Behavior Research Methods (2018) 50:115–133
https://doi.org/10.3758/s13428-017-1009-0

When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms

Geoff Hollis[1] · Chris Westbury[2]

Published online: 10 January 2018

Abstract

Large-scale semantic norms have become both prevalent and influential in recent psycholinguistic research. However, little attention has been directed towards understanding the methodological best practices of such norm collection efforts. We compared the quality of semantic norms obtained through rating scales, numeric estimation, and a less commonly used judgment format called best-worst scaling. We found that best-worst scaling usually produces norms with higher predictive validities than other response formats, and does so requiring less data to be collected overall. We also found evidence that the various response formats may be producing qualitatively, rather than just quantitatively, different data. This raises the issue of potential response format bias, which has not been addressed by previous efforts to collect semantic norms, likely because of previous reliance on a single type of response format for a single type of semantic judgment. We have made available software for creating best-worst stimuli and scoring best-worst data. We also made available new norms for age of acquisition, valence, arousal, and concreteness collected using best-worst scaling. These norms include entries for 1,040 words, of which 1,034 are also contained in the ANEW norms (Bradley & Lang, *Affective norms for English words (ANEW): Instruction manual and affective ratings* (pp. 1-45). Technical report C-1, the center for research in psychophysiology, University of Florida, 1999).

Keywords Semantics · Semantic judgment · Best-worst scaling · Rating scales · Numeric estimation

There have been numerous recent efforts to collect semantic norms across a large number of words, semantic dimensions, and languages (e.g., Brysbaert, Warriner, & Kuperman, 2014; Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014; Stadthagen-Gonzalez, Imbault, Sánchez, & Brysbaert, 2017; Warriner, Kuperman, & Brysbaert, 2013). Whereas previous semantic norms have contained entries for hundreds to thousands of words, new semantic norms contain entries for tens of thousands of words. These very large norms sets are useful for multiple purposes including, but not limited to, testing for effects across a variable's range (e.g., Baayen, Milin, & Ramscar, 2016), testing computational models of semantics (e.g., Hollis & Westbury, 2016), and quantifying the semantics of natural language in situations where word choice is more varied than typical laboratory conditions (e.g., Hollis, Westbury, & Lefsrud, 2017; Kiritchenko & Mohammad, 2016).

Crowdsourcing platforms like Mechanical Turk and CrowdFlower have presented new methodologies for collecting semantic norms at scale. Crowdsourcing platforms are online services where users post jobs for workers to complete. These jobs consist of a series of simple, well-defined decisions. For example, a user might post a series of pictures of a city taken from sidewalks in a downtown area. Workers would be asked to identify, for each picture, whether it contains a sign for a store. In this example, the labelled pictures might then be used to train a machine learning algorithm to recognize store signs and generalize to a new set of images. This example is a fairly typical use of crowdsourcing in machine vision.

In psychology, crowdsourcing has recently been used to collect estimates of the semantic properties for tens of thousands of words. Participants are presented with a word, a rating scale, and a semantic dimension along which to rate the

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-017-1009-0) contains supplementary material, which is available to authorized users.

✉ Geoff Hollis
hollis@ualberta.ca

[1] Department of Computing Science, University of Alberta, 3-57 Athabasca Hall, Edmonton, AB T6G 2E8, Canada

[2] Department of Psychology, University of Alberta, P217 Biological Sciences Building T6G 2E9, Edmonton, AB, Canada

word. This is not unlike traditional methods for collecting norms in a laboratory setting. The difference is the volume of data that can be collected within a span of time. As example, we recently ran a series of experiments that required participants to make 72,800 individual decisions about words. We estimated that if we were to run this in a laboratory setting, it would require 92 h of labor on behalf of research assistants. Data collection would have taken multiple months to complete, even with motivated assistants. Instead, we deployed the experiments over CrowdFlower and completed data collection in 4 days with minimal intervention from research assistants. Because of the volume of workers available, crowdsourcing can result in data collection efforts taking days whereas otherwise it would take months.

Few people are motivated to make borderline-inane decisions about words out of mere curiosity or goodwill. Crowdsourcing workers need to be reasonably compensated for their work. Herein lies one of the main barriers to collecting large-scale semantic norms: it is a cost-prohibitive endeavor. Very large norms sets such as Warriner, Kuperman, and Brysbaert (2013) and Brysbaert, Warriner, and Kuperman (2014) cost tens of thousands of dollars to collect if workers are adequately compensated for their time. Due to the high costs of such research, there is tangible value to be had in improving the efficiency of data collection methods.

Previous efforts have exclusively relied on ratings scales (e.g., Warriner et al., 2013) or numeric estimation where rating scales are an inappropriate format for data collection (e.g., Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). Very little work has been done in finding ways to reduce measurement error of these methodologies in a crowdsourcing environment, or alternatively of devising different ways of collecting human judgments. One exception is the work of Hollis (2017), who developed scoring methods for a response format called best-worst scaling. In a series of simulation experiments conducted by Hollis (2017), best-worst scaling showed promise for being a more efficient response format than rating scales for collecting human semantic judgments (see also, Kiritchenko & Mohammad, 2016, 2017).

In best-worst scaling, participants are presented with N items and instructed to choose the superior and the inferior of the N items along some specified latent dimension. For instance, in a valence estimation task, participants might be shown the words *mold*, *love*, *soup*, and *phone* and have to choose the most and least pleasant words among the set. N is typically four.

