 from XT 2007 Psych Review), also called 'strong sampling' in XT 2007 Developmental Science (and here). In their own words, "hypotheses with smaller extensions assign greater probability than do larger hypotheses to the same data, and they assign exponentially greater probability as the number of consistent examples increases." [emphasis added]. Critically, the primary data that test the size principle are from the 3-sub trials—the size principle doesn't modulate anything if there is only one exemplar—and how generalization narrows as a function of the number of exemplars at a particular generalization level. This logic is evident in, for instance, XT 2007 Dev Science where they only had children generalize with multiple subordinate items (i.e., there were no 'one exemplar' trials).*
>
> *If I look at the data here, the 3-sub trials do move around, increasing in basic level generalization the sequential 3-1 order condition (see point 3 above). This effect is small but significant. It replicates the key effect from SPSS. It is not explained by XT's model.*

The biggest effect, of course, is on the one exemplar trials. There, generalization narrows or broadens based on the trial order. But here's the rub—XT's model and strong sampling doesn't predict this effect AT ALL. The authors have come up with a reasonable explanation for this effect based on pragmatics, but scientifically, the authors should acknowledge that this effect is not consistent with XT's model. There is nothing in the size principle or the equations in XT's Bayesian model that says that generalization on the one exemplar trials should bop around as a function of trial order.

Concretely, if we agree on the facts above and if we agree on how equation 5 from XT's model creates the SC effect, then here are the conclusions I see…

1. *The basic SC effect is replicated in the 1-3 order regardless of the stimulus presentation details [XT's model explains this quite well; the data here provide an important qualification on the SPSS interpretation]*
2. *The basic SC effect is not replicated in the 3-1 order with the effect impacting the one exemplar trials only. [This is not explained by XT's model; SPSS showed the same effect but did not offer an interpretation]*
3. *The SC effect is reversed in the 3-1 ordering as a function of the presentation timing with broader generalization on the 3-sub trials with sequential presentation. [This is not explained by XT's model; this replicates the effect reported by SPSS]*

In summary, then, the data reported here do not support XT's model—2 of the 3 effects are not consistent with the model. I would like to see the authors at least acknowledge this.

This is a very reasonable point. We now acknowledge this directly (see below).

Beyond that, the authors are, of course, free to offer new explanations for the data. Their pragmatic interpretation seems reasonable. I thought the explanation of the reversal to be a bit weaker.

Below I provide several specific comments. But let me be clear about the key revisions I would like to see:

1. *Move the stimulus presentation timing result from the GD to the results (see point 3).*
2. *Acknowledge that findings 2 and 3 from the paper are not consistent with the size principle (strong sampling) from the XT model.*

Your points 2 and 3 correspond to two significant predictors in our meta-analytic models: presentation timing (Point 2) and the interaction of presentation timing with trial order (Point 3). The pragmatic explanation we offer in the paper (the single exemplar in the 3-1 order is

effectively a fourth exemplar) accounts for the main effect of presentation timing. We offer a possible account for the interaction effect, but agree that this account is highly speculative.

How do the accounts we offer relate to the theories posited by XT and SPSS? We agree with your point that the XT model, as presented in the paper 2007a paper, does not specifically predict the influence of pragmatic factors, and we now more directly highlight this point. But there is a broader class of models in the Bayesian framework that do predict effects that emerge from reasoning about the intention of the speaker (e.g. Frank, Goodman, & Tenenbaum, 2009), and would predict this type of pragmatic reasoning. Most relevant to the current work, there is also 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). We have clarified in the GD the relationship between the presentation ordering finding to the XT model.

In the case of the other effect – the interaction between presentation timing and trial order – it is not obvious how the XT model would account for this effect. But the SPSS effect does not predict this interaction either; it predicts a main effect of presentation order, which we do not find evidence for. We have added to the GD to highlight the fact that the XT model is not able to account for the observed interaction effect:

"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."

Detailed comments (note: page numbers are PDF page numbers...)

- *Abstract: "Yet both children and adults successfully learn noun meanings at the correct level of abstraction from similar evidence." What's the 'correct' level in this example? It depends. It could be Dalmatian, could be dog, could be animal. And what's the 'true' underlying category? There isn't one. Finally, 'making certain patterns of examples more consistent with a subordinate meaning than others'...this is a really loose summary of the SC effect. Please be more precise.*

Thank you, we have revised the abstract to clarify that by "correct level" we mean the level intended by the speaker. We have also clarified the description of the SC effect in the abstract.

