Experiments test whether sequential vs. simultaneous presentation (1) or identical exemplars (2) modulates the "suspicious coincidence" effect reported by Xu & Tenenbaum (2007).

Adele E. Goldberg, Lauren L. Emberson, Isaac N. Treves
Princeton University, Psychology Department

We replicate Xu & Tenenbaum (2007)'s "suspicious coincidence" effect, regardless of whether three exemplars are presented sequentially or simultaneously, or whether the exemplars are identical to one another or distinct. Our replication of Xu & Tenenbaum is a partial failure to replicate Spencer et al. (2011). Differences between our design and previous ones: ours was massively between subjects using participants on Mechanical Turk in order to avoid possible effects. Specifically, each of 511 participants witnessed a single trial. We used instances of categories that were distinct from those of either previous study (dog, fish, flower, bird *vs.* dog, truck, pepper). Our work does not investigate generalization to the higher, superordinate level, as generalizations to that level on the basis of a single exemplar are uncontroversially rare.

Xu and Tenenbaum (2007) manipulated the statistics of participants' experience with novel words. When adults and 3-4 year old children were shown a single exemplar of a category (e.g., a picture of a Dalmatian dog labeled a *fep),* they were often willing to extend the label to any instance of the corresponding "basic level" category (here, *fep* = "dog"). But after witnessing **three** different *feps*, each of which was a Dalmatian, participants were much more likely to apply the term more narrowly to only Dalmatians (i.e., the subordinate level category) and not to other dogs. They suggest that children and adults are aware that selecting three instances of the same narrow subcategory presents the learner with a "suspicious coincidence," given the assumption that the three exemplars are chosen randomly. This suspicious coincidence is resolved by assuming that the label only refers to members of the narrower subcategory (here, "Dalmatian"). Thus the Xu & Tenenbaum study suggests that the statistics in the input play a crucial role in determining which level of categorization a novel term applies to (see also Gweon, Tenenbaum, & Schulz 2010; Lawson 2014; Xu & Denison 2009).

We consider a current controversy in the literature surrounding the basic narrowing effect found by Xu & Tenenbaum (2007). Spencer et al. (2011) have argued that the narrowing effect of witnessing three similar exemplars was at least partially due to the low-level attentional demands of the experiment. Adopting an associationist perspective, they argued that presenting three items simultaneously, as Xu & Tenenbaum (2007) had done, increased the opportunity for a fine-grained comparison among the exemplars, which led to better memory of the shared features. The narrowing effect was then, on this view, a result of the narrow features being made more salient (see also Garner 1974; Gentner & Namy 2006; Gibson 1969; McMurray, Horst, & Samuelson, 2012; Samuelson, Schutte & Horst 2009, Sandhofer & Smith 2001). In a series of studies, Spencer et al. (2011) found that the narrowing effect of witnessing three exemplars was eliminated when the three exemplars were presented sequentially instead of simultaneously, apparently because, when the exemplars were presented sequentially, they could not be compared as easily. In fact, Spencer et al. (2011) found that the narrowing effect was reversed in the sequential condition: participants were more likely to generalize from three exemplars presented sequentially than from a single exemplar.

This finding by Spencer et al, that the narrowing effect was eliminated—even reversed—when three exemplars were presented sequentially is somewhat surprising, even if one adopts a fully associationist perspective. Participants witnessed the three exemplars immediately following one another within a span of 3 seconds, so we might expect them to be capable of remembering and comparing features across exemplars. That is, while sequential presentation may be expected to reduce the comparison across exemplars, it is not clear why it should eliminate such a comparison. As long as the three simultaneously presented entities are compared at all, one might expect their shared attributes to be made more salient, which could, on the associationist view, lead to a more narrow interpretation of the category. It is especially difficult to explain why three sequentially exemplars should show more generalization than a single exemplar, a finding reported but without detailed explanation by Spencer et al.

With great respect for Spencer's group, we aimed to better understand their results, but we make no headway. Experiment 1 results replicate Xu and Tenenbaum (2007), regardless of whether the presentation is simultaneous or sequential. An additional condition (Experiment 2) was run in which all three exemplars were identical to see if that might lead to a reduction or elimination of the suspicious coincidence effect. We find that it did not.

Experiment 1

Participants

We collected responses to a single question from 411 self-identified native English speakers from the US, over the age of 18 (*mean* = 31.79, 19-75) using Amazon's Mechanical Turk. Using a 2 x 2 design we varied whether participants witnessed one or three exemplars and whether the exemplars were witnessed simultaneously or sequentially. In order to avoid order effects, each participant received only a single trial (i.e., viewed 1 or 3 pictures and then made a response). In experiments reported, the protocol was approved by the Princeton University IRB.