The main value of best-worst scaling comes from the fact that it generates implicit rank information in multiple ways. With two decisions, rank information is available for five of the six pairwise items in a particular trial of four items; if *love* is chosen as most pleasant and *mold* as least, it can be determined that *love* is more pleasant than *soup*, *phone*, and *mold*, that *phone* is more pleasant than *mold*, and that *soup* is more pleasant than *mold*. The only pair for which no information is available is the pair of unchosen words, *soup* and *phone*. Additional implicit information becomes available when multiple trials share at least one word. Suppose a second trial contains the words *bladder*, *balloon*, *phone*, and *glasses*. Although *bladder*, *balloon*, and *glasses* never appeared with *soup*, *mold*, or *love*, information about the rank relationships of these sets of words is still available indirectly through their known relationships to the word, *phone*, which did occur in both trials.

The primary contributions of Hollis (2017) were in identifying and assessing methodological considerations for using best-worst scaling. This included how to construct sets of N items to provide maximal information per decision, and how to use results of best-worst trials to place items along an ordinal scale that is informative of the underlying latent dimension that judgments are being made over. These best practices were inferred from the results of simulation experiments. Hollis (2017) did provide some preliminary empirical evidence to validate his findings from simulated experiments, but compelling empirical evidence supporting best-worst scaling as a useful judgment format for collecting semantic norms is still lacking (however, see Kiritchenko & Mohammad, 2017, who find that split-half reliabilities of sentiment annotations performed with best-worst scaling are higher than if rating scales are used). The purpose of this research was to provide further validation of best-worst scaling as a useful judgment format for collecting semantic norms. This is primarily accomplished by comparing the quality of norms collected with best-worst scaling to norms collected with other response formats. Judgment formats are compared for multiple semantic dimensions that are of interest in psycholinguistic research, and for which large norms sets are already available: valence, arousal, concreteness, and age of acquisition.

Method

Stimuli The 1,034 words from the Affective Norms for English Words (ANEW) norms (Bradley & Lang, 1999) were used in this experiment, plus an additional six filler words to raise the stimulus count to a round number divisible by four. Divisibility by four was required, because the words were randomly sampled to form lists of 4-tuples and each word needed to appear an equal number of times. For 1,040 words, 260 4-tuples allowed for full coverage over the entire word list. This process was repeated 128 times to create the master stimulus set over which participants would make best-worst judgments. Four-tuples were created according to the design suggestions laid out by Hollis (2017): (1) no two words appeared together in more than one 4-tuple, (2) every word appeared in an equal number of 4-tuples, and (3) no 4-tuple was viewed more than once across all participants. These features each increase the amount of information gained per judgment made.

Participants Data collection was carried out under laboratory conditions or by using crowdsourcing software. Laboratory participants were drawn from the undergraduate psychology research pool at the University of Alberta. They received partial course credit for an introductory psychology course in exchange for their effort. Crowdsourced participants were acquired via CrowdFlower. The sampling frame was limited to users who spoke English and resided in Canada or the USA.

Apparatus In laboratory conditions, stimuli were presented on Apple G4 Macintosh Minis. Three separate computers were used to run the experiment. Two had 17.1-in. Samsung Syncmaster 713v monitors. The third had a 17.1-in. BENQ FP767 monitor. All screen resolutions were set to 1,280 × 1,024 and all text was presented in 36-point font. In crowdsourcing conditions, text was likewise presented in 36-point font. Display conditions would depend on the computer used by participants to access CrowdFlower.

Procedure The data collection procedure varied slightly depending on whether data collection was conducted in the laboratory or through crowdsourcing**.** Data for valence and age of acquisition norms were collected under laboratory conditions. Data for arousal and concreteness norms were collected via crowdsourcing.

> *Lab procedure*. On each trial, participants were presented with a 4-tuple and instructed to choose the “best” and the “worst” item from the tuple along a described latent dimension. Prior to their first decision, participants were provided a description of the latent dimension they were to make judgments over. If participants tried to select the same item for “best” and “worst”, they were informed that they had to choose different options. Participants saw a total of 260 unique 4-tuples, and were exposed to each word exactly once. Every 52 trials, participants were instructed to take a self-timed break. The entire procedure took approximately 20 min. Exact instructions were modelled after instructions from previously collected norms (detailed below).
>
> *Crowdsourcing procedure*. Participants were presented with a webpage containing a vertically-oriented sequence of 20 4-tuples. For each 4-tuple, participants were instructed to choose the “best” and “worst” item along a specified latent dimension. The latent dimension in question was described in a set of brief instructions that occurred at the very top of the webpage. Participants were paid US$0.20 for completing each page of 20 trials. Each page took approximately 60 s to complete. Thus, the expected wage of participants was US$12/h. Participants could complete up to 15 pages (300 trials) worth of best-worst judgments per latent dimension.

Examples of the presentation format for a trial, both lab-based and crowdsourced conditions, can be found in Fig. 1.

Latent dimensions and instructions Participants made judgments over one of four latent dimensions: valence, arousal, concreteness, or age of acquisition. In each case, instructions were modelled off the instructions used to collect previous norms (i.e., Brysbaert, Warriner, & Kuperman, 2014; Kuperman et al., 2012; Warriner et al., 2013). Exact instructions are supplied in Appendix A. Valence and age of acquisition data were collected under laboratory conditions. Arousal and concreteness data were collected via crowdsourcing.

