Behavioral Feature Rating Methods

Participants
We recruited 1065 participants on Amazon Mechanical Turk (AMT). All participants self-reported being between 18 and 40 years old and native English speakers. Participants were paid $7/hour for the their time, prorated to the nearest quarter hour. Experimental methods were approved by the New York University Committee on Activities Involving Human Subjects/IRB and the Air Force Research Laboratory.

Stimuli and task
Each participant rated how well a series of words were described by a set of semantic features. The entire word set was composed of 619 common English words and there were 21 semantic features rated. The features spanned a range of perceptual, motor, and abstract characteristics that might characterize a concept (see Table in Jeff's supplemental results). Each participant rated five features for each of twenty words using a web interface controlled by Javascript and the psiTurk online experiment platform (Gureckis et al., in press). As no single participant was making all word-feature ratings and given the ongoing unpredictable nature of online data collection, the selection of words and features for each participant happened according to the following scheme: The existing set of ratings was queried to find those word-feature combinations with the fewest total ratings already made and the twenty words with the lowest rating count across word-feature combinations were selected for presentation for that participant.

Each "trial" involved presentation of a single word and five features to rate, and so the twenty selected words were randomized to provide the trial order. Within each trial, one feature to rate was the low count cell used to select the word for presentation. The remaining four features were selected as the features for the current word with fewest ratings. Then, the top to bottom screen position of the five features for each word was randomized to minimize effects of consistently rating two features in the same order (e.g., always rating the "appearance" feature before the "sound" feature). Each trial of the task consisted of a word presented at the top of the screen, a Wordnet definition underneath the word to provide a guide to the intended meaning, and a list of five features to rate. Participants indicated their rating as to the applicability of a feature to a word from "*Not at all"* to "*Very much so*" using a slider controlled by the mouse and these ratings produced continuous raw data outputs from 0-100. Within each trial, participants were free to adjust their responses until they indicated that they were finished by clicking "next" to receive the next trial. The mean number of ratings per word-feature pair was 7.4 and the feature rating values used in subsequent analyses were calculated as the mean across participant rating within feature for each word.

References
Gureckis, T.M., Martin, J., McDonnell, J., Rich, A.S., Markant, D., Coenen, A., Halpern, D., Hamrick, J.B., Chan, P. (2015, in press)
psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*
DOI: 10.3758/s13428-015-0642-8