

Similarity to Centroid Analysis Using Word Embeddings

As a replication of the typicality ratings analysis in Section 2.8.2, 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. For example, we calculated the centroid of the term *desserts* by averaging the individual vectors for *chocolate cake*, *lemon meringue pie*, *chocolate pudding*, *strawberry shortcake*, and four other typical desserts. We then calculated the similarity between each individual response and the term centroid.¹ This measure complements the typicality ratings analyzed in Section 2.8.2, as similarity in word embedding space is based on statistics of word usage, rather than intuitions about typicality.

The mean similarity of each response relative to the centroid for each superordinate term is shown in Figure S5. As with the typicality analysis, we fit separate models for the Label/Exemplar vs. Label/Definition data, using sum contrast coding for the Cue Type variable. In the Label vs. Exemplar model, responses in the Label condition were more similar to the centroid than responses in the Exemplar condition ($\beta = .047$, $CI_{95} = [.016, .078]$, $t = 3.00$, $p < .01$) and responses decreased in their similarity to the centroid as Response Number increased ($\beta = -0.047$, $CI_{95} = [-0.071, -0.023]$, $t = -3.84$, $p < .001$). In the Label vs. Definition model,

¹ We adopted this approach rather than compare each response to the term itself (e.g., the similarity between *cheesecake* and *dessert*). As the superordinate terms are in a hierarchical level above the responses, the superordinate terms are not necessarily used in the same contexts as typical examples of those terms (e.g. *cake* and *desserts* are used in different contexts). For this reason, the average of highly typical examples is likely a better representation of the central tendency of the category in word embedding space than the word itself. See Rissman and Lupyan (2021) for a comparison of the similarity between the responses and the superordinate terms. They found that responses were more similar to the superordinate term in the Label condition than in the Exemplar condition.

responses in the Label condition were more similar to the centroid than responses in the Definition condition ($\beta = .087$, $CI_{95} = [.033, .141]$, $t = 3.17$, $p < .01$) and responses decreased in their similarity to the centroid as Response Number increased ($\beta = -.21$, $CI_{95} = [-.23, -.19]$, $t = -24.20$, $p < .001$). In both models, adding the interaction between Cue Type and Response Number did not increase model fit. As with the typicality ratings, 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.

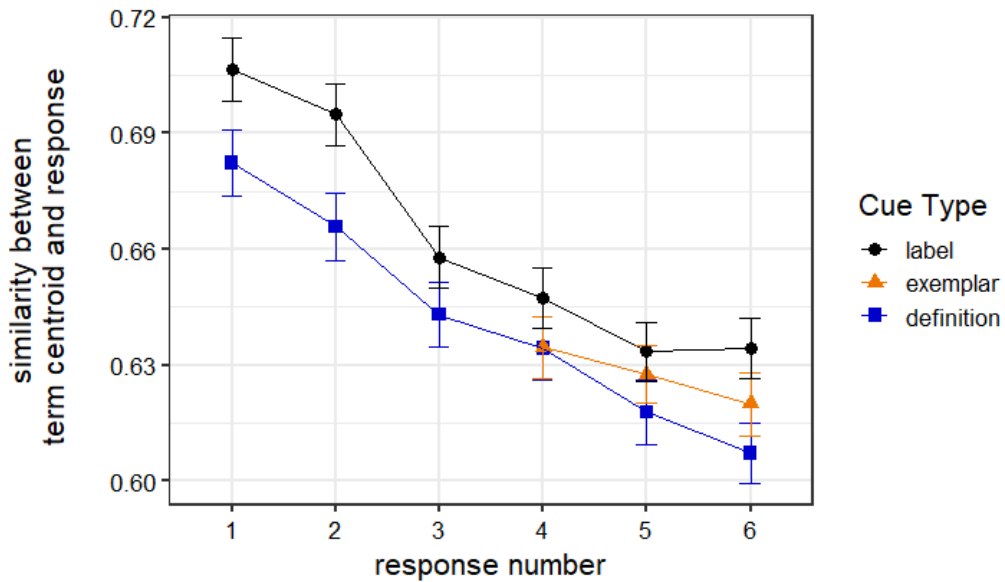


Figure S5. Mean centroid similarity of responses, grouped by Cue Type and Response Number (Experiment 1A: Label and Exemplar conditions; Experiment 1B: Definition condition). Error bars show 95% confidence intervals of the mean.