Detecting noncompliance It was determined that a small portion of both lab-based and crowdsourced participants produced noncompliant behavior. For laboratory participants, following the procedure described in Hollis (2017), noncompliance was detected by first using all raw trial data to calculate item scores for whichever latent dimension was being analyzed. After item scores were calculated, a participant’s proportion of “best” and “worst” choices consistent with item scores were calculated. If a participant was making choices randomly, their “best” and “worst” choices should be consistent with expectations 50 % of the time. Given that each participant completed 260 trials, the 95 % confidence interval for random guessing would be ± 6.54 %. Thus, we marked any participant with a compliance rate less than 56.54 % as

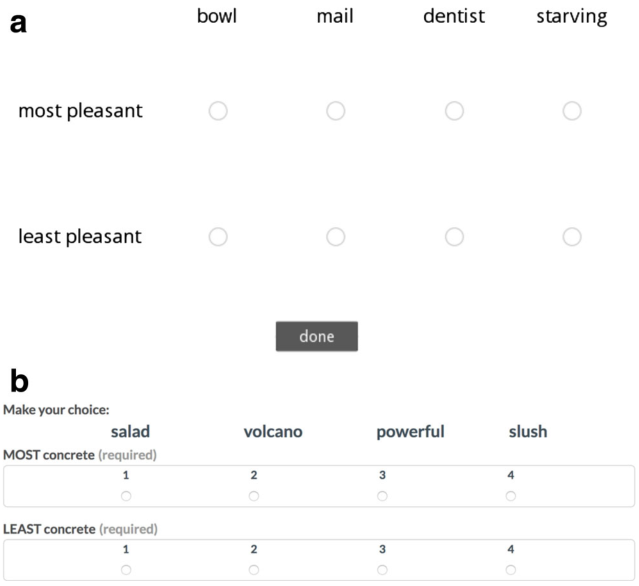

Fig. 1 Example trials for (**a**) lab-based and (**b**) crowdsourced data collection. In the task instructions, participants would be told to choose the “best” item and “worst” item according to their ordering along a specified latent dimension (e.g., most pleasant and least pleasant item)

"noncompliant." Most laboratory participants had very high compliance rates (median compliance = 96.92 %). A total of two participants had their data removed from the valence trials (compliance = 50.00 %, 53.07 %). One participant had their data removed from the age of acquisition trials (compliance = 53.46 %). After the data of noncompliant participants were removed, scores were recalculated with the remaining data.

CrowdFlower offers built-in tools for detecting participant noncompliance. For each latent dimension, a set of 42 test trials were created. For each of these test trials, the two most plausible options for "best" and two most plausible options for "worst" responses were hand-flagged by the researcher based on intuition. The first time a participant performed the task (i.e., a page of 20 4-tuples), they would see eight of these test trials randomly. Each additional time they performed the task, they would see one new random test trial. If a participant's answers on the test trials conflicted with the predetermined plausible options for more than 30 % of cases (default threshold suggested by Crowdflower), the participant was flagged as noncompliant. Noncompliant participants were prevented from contributing any further judgments, their data were filtered prior to analyses, and new participants were recruited to recollect any data that the noncompliant participant had provided.

Crowdflower provides an interface for viewing test trials and the distribution of participant responses to them. After every 1,000th trial for the first 5,000 trials, a researcher updated the plausible options for each test question to be consistent with the majority answers from crowdsourced participants. This was done to ensure that the selection of best-worst items was not biased by the intuitions of the researcher but rather reflected the sample average. In practice, very little adjustment of the test trials was required; participants were clearly performing the task as expected (compliance > 90 %) or clearly producing noncompliant behavior (compliance < 60 %). However, the noncompliance rate was high (roughly 30 % of participants). Because of high noncompliance rates, verification is strongly recommended for crowdsourcing human judgments.

Data volume For valence estimates, data from 110 participants were collected (n=28,600 trials). For age of acquisition estimates, data from 70 participants were collected (n=18,200 trials). Both of these participant counts are prior to filtering participants based on noncompliance. For each of the crowdsourced dimensions (concreteness, arousal), 18,200 trials from compliant participants were collected (i.e., the equivalent of 70 participants worth of data in the laboratory conditions). These data came from 197 and 166 unique participants, respectively.

Scoring best-worst data A detailed explanation of the various ways to convert best-worst data into estimates of latent dimensions (i.e., scoring methods), along with their respective benefits and shortcomings, is provided by Hollis (2017). For the current analysis, each of the five scoring methods described in Hollis (2017) were applied to each of the four latent dimensions being estimated. The method that produced scores most strongly correlated with lexical decision times from the English Lexicon Project (Balota et al., 2007) was selected as the appropriate scoring method for a particular latent dimension.

For valence, arousal, and age of acquisition, a method called *value scoring* presented itself as the most appropriate scoring method. Value scoring is a method closely related to the Rescorla–Wagner update rule (Rescorla & Wagner, 1972), that assigns values to words based on the outcome of a best-worst trial. A value of 1 is assigned to a word for being identified as ordinally higher compared to another item in a particular data point, and 0 for being identified as ordinally lower compared to another item in a particular data point. A word's expected value is then updated based on the observed discrepancy between current empirical value and previously learned expected value. Each data point allows for ten updates – two updates (one a 0, one a 1) for each of the five of six-word pairs for which ordinal information is available in a single best-worst judgment.