Methods

A survey was created using Qualtrics. Participants gave informed consent.

Each participant initially saw the following instructions:

"Mr. Frog speaks a different language, that he'll try to teach you. You will be shown a picture and told what it's called in his language. You will then need to pick out other examples."

Participants were shown a screen shot of a single exemplar or a set of three different instances of the same subordinate level category (e.g., three different golden retrievers). They were then told, "Here is a *fep*" or "Here are three *feps*". The nonsense words used were *feps* (dogs), *zecks* (birds), *nats* (fish), *galts* (flowers).

Participants were randomly and roughly equally assigned to each of 4 conditions and a single stimulus within the condition**.**

Participants were asked to "check the box(es) for any other feps that you find in the pictures below." They were always provided with 16 pictures including 2 subordinate level matches (e.g., two new golden retrievers), 2 basic level matches (e.g., a Labrador and a beagle), and 12 distractors (pictures of categories used with other participants). The order of the pictures was randomized.

Results

The results from 1 vs. 3 exemplars presented simultaneously and sequentially are provided in Figure 3. The grey bars represent the percentage of times participants selected other instances of the same subordinate category as the witnessed exemplar(s); e.g., how often participants considered a different golden retriever a *fep* when shown one or three *feps* that were golden retrievers. As expected, this number is reassuringly near ceiling across conditions. Also as expected, and as indicted by the virtual absence of black bars in Figure 3, participants virtually never selected distractors as instances of the novel category. For example, they did not choose any flower or fish as an instance of a *fep,* if *fep* had been illustrated with a golden retriever.

Notice that it does not make sense to specify whether one exemplar is presented simultaneously or sequentially: only a single exemplar is presented. Nonetheless, for the sake of symmetry, we ran the 1-exemplar condition twice. This provides essentially a replication, and the results in both 1-exemplar conditions are reassuringly non-distinct.

The key comparisons are between the blue bars on the left and right panels. These represent the percentage of the time participants chose entities from the same basic level (and distinct subordinate level) from the witnessed exemplar; e.g., the percentage of the time participants selected a Labrador or a poodle as a *fep* when shown one or three *feps* that were golden retrievers.

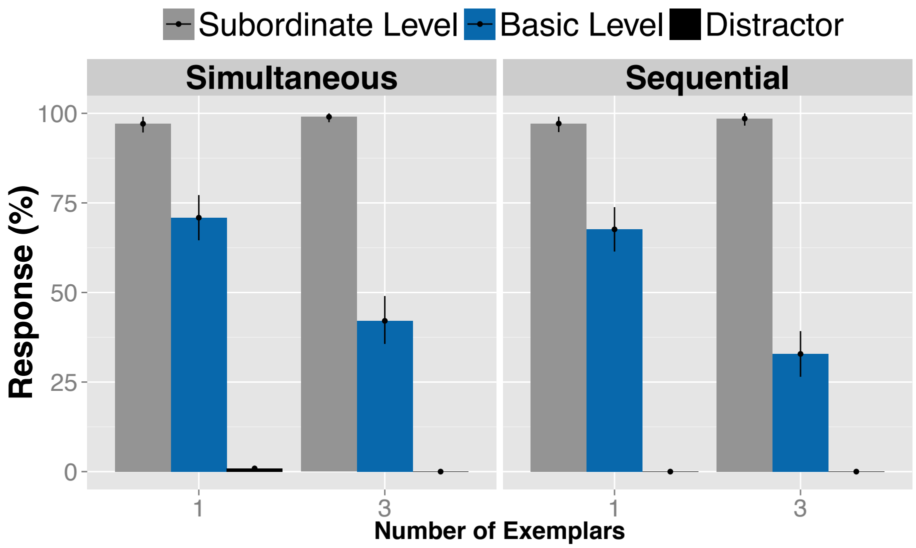

Figure 1: Percentages of responses to a request concerning what a novel word (e.g., *galt*) referred to: subordinate level choices (grey bars), basic level choices (blue bars), and distractors (black), after viewing one or three exemplars illustrating the novel word, presented simultaneously (left panel) or sequentially (right panel). Subordinate level choices are at ceiling; choices of distractor items, at floor.

This and all models reported, unless otherwise specified, control for any effects of category by including an additional fixed effect for category of pictures (dogs, flowers, fish, birds). Category was included as a fixed effect instead of a random effect (i.e., conducting mixed effects model) because of the small number of categories and, moreover, the dog category was consistently found to exhibit a higher degree of generalization compared to other categories. Thus, category violates the assumptions of normally distributed intercepts necessary to include it as a random effect. Subjects were also not a random effect since each participant supplied a single data point.

