Estimating the semantic properties of words with simple mean ratings can be misleading
Van Rynald T. Liceralde & Peter C. Gordon (University of North Carolina at Chapel Hill)
vliceral@live.unc.edu

Semantic properties of words (e.g., emotionality, concreteness, embodiment, plausibility, animacy) are acknowledged to have wide-ranging effects on sentence processing. Many of these semantic properties are estimated by asking subjects to rate words on some semantic dimension (e.g., For visual perceptual strength: “On a scale of 1 (not experienced) to 6 (experienced greatly), to what extent do you experience WORD by seeing?”) and computing the mean of these ratings. These mean ratings are then assumed to reflect the location of the word in the continuum of the latent semantic dimension, e.g., a mean of 3 indicates that a word has moderate visual strength.

This assumption about mean ratings has at least two psychometric corollaries. First, it assumes that in this context, treating an ordinal scale as interval or ratio (henceforth, metric) by taking the average of ratings is innocuous. Second, it assumes that words with the same mean ratings have roughly the same location in the continuum. In this work, we demonstrate that these corollaries do not necessarily hold, in that words with the same metric mean ratings may exhibit different latent estimated means when the ratings are modeled as ordinal and vice versa.

Using data from the Lancaster Sensorimotor Norms1, where the distribution of sensorimotor perceptual ratings given to ~36,000 words are publicly available, we fit two types of hierarchical Bayesian models to the visual perceptual rating subset of the data to estimate mean ratings (μ) and standard deviations (σ) for each word: (1) a standard linear (metric) model that assumes the ratings are directly generated from a latent normal distribution of perceptual strength and (2) an ordinal probit model that treats the ratings as ordinal and assumes that they are linked to the latent normal distribution by a function that maps thresholds in the latent distribution to the ratings based on the observed frequency distribution of ratings2. Typically, the latent distribution in the probit model is Normal(0,1) but we specified our probit model to be comparable in scale to the metric model. Thus, we have two sets of μ and σ that were estimated for each word: metric μs and σs and probit μs and σs.

To illustrate the core finding of this work, Figure 1 shows a scatterplot of the estimated metric μ against the probit μ for 50 words. Words that are aligned on a vertical line are words for which the metric μ are estimated to be different but the probit μ are estimated to be similar (e.g., Words 2 & 20; 22 & 23; Figure 2). In contrast, words that align on a horizontal line are words for which the metric μ are estimated to be similar but the probit μ are estimated to be different (e.g., Words 22 & 10; 23 & 31; Figure 3). One can imagine points within the region of the gray curves (which represent the population of words with the highest and lowest estimated probit σ) that align vertically and horizontally to picture the many possibilities of misalignment between metric and probit μ that could be present in words that researchers are considering as target stimuli on the basis of metric mean ratings. Given that the posterior predictive distributions indicate that the probit model provides a better fit to the data than the metric model (Figures 2 & 3), these results show that treating ratings of semantic properties as metric is not innocuous. Doing so could give a misleading impression that words that have been controlled based on metric mean ratings lie on the same location in the semantic dimension of interest when they do not.

We anticipate that the same issues extend to other semantic properties that are estimated using subjects’ ratings and would be exacerbated for words and properties where subjects use idiosyncratic criteria to provide their ratings (e.g., see the bimodal distribution of the visual perceptual strength of incremental in Figures 2 & 3). This work has implications for the statistical and experimental control of these properties in psycholinguistic research and for work in computational linguistics that use semantic property norms that are based on metric mean ratings to estimate characteristics of language samples. The better fit of the ordinal probit model to the data suggests that it is worthy to consider probit μ as an alternative to the metric μ.
Figure 1. Scatterplot of words' posterior modal metric $\mu$ against posterior modal probit $\mu$. Line segments intersecting dots indicate 95% highest density interval. Curves indicate the estimated locations of metric $\mu$ as a function of probit $\mu$ for words that would have the smallest estimated probit $\sigma$ (steeper/higher curve) and words that would have the largest estimated $\sigma$ (shallower/lower curve).

Figure 2. Specific cases from Figure 1 (boxed in red) showing words that align vertically have different estimates for metric $\mu$, but similar estimates for probit $\mu$. Posterior predicted normal distributions (for metric $\mu$) and posterior predicted probabilities (for probit $\mu$; with vertical segments indicating 95% HDI) are superimposed on the data. The fit of the ordinal probit model is substantially better than the metric model, as the probit predictions match the observed distributions of ratings more. $N =$ number of ratings; $\mu =$ modal mean; $\sigma =$ modal standard deviation.

Figure 3. Specific cases from Figure 1 (boxed in purple) showing words that align horizontally have similar estimates for metric $\mu$, but different estimates for probit $\mu$. See Figure 2 description for interpretation of figure.

References:
1. Lynott, Connell, Brysbaert, Brand, & Carney. (in press). BRM.