For concreteness, the best performing scoring measure was *the analytic best-worst solution*. It is thus used for reporting concreteness results. The analytic solution assigns a value to each word as log[(1 + the unit normalized best-worst score)/(1 - the unit normalized best-worst score)], where an item's best-worst score is the proportion of times an item was chosen as the worst option subtracted from the proportion of times it was chosen as the best option.

These results corroborate Hollis (2017), who found that value scoring generally produces scores that best capture variability of an underlying latent dimension, but that other scoring methods sometimes give better results for certain types of latent distributions and noise levels. We note that the differences in performance between the various scoring methods were quite marginal (tenths to hundredths of a point of variance accounted for) and any could have been used with negligible loss in quality of estimates.

Results

Descriptive statistics

We start by providing descriptive statistics for each of the best-worst norms. Histograms for each of the four variables are presented in Fig. 2. Means, standard deviations, and ranges are presented in Table 1.

Readers will note that values for the different variables are on different scales. This is because there are a variety of ways

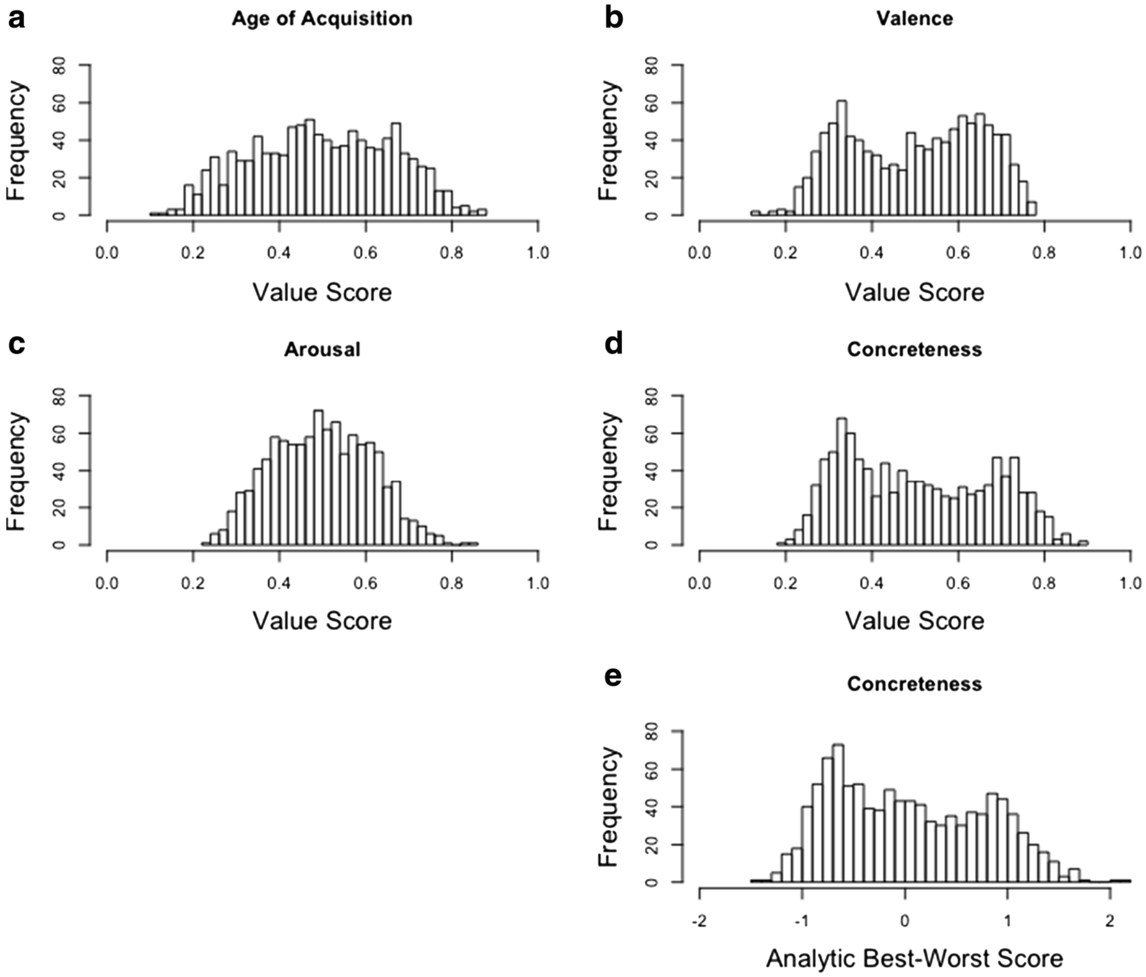

Fig. 2 Histograms for (**a**) age of acquisition, (**b**) valence, (**c**) arousal, and (**d**) concreteness, when best-worst data are scored using value scoring. Also present are (**e**) concreteness, when best-worst data are scored using the analytic best-worst (ABW) method. Concreteness ABW scores better fit behavioral measures of lexical processing than did value scores, and are thus used for analyses. Concreteness value scores are presented in (**d**) to place the variable on a common scale with the other three variables, to aid in comparisons

to score best-worst data, and each scoring method places items along scales of different ranges. Three methods described by Hollis (2017) place items along a scale of (-∞, ∞): Elo, best-worst scoring, and analytic best-worst scoring. Two methods place items along a scale of [0, 1]: Rescorla-Wagner scoring and value scoring. Because concreteness norms were scored using the analytic best-worst method (it was determined this method provided the best fit to behavioral measures of lexical processing), their scale differs from the other three variables, which were all scored with value scoring. To facilitate comparisons between variables, concreteness values are also presented when data were scored using value scoring.

