

Supplementary Material A: Simulation-based power analysis – Details

For the power analysis, we assumed the following probabilities for observing a novel word per condition. Differences between two conditions always indicate differences in the distributions from which data for these conditions are sampled, and thus the presence of an actual effect that has to be detected in the analysis (Wickelmaier, 2022). A graphical illustration of these effects is presented in Figure 3.

1. We start with base probabilities of .025, .20, .22, and .24 for the four levels of conceptual distance (0 to 3 modifications), respectively. We choose the value of .05 for the base concepts since we don't want to rule out that the present experimental setting (with many novel stimuli) might prompt some participants to still produce novel responses for these. For Experiment 3, we instead start with base probabilities of .20, .24, .22, and .15 for the four levels of conceptual distance, respectively.
2. For all items in the open response format, except those in Experiment 1 and 2 with a base concept as target, we then add an effect of target frequency that always increases by .05 per repetition of the target item. For Experiment 3, this means adding a main effect of target frequency in the open response condition; for the open response format in Experiments 1 and 2, this also means adding an interaction. This also introduces an interaction between target frequency and response format in all three experiments, since the effect is not added in the one-word condition
3. For Experiment 2, for all items except those with the base concept as a target, we then add an effect of distractor frequency that always increases by .015 for each time the target item has been presented as a distractor before (i.e., a total of .12 for the maximum number of 8 repetitions)
4. We then add a main effect of task setting by increasing all probabilities by .10 for Experiment 1 and .05 for Experiment 3, but with no increase for Experiment 2

5. We then add to the main effect of response format by increasing all probabilities in the one-word format by an additional .15
6. Finally, to model our expectation that speakers will almost exclusively produce existing words for the base concept targets in Experiment 1 and 2, irrespective of the other variables (leading to the various interactions with conceptual distance discussed in the Hypotheses), we re-set all probabilities for the base concepts in Experiment 1 and 2 to .025.
7. As random effects, we add random intercepts for each participant (sampled from a normal distribution with $M = .04$, $SD = .02$) and for each target (sampled from a normal distribution with $M = .05$, $SD = .03$) to each data point. These effects are sampled again in each simulation.
8. In each simulation, resulting probabilities $< .001$ were set to .001 and probabilities $> .999$ were set to .999. Then, the probabilities were converted into binary outcomes by using them to sample from a binomial distribution with that probability of observing a novel word.

Note again that these values are not based on informed assumptions or expectations about actual effect sizes, but instead reflect the minimal effects that we consider to be of psychological relevance (following Wickelmaier, 2022; see also Lakens, 2022); these effects are already so small that any smaller effects would be practically irrelevant. Following the same reasoning, for the absolute size of the probabilities we oriented ourselves to the findings by Pugacheva and Günther (2024), who (across all targets) observed about 26% novel word responses in their Experiment 1 (where participants could choose to produce either existing or novel words). We expect a much higher novel word rate in the present study (as participants have to refer to novel concrete objects rather than concrete and abstract familiar concepts), so this rate provides a lower boundary for possible probabilities for novel words.

In each simulation, we randomized the trial order according to the Procedure outlined in the Method section, added the random intercepts for participants and the 64 different targets, and sampled binary outcomes (“novel word produced” vs. “existing word produced”) from these probabilities. Then, on each simulated data set, we performed all the analyses outlined in the Results section. To estimate power, we checked the proportion of 100 such simulation runs for which the test result in question was significant ($p < .05$; Wickelmaier, 2022). The results are presented in the “Sample size justification” section.