A comparison across conditions replicates the narrowing effect for 3 exemplars compared to 1 exemplar found by Xu and Tenenbaum (2007). Specifically, a general linear model was fit to the data (RStudio, 0.98, R 3.1.1), in which a logistic regression predicted a number of Basic

Level responses each participant selected (out of 2) based on a single fixed effect of the number of exemplars (1 vs. 3). Predicting the number of Basic Level responses based on exemplar confirmed systematic and robust differences in generalization to the basic level with the presentation of 3 *vs.* 1 exemplars (β= 0.3245, Z = 5.25, $p < 0.0001$).

Having replicated the narrowing of generalization reported in Xu and Tenenbaum (2007), we considered whether there were differences in generalization across presentation type. While there is a numerical difference in the expected direction, the result did not approach significance. First, we added the fixed effect of presentation (sequential vs. simultaneous) to the model reported above to determine whether including this variable accounted for a significant amount of the variance. As with this previous model, this augmented model had a significant fixed effect for number of exemplars viewed (β = 0.3262, Z = 5.27, $p < 0.0001$), but relevantly, the fixed effect of presentation type did not have a significant intercept (β = -0.15, $p > 0.2$) and the inclusion of this effect did not significantly increase the amount of variance accounted for by the model (χ^2 =1.58, $p > 0.2$).

Because presentation type affects only participants viewing 3 exemplars, we then restricted the model to focus on just the responses to 3 exemplars (see Figure 4). Again, we found no significant effect of presentation type (β = 0.14, Z = 1.48, $p > 0.1$) and that the inclusion of presentation type did not significantly increase the amount of variance accounted for by the model over a model a fixed effect to control for category (χ^2 = 2.2, $p > 0.1$). Thus, we do not find evidence for differences in generalization across presentation type comparing sequential and simultaneous directly.

Discussion

Experiment 1A replicated Xu & Tenenbaum's narrowing effect when three exemplars of a category are witnessed instead of one. This effect was not modulated by whether the exemplars were presented simultaneously or sequentially. This is in contrast to Spencer et al. (2011) who found a narrowing effect in the simultaneous but not in the sequential condition. Thus, we cannot explain the lack of a narrowing effect in the Spencer et al (2011) studies, nor its reversal.

Experiment 2: using identical vs. distinct exemplars

It is conceivable that Spencer et al. did not find a narrowing effect of three exemplars compared with one exemplar because participants may have misidentified their three exemplars as three presentations of a single exemplar. In particular, each exemplar was displayed on the same background, and all three were very close in appearance. If the exemplars were construed as a single instance that simply moved across the screen sequentially from left to right—despite the label that described them as "three wugs"—that could explain why the narrowing effect was eliminated: perhaps participants viewed the *one wug* and the *three wug*s conditions as identical. Spencer et al. themselves tried to address this concern with an additional study (their experiment 3), where each of the three exemplars was displayed in the same location, sequentially. They again found no evidence of narrowing.

We also address the possibility that participants construed the three *wugs* as being an instance of the same *wug* in our replication with two conditions in which three exemplars are shown sequentially: in a non-identical condition, each of the three exemplars was obviously unique (in terms of shading, pose, background, and orientation); in the "identical" condition, all three exemplars were identical.

Participants

We collected responses from 100 self-identified native English speakers in the US, over the age of 18 (*mean* = 29.9, 19-55) using Amazon's Mechanical Turk. The age range of this group was not different than the ages of either 3-exemplar groups of Experiment 1, used for comparison (χ^2 = 4.42, *df* = 2, *p* = 0.11). In this experiment, we presented participants with three identical exemplars of the same subordinate level category, presented sequentially. We did not include a condition in which three identical exemplars were witnessed simultaneously as that seemed to be quite unnatural. Participants were assigned randomly and roughly equally to one of four categories. As in Experiments 1 and 2, each participant received only a single trial.

Methods

A survey was created using Qualtrics. Participants gave informed consent.
Each participant saw the same instructions as in Experiment 1: "Mr. Frog speaks a different language, that he'll try to teach you. You will be shown a picture and told what it's called in his language. You will then need to pick out other examples."
Below we compare results from participants exposed to 3 identical exemplars presented sequentially, with the data from participants who witnessed 3 non-identical exemplars presented either sequentially or simultaneously. The latter data sets were taken from Experiment 1.

Participants

We collected responses from 100 self-identified native English speakers in the US, over the age of 18 (*mean* = 29.9, 19-55) using Amazon's Mechanical Turk. The age range of this group was not different than the ages of either 3-exemplar groups of Experiment 1, used for comparison (χ^2 = 4.42, *df* = 2, *p* = 0.11). In this experiment, we presented participants with three identical exemplars of the same subordinate level category, presented sequentially. We did not include a condition in which three identical exemplars were witnessed simultaneously as that seemed to be quite unnatural. Participants were assigned randomly and roughly equally to one of four categories. As in Experiments 1 and 2, each participant received only a single trial.

