

## Supplementary Materials:

### Replication of analyses using subs2vec (subtitles) embeddings

In the main manuscript, we calculated similarity between responses using word embeddings trained on English Wikipedia+Statmt news corpus (Bojanowski, Grave, Joulin, & Mikolov, 2017). As additional verification, we replicated all our analyses using the subs2vec embeddings derived from movie and TV show subtitles (van Paridon & Thompson, 2020).

#### 1. Semantic centrality of responses (see Appendix)

We used word embeddings to analyze the centrality of the responses to the superordinate terms. We calculated the centroid of each superordinate term by selecting the eight most typical responses for each term (as indicated by the typicality ratings) and averaging the vectors of these responses. We then calculated the similarity between each individual response and the term centroid. We fit separate models for the Label/Exemplar vs. Label/Definition data, using sum contrast coding for the Cue Type variable.

Mean similarity to centroid for each Cue Type and Response Number are shown in Figure S1. In the Label vs. Exemplar model, responses in the Label condition were more similar to the centroid than responses in the Exemplar condition ( $\beta = .032$ ,  $CI_{95} = [.001, .062]$ ,  $t = 2.00$ ,  $p < .05$ ) and responses decreased in their similarity to the centroid as Response Number increased ( $\beta = -0.047$ ,  $CI_{95} = [-0.069, -0.025]$ ,  $t = -4.12$ ,  $p < .001$ ). In the Label vs. Definition model, responses in the Label condition were more similar to the centroid than responses in the Definition condition ( $\beta = .12$ ,  $CI_{95} = [.061, .18]$ ,  $t = 3.17$ ,  $p < .01$ ) and responses decreased in their similarity to the centroid as Response Number increased ( $\beta = -.19$ ,  $CI_{95} = [-.21, -.17]$ ,  $t = -23.39$ ,  $p < .001$ ). In both models, adding the interaction between Cue Type and Response Number did not increase model fit. As with the centroid analysis based on the Wikipedia embeddings, we see a Label Advantage: when participants were not given labels for the superordinate categories, their responses were less central to the structure of the category.

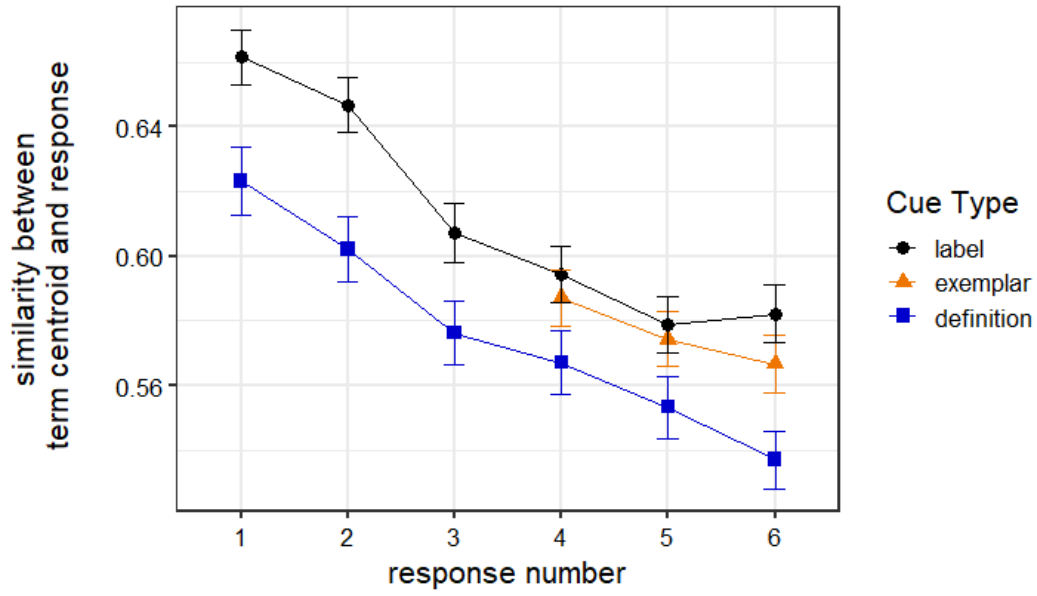


Figure S1. Mean centroid similarity of responses, grouped by Cue Type and Response Number (Experiment 1A: Label and Exemplar conditions; Experiment 1B: Definition condition). Error bars show 95% confidence intervals of the mean.