- *P5, line33: should be basic memory and comparison processes (or perceptual processses) [the SPSS account did not focus solely on memory]*
- *P6, line 16: should be 'children and adults' [since the experiments reported here included adults]*

We have clarified both of these points

.

- *P6, line 16: not sure why Lawson goes against SPSS—if I highlight commonalties and all three exemplars are chili peppers, wouldn't that lead to subordinate level generalization which is what we all find?*

Thanks for this comment. In SPSS 2011, our understanding of the account for the observed difference between the simultaneous and sequential presentation conditions is that simultaneous presentation leads to "increased discrimination" (p. 1050) compared to sequential presentation, making the representation of individual exemplars more precise and thus leading to more subordinate generalization. In contrast, Lawson (2014) argues that simultaneous presentation "supports alignment of shared features" whereas sequential presentation "supports detection of differences" (p. 148). It seems that both claims can't be true - simultaneous presentation cannot both make exemplars more different and more similar to each other.

- *P7, line 57: something is off here. Perhaps it should be: 'participants are aware of the exemplars from the previous trial and therefore do not interpret the single exemplar as the only observed exemplar from the target category."*

Thank you for spotting this. We have fixed this.

- *P 9, line 42: suggested edit: 'while in the key conditions from SPSS, participants saw...' This is important because we did replicate XT in the other experiments.*

Noted, we have clarified this.

- *P 10, line 56: typo (an extra 'not')*

Fixed.

- *P 14, line 10: the SC effect suggests a 'powerful mechanism" by which learners might overcome... 'Powerful' seems like a stretch to me since the effect is pushed around across conditions, but ok…*
- *P 15, first full paragraph: the authors replicate our effect, but offer a new interpretation. This is ok, but perhaps the authors could at least acknowledge our explanation of the same effect?*

Thank you, we have added a sentence pointing to the fact that SPSS attributed this effect to sequential timing of the exemplars.

Reviewer: 2

This ms is focused on an important question: How do adults and children interpret words referring to object categories, despite the inherent ambiguity in potential scope? More

specifically, how do they use sampling distributions to learn words for non-basic level meanings, including especially subordinates ("chili pepper"; "collie")?

Two major concerns limit the impact of the current revision: insufficient treatment of the existing evidence (from children, in particular) and the authors' interpretation of the results.

Existing developmental literature.

There is considerable work on sampling diversity and naming in children. There is also ample work on what kinds of information children need in addition to evidence from sampling distribution if they are to establish subordinate level categories.

The authors have a choice. If they want to claim, as they now do, that this ms is about development (see abstract, intro, etc), then it is essential that they describe the evidence concerning children's performance in cases with sampling distributions so very like the conditions of the experiments described here. If instead the authors choose not to succinctly discuss the existing evidence from children and relate this work to that in the current ms, then the ms should be revised considerably to remove claims about development or acquisition.

The authors argue that their decision to focus on the 'suspicious coincidence' is more than a matter of resolving a methodological matter (XT vs SPSS). If resolving issues of acquisition are at issue, as the authors claim, then it is crucial that they acknowledge, however briefly, the highly relevant evidence on this very topic from children. Although a few developmental papers are cited, this is not sufficient: the authors should convey the striking parallels between those developmental designs and the design featured in the current ms (as in XT and SPSS) and the strength of the findings (that children tend to extend novel names to the basic level category, rather than subordinate level category, unless they are given more evidence than sampling distribution alone to do so).

As I mentioned in my first review, Waxman 1990 is especially germaine. But now, let me be more clear: In this task, children saw three exemplars of a given subordinate level category (as in the current ms design); these exemplars were either named or not named. (They also saw three exemplars of basic level categories; and superordinate level categories, as part of this comprehensive design). This strong parallel in design, in manipulating sampling diversity, makes this work directly relevant to the current ms. It is relevant enough to be described so as to permit readers to see the parallels. XT's 2007 paper was motivated in part by preschoolers' performance in this task (children tended to focus on the basic, rather than subordinate level, in this very similar design). Also see Waxman & Hatch; Waxman Lynch Casey & Baer, etc for other designs and converging evidence of children's performance with different sampling distributions.