Methods

A survey was created using Qualtrics. Participants gave informed consent.
Each participant saw the same instructions as in the replication above: "Mr. Frog speaks a different language, that he'll try to teach you. You will be shown a picture and told what it's called in his language. You will then need to pick out other examples."
Below we compare results from participants exposed to 3 identical exemplars presented sequentially, with the data from participants who witnessed 3 non-identical exemplars presented either sequentially or simultaneously. The latter data sets were taken from experiment above.

Results

We again modeled subject responses to basic level exemplars while controlling for category differences but now considering responses across 3 levels of presentation type (simultaneous, sequential and sequential-identical). Three identical exemplars sequentially did not show a different effect than showing three non-identical exemplars sequentially. Specifically, we used (non-identical) sequential as the intercept, and found no significant difference between conditions ($Z = 0.51, p > 0.5$). With simultaneous presentation as part of the intercept, we again find no significance effect for either sequential presentation of 3 different or 3 identical exemplars ($Zs < |1.5|, ps > 0.1$). In addition, we compared a model including presentation type with a model that only controls for category and do not find that the inclusion of presentation type significantly increases the amount of variance accounted for in the data ($\chi^2 = 2.61$, $p > 0.25$). Thus, we do not find evidence that presentation type significantly explains variation in subject responses nor do we find evidence for differences of sequential presentation of either 3 identical or 3 different exemplars compared to the simultaneous presentation of 3 different exemplars.

Thus, we see no evidence to suggest that participants in Spencer et al (2011) mistakenly assumed that the three instances of *wugs* were actually a single instance presented three times, since in the present study, the three instances *were* actually identical, and yet it did not significantly affect participants' willingness to generalize to the basic level. Thus, our results undermine the idea that Spencer et al.'s findings are a result of either simultaneous *vs*. sequential presentation (Experiment 1), or a result of the exemplars being perceived as identical (Experiment 2; see also Spencer et al.'s Experiment 3).

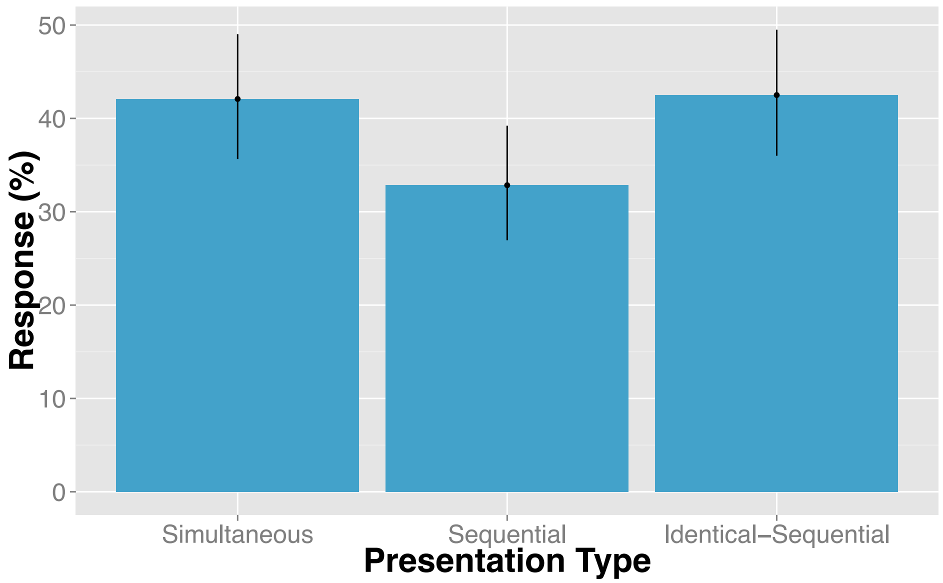

Figure 2: Percentage responses at the basic basic-level to a request concerning what a novel word (e.g., *galt*) referred to. Participants viewed three exemplars illustrating the novel work in three types of presentation: Simultaneous, Sequential and Identical-Sequential (1 exemplar presented 3 times sequentially, the other two conditions had 3 unique exemplars).

Conclusion

While we replicate Xu & Tenenbaum's (2007) "suspicious coincidence" effect, demonstrating that learners assume a more narrow interpretation of a novel label when three exemplars of the same subordinate category are witnessed instead of only one, we are unable to ascertain what led to the Spencer et al. (2011) failure to replicate that result, as we were unable to replicate the failure to replicate.

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.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–72. doi:10.1037/0033-295X.114.2.245.