Figure S2 shows, for individual terms, the mean similarity between responses and category centroids for each Cue Type.

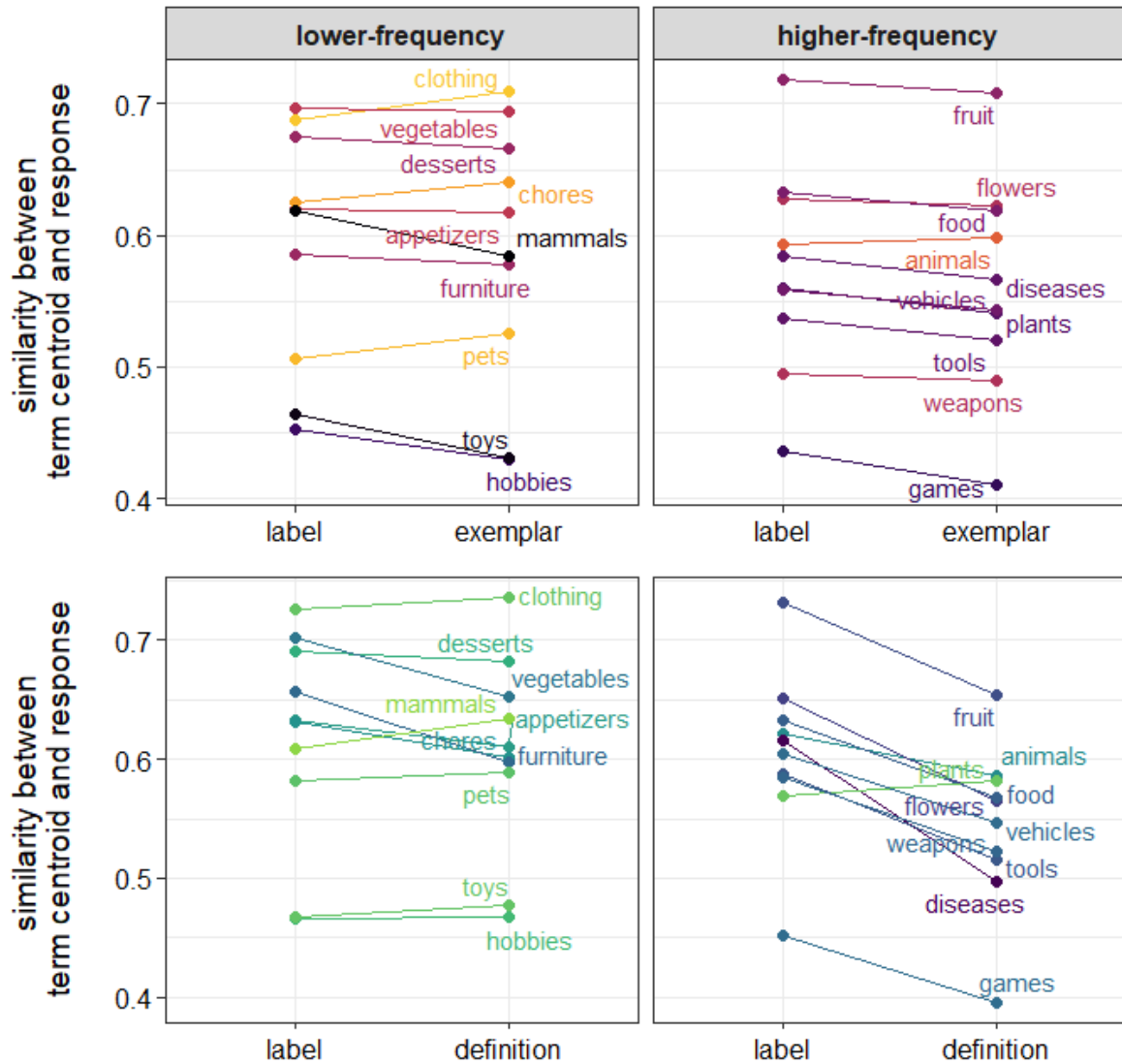


Figure S2. Mean centroid similarity of responses, grouped by Cue Type and superordinate term (Experiment 1A: Label and Exemplar conditions; Experiment 1B: Definition condition). Terms are grouped by lower vs. higher frequency. Darker colors indicate a stronger Label Advantage for that term.