Table 1 Descriptive statistics for age of acquisition, valence, arousal, and concreteness when best-worst data are scored using value scoring. Descriptive statistics are also presented for concreteness when best-worst data are scored using the analytic best-worst method (ABW). Concreteness ABW scores better fit behavioral measures of lexical processing than did value scores, and are thus used for analyses. Concreteness value scores are presented to place the variable on a common scale with the other three variables, to aid in comparisons

Variable	Mean	SD	Min	Max
Age of acquisition	0.499	0.161	0.118	0.873
Valence	0.497	0.153	0.125	0.772
Arousal	0.499	0.114	0.236	0.859
Concreteness (value score)	0.506	0.168	0.183	0.896
Concreteness (ABW)	0.035	0.729	-1.493	2.155

Comparing best-worst scaling to alternative response formats

We were interested in how norms collected through best-worst scaling compared in quality to norms collected with other response formats. We compared best-worst scaling norms to previous 9-point rating scale valence and arousal norms collected by Warriner et al. (2013), 5-point rating scale concreteness norms collected by Brysbaert et al. (2014a), and numeric estimation age of acquisition norms collected by Kuperman et al. (2012). Generally, there was high correspondence between best-worst norms and their comparison norms. Correlations were as follows: age of acquisition r(996) = 0.91 (p < 2.2e-16), arousal r(1036) = 0.77 (p < 2.2e-16), valence r(1036) = 0.96 (p < 2.2e-16), and concreteness r(1020) = 0.93 (p < 2.2e-16).

Despite high inter-norm correlations, best-worst norms were differentiated from their comparison norms in numerous quantitative ways (described below). We were primarily interested in three considerations when comparing the current norms to previously collected norms: how the current norms compare to previously collected norms in terms of (1) their ability to predict behavioral measures of lexical processing, (2) the degree to which they provide an accurate measure the underlying construct of interest, and (3) the cost efficiency of data collection.

For purposes of readability, we start by presenting detailed analyses for just age of acquisition. We then present analyses of the other variables in condensed form, following the same style as age of acquisition. We chose to lead with age of acquisition because it provided the most highly contrastive case compared to previously collected norms for the same construct. Following these comparisons, we provide additional analyses of all four variables in terms of where along their range of behavioral measures norms diverge most in their predictive validities.

Predictive validity: Age of acquisition

A log transformation of age of acquisition values provided an improved fit to behavioral measures of lexical processing, both for best-worst data and numeric estimation data. All results pertaining to acquisition data therefore use log-transformed values.

Analyses were limited to the intersection of the word list used for the current research and the word list used in Kuperman et al.'s (2012) age of acquisition norms. Over these 988 words, our estimates correlated with lexical decision reaction times (LDRTs) from the English Lexicon Project (Balota et al., 2007) at r(986) = 0.66 (p < 2.2e-16), accounting for 43.56 % of the variance in LDRTs. In contrast, the Kuperman et al. norms correlated with LDRTs at r(986) = 0.59 (p < 2.2e-16), accounting for 34.86 % of the variance in LDRTs. A comparison of correlations through Fisher's r-to-z transformation indicated that these coefficients are reliably different in magnitude (z = 2.56; p = 0.005). Thus, our age of acquisition norms have higher predictive validity than those of Kuperman et al. (2012) when no other variables are taken into consideration.

Variables that predict lexical access measures are ubiquitously inter-correlated. A more restrictive, and perhaps informative, test of the predictive validity of two competing norms sets is their ability to account for variation in lexical access measures beyond the influence of other variables, including other norms measuring the same construct. We constructed regression models of LDRTs that included six variables that are commonly studied in the field of psycholinguistics and that are available for a wide range of words: word length, orthographic neighborhood size, word frequency, valence, arousal, and concreteness. Word frequencies were taken from the Twitter frequency norms (Herdağdelen & Marelli, 2017). Word frequencies calculated from social media have recently been demonstrated to account for more variance in lexical decision tasks than frequency measures calculated from other sources (Herdağdelen & Marelli, 2017). Valence and arousal measures were taken from Warriner et al. (2013). Concreteness measures were taken from Brysbaert et al. (2014, b). We additionally included in a hierarchical fashion, either the Kuperman et al. (2012) age of acquisition norms, or our best-worst norms, followed by the other age of acquisition measure left out in the previous step. We then performed model comparisons between models containing one or two age of acquisition norms to determine the extent to which either age of acquisition norm accounted for variance in LDRTs above and beyond other age of acquisition norms and the other predictors listed above.

The best-worst age of acquisition norms account for additional variance in lexical decision times, above the age of acquisition norms of Kuperman et al. (2012) and the other six variables considered in this analysis (F[1,969] = 30.57, p = 4.14e-10). However, the Kuperman et al. (2012) age of acquisition norms do not account for any unique variance above the best-worst norms and the other six variables considered in this analysis (F[1,969] = 2.71, p = 0.10). By this highly restrictive test, the best-worst age of acquisition norms have superior predictive validity compared to those reported by Kuperman et al. (2012).

