**Online Supplement**

As stated in the text, the main experiment exposed children to novel names applied to typical exemplars before atypical items. This ordering was motivated by the fact that Spencer et al. (2011) had found that a narrow interpretation of one novel word tended to narrow participants’ interpretation of other words on subsequent trials, while the reverse was not found (see also Lewis & Frank 2018). Given that the basic-level interpretation of a single example has been widely documented (Waxman 1990; Archambault, Gosselin & Schyns 2000), our interest is in whether children systematically reduce the number of exemplars selected when presented with an atypical member of the category. We therefore ordered typical before atypical trials so that participants would need to retreat from the expected basic-level interpretations on typical exemplars to a more narrow focus only on the atypical items.

However, in response to a reviewer’s concern, we ran a new group of children on the same material but with the order of typicality reversed (atypical trials now precede typical trials). Reassuringly, the blowfish effect is significant in the new data set as described below. This new set of data confirmed that it is not trial order (typical followed by atypical) that produced a narrower interpretation of a novel word but the typicality of exemplars (either 1 or 3 exemplars).

**Participants**

The final sample included twenty (20) children between the ages of 4 and 6 (*M*= 1859 days, 5.09 years, *SD* = 288.5 days; 10 female, 10 male; 99.6% exposure to English in the home, range = 95-100% English exposure) were tested in the Princeton University Baby Lab. Three additional children were run but excluded because they spoke less than 80% English in the home (2 children) or for not following instructions (1 child, they selected 6 or more distractors on the last 50% of the trials), although the results do not change if they are included.

Procedure

The follow-up procedure was identical to the main task with one exception: the order of atypical and typical exemplars was reversed with four trials of each atypical and typical. That is, each child witnessed four atypical trials, chosen in random order from the four semantic categories (e.g., an atypical fish, atypical flower, atypical bird, atypical dog) and *then* four typical exemplars in a new random order of the four categories (e.g., typical dog, typical bird, etc.). A new experimenter tested the children. As before, the number of items (1 vs. 3) varied between-subjects.

Results

The same analysis was run that had been used for the main analysis. Specifically, a logistic regression predicted the number basic-level responses (out of 2) for each participant. Exemplar typicality (atypical vs. typical) and the number of exemplars (1 vs. 3), and their interaction were fixed effects, and category was included as a random effect in order to control for any differences in generalization across categories. The binary variables of typicality and number of exemplars were again contrast coded.

 Critically, Exp. 1A replicates the typicality (Blowfish) effect (*β* = -1.61, *Ζ* = -5.69, *p* < 0.0001, see Figure A.1). In fact, consider the beta values, this numerically effect is larger than in the main experiment despite the smaller sample size. When considering the mean responses for the atypical exemplars in the current dataset compared to the atypical trials in the dataset reported in the main text, we find highly similar means. In the current dataset, mean responses for 1 atypical exemplar is 0.0208 (*SD* = 0.144) and for 3 atypical exemplars is 0.156 (*SD* = 0.515). In the dataset from the main text, mean responses for 1 typical exemplar was 0.32 (*SD* = 0.66) and for 3 atypical exemplars was 0.07 (*SD* = 0.34). Interestingly, inspection suggests that it is the responses to the typical exemplars that differed and led to the greater beta value. In this dataset, mean responses for 1 typical exemplar was 1.35 (*SD* = 0.837) and 3 typical exemplars was 1.53 (*SD* = 0.671). In the full dataset in the main text, mean responses to 1 typical exemplar was 0.66 (*SD* = 0.91) and for 3 typical exemplars was 0.28 (*SD* = 0.59). The mean responses to typical exemplars in this supplementary dataset are more in line with the responses made by adults (as reported in the main text). While further investigation would be necessary to make this claim conclusively, responding to typical exemplars after atypical exemplars increased children’s basic level interpretation of novel words.

While this supplementary dataset confirms the effect of typicality on basic level interpretation of novel words, we do not see the narrowing effect of 3 similar exemplars in this smaller dataset. In fact, there was a marginally significant result in the opposite direction: The multiple exemplars were more likely to be interpreted at the basic level (*β* = 0.535, *Ζ* = 1.84, *p* = 0.060).  Moreover, we find a marginally significant interaction between these two variables (*β* = 0.473, *Ζ* = 1.67, *p* = 0.096) unlike the main dataset. We do not wish to overinterpret the lack of a narrowing effect in the case of multiple exemplars (or the apparently larger typicality effect), as the effect of multiple similar exemplars has been replicated both in our main findings and elsewhere (Lewis & Frank, 2016). In particular, we note that the current manipulation was between-participants, with a small sample size (20 subjects total). Because of the power difference, we did not perform a comparison across experiments.

To summarize, we had originally been concerned about eliminating the basic-level bias by providing atypical exemplars before typical exemplars, because we thought that participants might continue to associate a narrow interpretation with a novel label once they had done so for atypical exemplars. If this were to happen, the standard basic-level bias would not be in evidence. However, we now see that this concern was unfounded. In fact, if anything, the basic-level bias is stronger for typical exemplars *after* exposure to atypical exemplars, even though exemplar members were always between distinct categories. For example, after witnessing e.g., *fep* labeling a blowfish (atypical), participants tended to interpret *fep* narrowly, as only referring to other blowfish, and then when the same participants witnessed, e.g., *galt* to label a Labrador (typical) they generally interpreted *galt* at the basic-level to include all dogs. This is predicted by the idea that atypical exemplars lead to markedly more narrow interpretations of novel words than typical exemplars. Thus, Exp. 1A clearly puts to rest the possibility that the typicality effect in the main experiment was due to the ordering of atypical and typical trials.



Figure 1A: Data collected for Experiment 1A. Children’s mean # of basic-level selections (out of 2) when asked to find matches of a novel word (e.g. *galt*) when witnessing 1 or 3 typical or atypical exemplars. Error bars represent bootstrapped 95% confidence intervals.