## 2. How diverse were the responses across different Cue Types? (See Section 2.8.3)

We calculated the similarity between the responses for each pair of participants in word embedding space. We then calculated the mean similarity between each pair of participants for each Cue Type. For this analysis, we used treatment coding for the Cue Type variable, with the Label condition as the reference level.

Pairwise similarity was lower in both the Exemplar and Definition conditions than in the Label condition (Exemplar:  $\beta = -.10$ ,  $CI_{95} = [-.19, -.015]$ ,  $t = -2.29$ ,  $p < .05$ ; Definition:  $\beta = -.26$ ,  $CI_{95} = [-.45, -.067]$ ,  $t = -2.64$ ,  $p < .05$ ) (Figure S3).

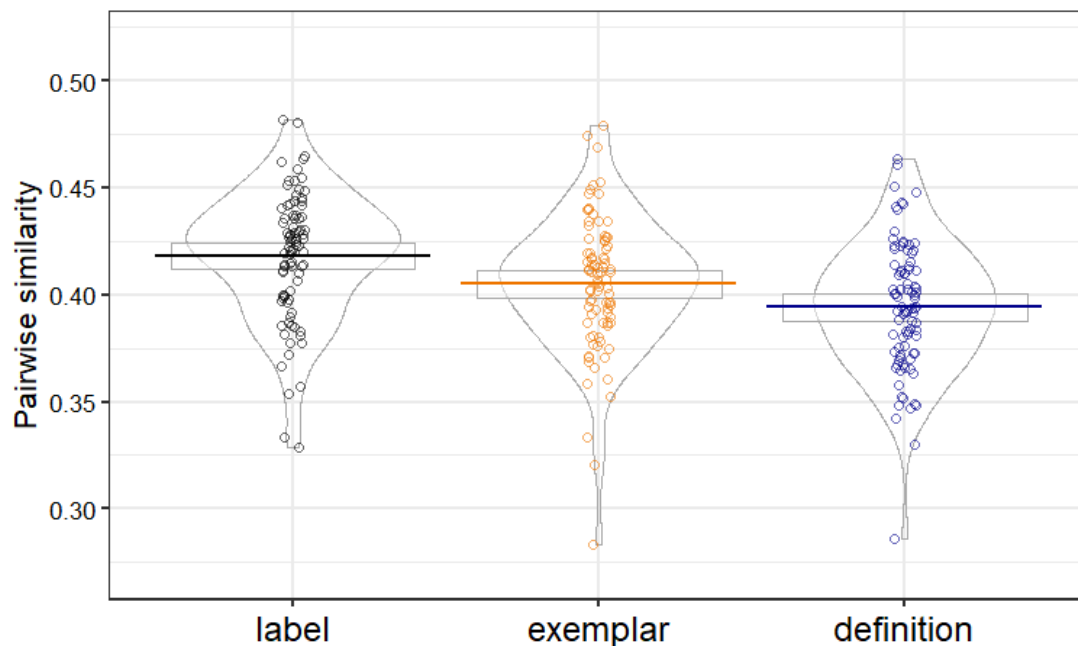


Figure S3. Mean similarity in word embedding space between each pair of participants for each Cue Type. Each point is a participant (indicating the mean similarity between that participant and every other participant).

Consistent with the analysis using the Wikipedia embeddings, these results suggest that when participants viewed an exemplar list, they interpreted the category structure of the list in divergent ways, leading them to align less with one another than when viewing a superordinate term. The significant difference between the Label and Definition conditions should be interpreted with caution, as this difference was not replicated using the Wikipedia embeddings.

Figure S4 shows the Label Advantage for individual terms.