The previous analysis was repeated for word naming times from the English Lexicon Project as well as lexical decision times from the British Lexicon Project (Keuleers, Lacey, Rastle & Brysbaert, 2012). In both cases, the same pattern was seen: reliable unique effects for the best-worst age of acquisition norms (F[1,969] = 7.37, p = 0.007; F[1,755] = 29.95, p = 1.65e-7, respectively), and no unique effects for the Kuperman et al. norms (F[1,969] = 0.215, p = 0.643; F[1,755] = 2.40, p = 0.12, respectively).

Measurement efficiency: Age of acquisition

We repeated the correlation analysis from the previous section with randomized subsamples of the entire data. For our data, each participant made judgments on 260 trials, providing full coverage over our word list of 1,040 items. We subsampled data for 4, 8, 16, 32, or 64 participants (of 69 participants total), calculated age of acquisition scores, and correlated scores from the subsampled data with LDRTs. The procedure was repeated 100 times and LDRT correlation results were averaged within each subsample size. The central question of measurement efficiency, as it pertains to issues like the cost of running an experiment, is how much data (in this case, how many individual judgments) are required to produce sufficient quality estimates of a latent value. We used the trial-to-word ratio (i.e., the number of trials divided by the number of words) as an index of data collection cost. Since each participant sees all (1,040) words once in 260 trials, the subsample sizes correspond to 1, 2, 4, 8, and 16 trials per word. For example, with four subjects 4 × 260 = 1,040 trials of data are available. Norms are being collected for 1,040 words, so in this example one trial of data is collected for every word for which norms are being collected, hence the trial-to-word ratio is 1:1. If two data collection methods produce norms of different predictive validity when the trial-to-word ratio is held constant, the method that produces the norms with higher predictive validity is the more efficient method at that particular volume of data collection.

We requested the raw numeric estimation data from Kuperman et al. (2012) and performed a comparable subsampling analysis on it. Over the words shared between these two datasets, their median number of responses per word was 20. We created random subsamples of their raw data such that each subsample had 1, 2, 4, 8, or 16 trials per word contained within it. This procedure was repeated 100 times. Each time, age of acquisition estimates were correlated with LDRTs and correlations were averaged for each subsample size.

Results of the subsampling analyses are plotted in Fig. 3a. For trial-to-word ratios of 4:1 or more, error bars denoting the sample error of the mean were too small to be visible. Thus, error bars are excluded from the figure. Two relevant points are evident from this figure. First, it is evident that best-worst scaling produces norms with higher predictive validity than numeric estimation across the range of trial-to-word ratios. Second, 16 trials per word appears sufficient to collect asymptotically maximum quality age of acquisition estimates using best-worst scaling.

Further comparison of measurement instrument validity: Age of acquisition

There is a quantitatively large gap in the predictive validity of our norms (accounting for 43.56 % of the variance in LDRTs) and those of Kuperman et al. (2012; 34.86 % of the variance). Although we do not have enough data to accurately extrapolate where the asymptotic maximum predictive validity of the numeric estimation norms will be, a visual inspection of Fig. 3a does suggest it will be a lower value than the asymptotic maximum of the best-worst norms. This is a particularly worrisome observation, as it suggests differences between the two measurement instruments is not merely a matter of measurement error, in which case we would see the same asymptote with different convergence rates. Rather, there is a discrepancy of *what is being measured* between the two measurement instruments.

We started our analysis by correlating each of the two norms sets with the orthographic variables (word length, word frequency, neighborhood size) and semantic variables (valence, arousal, concreteness) used in the section on predictive validity above to attempt to identify differences in the correlations of these norms with these measures. Results are presented in Table 2. The most notable difference between the two norms is in terms of how strongly they correlate with word length. Our norms correlate more strongly with word length (r[996] = 0.502) than do Kuperman et al.'s (r[996] = 0.436; r-to-z test: z = 1.89; p = 0.03). Shorter words are less common (r[35513] = -0.408), have smaller neighborhood sizes (r[70165] = -0.520), and are less pleasant (r[13,921] = -0.025), more arousing (r[13,921] = 0.108), and less concrete (r[39,952] = -0.292 than longer words. These patterns are consistent with the small differences in correlation strength and relative direction change between the two norms sets on those same five variables. Thus, it appears that our norms are more strongly influenced by word length than those of Kuperman et al. (2012).

It was unclear whether or not "more strongly influenced" indicates a source of undesirable bias. Early-acquired words do tend to be shorter than late-acquired words. It may be that numeric estimation is deficient in its sensitivity to this factor and that best-worst scaling more accurately captures the true relationship between age of acquisition and word length. Further clarification can be found by comparing the two age of acquisition norms in question to alternative norms that rely on a measure more closely coupled to the actual event of learning a word's meaning (henceforth collectively referred to as *test-based norms*; Brysbaert & Biemiller, 2017), rather than the retrospective estimation of adults.