By failing to engage this evidence, the current manuscript fails to meet its mark of addressing the very question regarding acquisition that they pose.

A succinct description of the designs and the child evidence would not add much space, but it would make it clear how children respond in tasks like the ones presented here.

We have added a paragraph in the introduction that more directly summarizes the developmental literature on word learning across at multiple levels of abstraction, and have also highlighted the methodological similarities between the present (and previous) work and Waxman (1990).

Remaining interpretive questions:

The motivation for the particular experiments conducted would benefit from additional clarification. The authors do not test an exhaustive 2x2x2x2, but instead have selected a subset of these possible combinations. Certainly, motivating each specific experiment would make the ms prohibitively long. However, a helpful compromise would be to explain that all factors are equally balanced across the 12 experiments (i.e., for each variable, each of its levels is reflected in 6 experiments). This would be helpful for interpreting the results.

Thank you, this is a helpful suggestion we have added a sentence explaining this in the Procedure section:
"Our set of experiments does not include all possible combinations of these design factors, but all levels are tested in at least one experiment."

One claim - that adults interpret the one-exemplar trial in 3-1 orders as a 4th exemplar from the same category -- warrants clarification. If this is the case, it is surprising that the introduction of a different label on the 1-exemplar trial does not diminish this effect. Adults adhere to the principle of contrast: different words should have different meanings (e.g., Clark, 1988). Thus, when Mr. Frog uses a new word on the 1-exemplar trial, the prediction is that adults should infer he has a different meaning in mind-yet participants appear to use the same category. Have adults forgotten the previously used words? Do they ignore words on these trials? More discussion, even if speculative, is warranted to interpret this surprising finding. It would be beneficial to address briefly the theoretical implications of null effects (eg in both label and blocking).

Thanks you for these suggestions. We have added a paragraph in the GD speculating about this pattern of findings, particularly the null effect of labeling.

More importantly, the authors claim that the 1-exemplar trial in 3-1 orders essentially functions as a 4th exemplar for the category. It is worth clarifying that this "4th exemplar"

is actually one of the original 3. Although clear in Fig 1, this should be clarified in the discussion of the effect.

Thank you - we have clarified this point.

Moreover, on the strong sampling hypothesis, it is unclear what evidential value a 4th exemplar holds when it is, in fact, a re-presentation of one of the original 3. If participants remember the original 3 exemplars, then simply seeing one of them again (labeled again as belonging to the category) may not yield additional information.

But the authors make a different claim: that the 1-exemplar trial (in the 3-1 order) is less likely to yield basic-level interpretations than the 3-exemplar trial (p. 14). This would suggest that participants are assigning additional evidentiary weight to that "4th exemplar." However, this effect is driven almost entirely by performance in the 3-1/Sequential experiments; there appears to be no effect in the 3-1/Simultaneous experiments (ds = .02, -.02, -.04). The authors then give a different, unrelated account of this 3-1/Sequential interaction effect (that sequential presentation leads to greater uncertainty, resulting in a default to basic-level meanings). In other words, the authors interpret both a main effect and its interaction when the former is solely attributable to the latter. It would be advantageous to focus their interpretation of the evidence on the interaction, setting aside the discussion of the 4th exemplar's additional evidentiary weight.

Resolving this interpretive issue is important. Doing so would not detract from the central claim that the 'suspicious coincidence' obtains across all situations but 3-1 orders, and this exception is likely due to participants viewing the 1-exemplar trial as referencing the same category as the previous 3-exemplar trial. The strong sampling hypothesis seems to make no clear prediction for those "4th exemplar" trials (except a rejection of the basic level interpretation); and there seems to be no clear pattern of participant behavior either.

To clarify, we argue that the 3-1 order is a main effect - in particular, that it leads to reduced generalization to the basic level in the single exemplar trials. We see this effect in *both* sequential and simultaneous presentation timings for the 3-1 ordering. This main effect is reflected in our meta-analytic model, and can be explained by the “fourth exemplar” account.

In addition, there is a second (much smaller) pattern to explain: Why there is increased generalization to the basic level in the three subordinate condition under sequential presentation and 3-1 trial order conditions. As we note in the GD, neither XT nor SPSS predict this interaction, and we only have a speculative explanation for this finding.