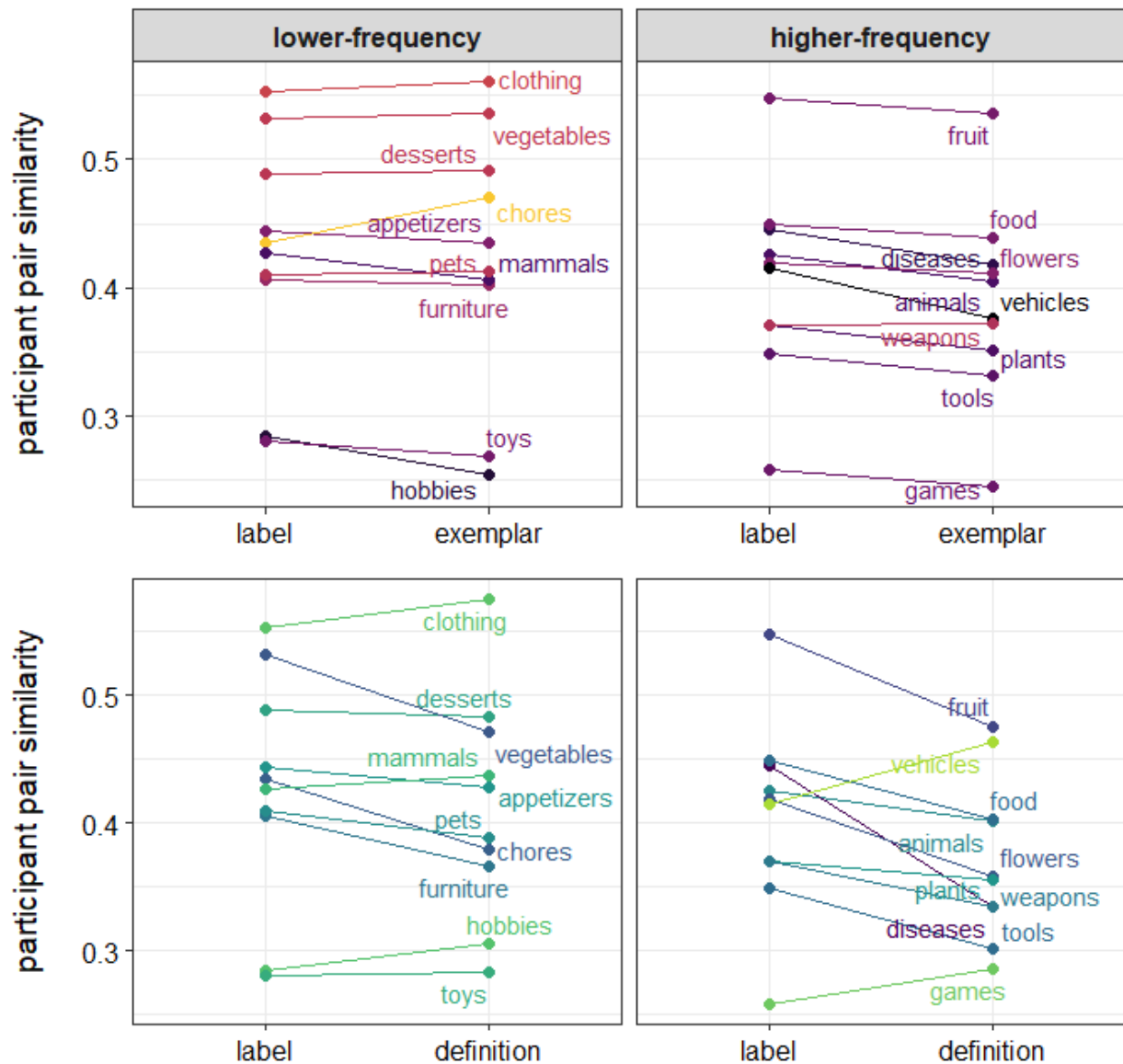


Figure S4. Similarity between participants in word embedding space, grouped by Cue Type and superordinate term (Experiment 1A: Label and Exemplar conditions; Experiment 1B: Definition condition). Terms are grouped by lower vs. higher frequency. Darker colors indicate a stronger Label Advantage for that term.

### 3. Similarity of responses within trials (see Section 2.8.4)

We used the word embeddings to calculate the mean semantic similarity between responses 4-6 and responses 1-3 in each trial. Similarity on this measure was significantly higher in the Exemplar condition than in the Label condition ( $\beta = .098$ ,  $CI_{95} = [.037, .16]$ ,  $t = 3.15$ ,  $p < .01$ ). Consistent with the analysis using the Wikipedia embeddings, this suggests that participants in the Exemplar condition were more strongly tethered to the particular responses 1-3 than participants in the Label condition were. Similarity was marginally lower in the Definition condition than in the Label condition ( $\beta = -.14$ ,  $CI_{95} = [-.30, -.018]$ ,  $t = -1.75$ ,  $p = .095$ ).

### References

- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135-146. doi:10.1162/tac1\_a\_00051
- van Paridon, J., & Thompson, B. (2020). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior research methods*, 53, 629-655. doi:10.3758/s13428-020-01406-3