We considered multiple sources of test-based age of acquisition norms for this comparison (see Brysbaert & Biemiller, 2017, for full descriptions and data). The Goodman, Dale, and Li (GDL, 2008) norms contain 562 items of parent-provided responses of whether or not their toddlers know words, and how old those toddlers are. Morrison, Chappell, and Ellis (MCE, 1997) report picture-naming age norms for 297 pictures. The American Teacher's Vocabulary List (ATVL; https://www.flocabulary.com/wordlists/) provides the approximate grades (K-8) at which 1,461 words are reliably encountered in grade-appropriate readings and grade-appropriate state exams. Finally, Dale and O'Rourke (DO, 1981) provide a 44,000-item dataset reporting at which grade level the majority of children know the meaning of an item, as tested with a three-item multiple choice test.

If the AoA estimates collected in the current experiment are "biased by" rather than "accurately influenced by" word length, we should expect this to be expressed by weaker

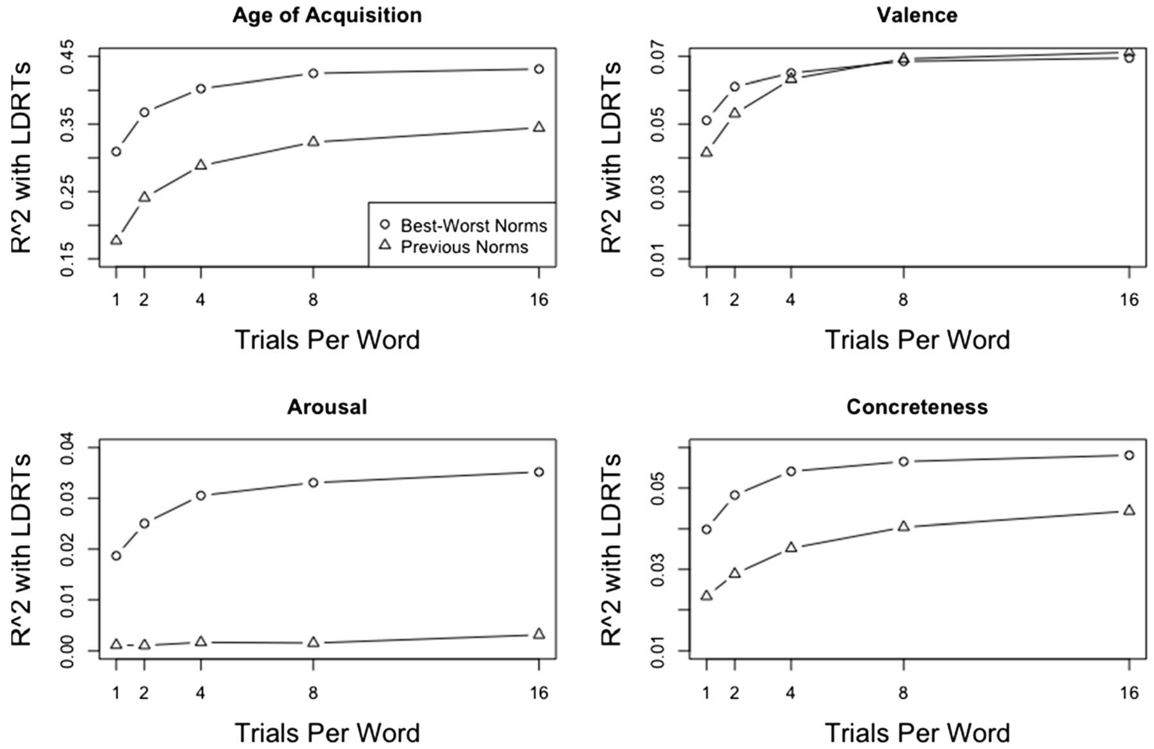

Fig. 3 Comparison of norms on their ability to predict lexical decision reaction times (LDRTs) from the English Lexicon Project (Balota et al., 2007). Variables described from left to right and top to bottom are (**a**) age of acquisition, (**b**) valence, (**c**) arousal, and (**d**) concreteness. Previous comparison norms are from Kuperman et al. (2012), Warriner et al. (2013), Warriner et al. (2013), and Brysbaert et al. (2014, b), respectively

correlations with the above four age of acquisition norms, compared to the norms of Kuperman et al. (2012). Little evidence was found for a difference in relationship strength between test-based age of acquisition norms and either the Kuperman et al. norms or our current norms. There was no difference in correlation strength with the GDL norms (current $r[137] = 0.438$, Kuperman et al. $r[137] = 0.394$; r-to-z $p > 0.05$), the MCE norms ($r[76] = 0.746$ vs. $r[76] = 0.730$; $p > 0.05$), the ATVL norms ($r[120] = 0.788$ vs. $r[120] = 0.738$; $p > 0.05$), or the DO norms ($r[970] = 0.673$ vs. $r[970] = 0.668$; $p > 0.05$).

These comparisons were repeated on the residuals of the test-based age of acquisition norms after accounting for effects of word length. Again, no difference in relationship strength was observed for the GDL data ($r[137] = 0.416$ vs $r[137] = 0.386$; $p > 0.05$), the MCE data ($r[76] = 0.661$ vs. $r[76] = 0.638$; $p > 0.05$), the ATVL data ($r[120] = 0.734$ vs. $r[120] = 0.703$; $p > 0.05$), or the DO norms ($r[970] = 0.583$ vs. $r[970] = 0.594$; $p > 0.05$). Even when first accounting for the relationship between test-based age of acquisition and word length, the current norms and the Kuperman et al. (2012) norms showed no difference in predictive validity of test-based age of acquisition measures. We thus conclude that by the current criterion (prediction of test-based measures of age of acquisition), both norms sets have equivalent validity. This holds true even in the case that most differentially disadvantages the current norms: first factoring out effects of word length.