Figure S6 shows, for individual terms, the mean similarity between responses and category centroids for each Cue Type. The advantage of labels over exemplars was largest for *toys* and *mammals* and smallest for *clothing* and *pets*. The advantage of labels over definitions was largest for *flowers* and *diseases* and smallest for *mammals* and *toys*. In models of the Label/Exemplar and Label/Definition data, neither the frequency and nor the generality of the terms interacted with Cue Type (p 's $> .1$). The ratings difference between our Turker-generated definitions and dictionary definitions did not significantly interact with Cue Type ($p > .1$).

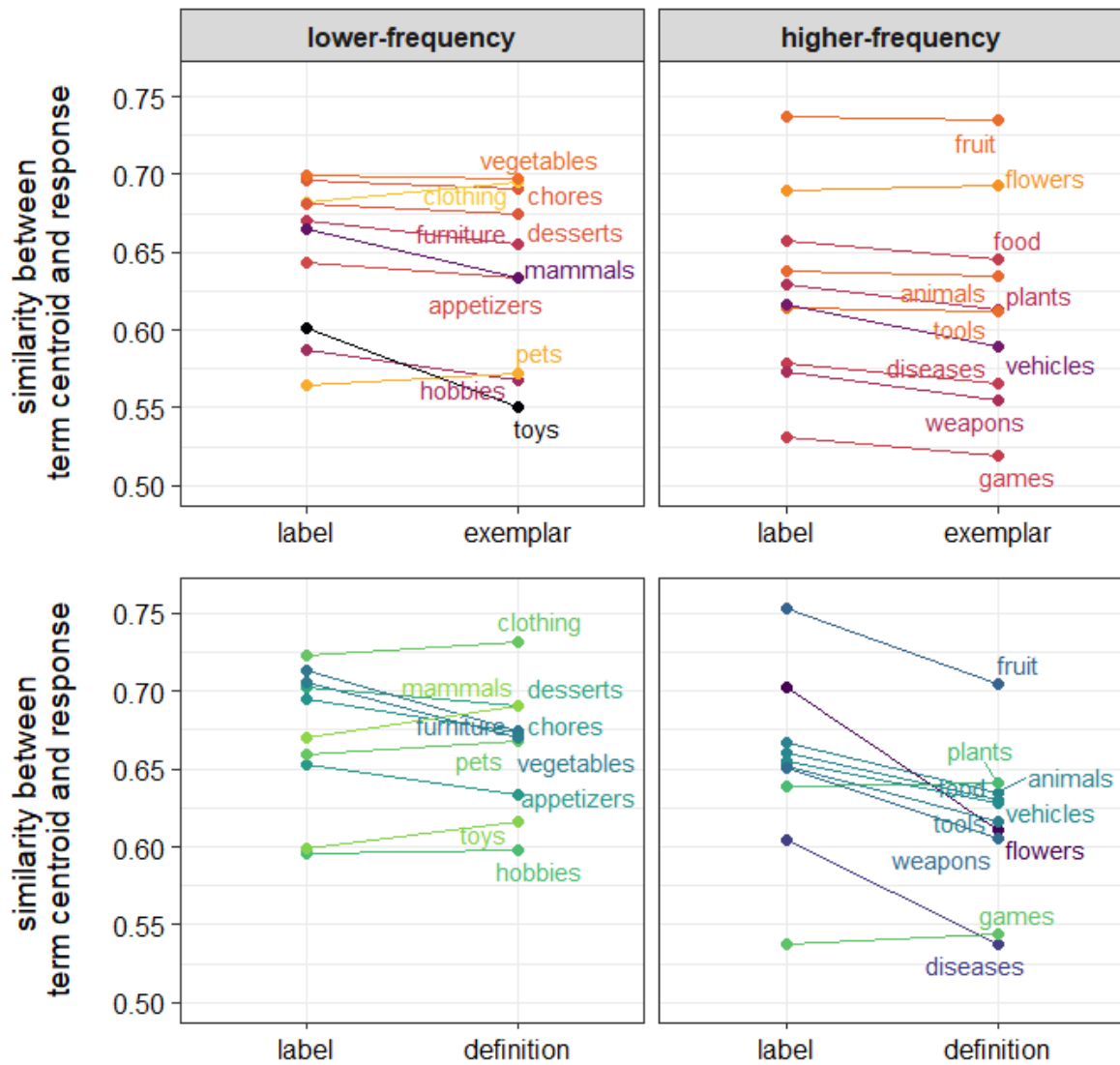


Figure S6. Mean centroid similarity of responses, grouped by Cue Type and superordinate term (Experiment 1A: Label and Exemplar conditions; Experiment 1B: Definition condition). Terms are grouped by lower vs. higher frequency. Darker colors indicate a stronger Label Advantage for that term.