Interim summary

The above analysis demonstrates that, in the case of measuring age of acquisition through adult retrospective estimates, best-worst scaling has multiple desirable properties over

Table 2 A comparison of two age of acquisition (AoA) norms on their correlational structure with six other variables. Word frequencies were obtained from Herdağdelen and Marelli (2017), valence and arousal measures from Warriner et al. (2013), and concreteness measurements from Brysbaert et al. (2014, b). Orthographic neighborhood sizes (Orthographic N) were calculated from the 78,000-word dictionary of Hollis, Westbury, and Lefsrud (2017)

AoA Norms	Comparison measure					
	Length	Log frequency	Orthographic N	Valence	Arousal	Concreteness
Hollis & Westbury	0.502	-0.668	-0.423	-0.363	0.211	-0.375
Kuperman et al.	0.436	-0.654	-0.394	-0.321	0.188	-0.418

numeric estimation. Best-worst scaling produces norms with higher predictive validity on lexical access tasks. The methodology also requires substantially fewer judgments to be made to produce norms of equivalent predictive validity. Both methods appear to produce norms of equivalent construct validity when validated against objective measures of age of acquisition.

Results from the subsampling analysis in the measurement efficiency section are troubling. The fact that the asymptotic predictive validity of best-worst scaling and numeric estimation norms appear to be different values suggests that the two measurement instruments are in fact measuring different things. If the two measurement instruments differed only in measurement error, they would have the same asymptotic predictive validity, but differing convergence rates. To complicate matters, comparisons involving test-based measures of age of acquisition did not differentiate best-worst norms from numeric estimation norms. Finally, we observed that best-worst norms correlated more strongly with word length than the numeric estimation norms did. However, the previous tests do not provide strong evidence one way or the other as to whether this is due to bias in the best-worst response format, or due to the fact that the best-worst format better captures the true relationship between age of acquisition and word length. This result points to the importance of not relying on a single measurement instrument for collecting semantic norms. Multiple, different, instruments should be employed. This will help with identifying idiosyncrasies specific to one particular instrument.

Predictive validity

Best-worst norms for valence, arousal, and concreteness correlated with LDRTs from the English Lexicon Project at $r(1024) = -0.265$, $r(1024) = 0.164$, and $r(1010) = -0.227$, respectively (all $p < 0.05$). For comparison, valence, and arousal measures from Warriner et al. (2013), and concreteness measures from Brysbaert, Warriner, and Kuperman (2014) correlated with LDRTs at $r(1024) = -0.266$, $r(1024) = 0.036$, and $r(1010) = -0.206$, respectively. Of these correlations between rating scale norms and LDRTs, valence and concreteness statistically reliably correlate with LDRTs (both $p < 0.05$), but arousal does not ($p > 0.05$). Best-worst arousal norms more strongly correlated with LDRTs than did the Warriner et al. (2013) arousal norms (Fisher r-to-z transformation $z = 2.93$, $p = 0.002$). In the cases of valence and concreteness, there was no difference in the correlation strength with LDRTs between best-worst norms and rating scale norms.

Predictive validities of the various norms were tested using model comparison, following the steps outlined above for age of acquisition. These analyses revealed no differences in the amount of unique variance accounted for in LDRTs by best-worst norms and comparison norms.

Previous research has suggested that affective measures have nonlinear relationships to processing times (e.g., Estes & Adelman, 2008; Vinson, Ponari, & Vigliocco, 2014), for example, because words may be processed faster the more affectively laden they are, regardless of the polarity of affect (Kousta, Vinson, & Vigliocco, 2009). By including only linear terms in regression analyses, we miss the possibility of differentiating norms based on nonlinear relationships. We repeated the above comparisons of valence and arousal norms and allowed for possible nonlinear fits using restricted cubic splines. Analyses were conducted using R's *gam* function from the *mgcv* package, applying the *s()* operator to fit a spline.

The model including a spline for rating scale valence norms, plus other variables entered (described in analyses of age of acquisition), accounted for 58.2 % of the variance in LDRTs (AIC = 10,437.92). The model including a spline for best-worst valence norms accounted for 58.2% of the variance in LDRTs (AIC = 10,438.79). The two models did not substantially differ in their ability to reduce information loss in LDRTs (relative likelihood = 0.65). The model including splines for both norms also accounted for 58.2 % of the variance in LDRTs (AIC = 10,439.84). It did not substantially differ in its fit from the rating-scale-only model (relative likelihood = 0.38), nor the best-worst-only model (relative likelihood = 0.59). AIC values for the arousal models were as follows: using rating scale only AIC = 10,437.92, using best-worst only AIC = 10,438.93, and using both norms AIC = 10,439.91. Like with the comparison of valence norms, none of these models substantially differed in fit based on relative likelihoods. We repeated these analyses for the concreteness norms: rating scale model AIC = 10,437.92, best-worst model AIC = 10,436.44, and the model using both splines AIC = 10,438.29. Again, these models were not differentiated based on relative likelihoods. We are left with the conclusion that, by our criterion of being able to predict unique variance in lexical decision times while accounting for effects of other relevant variables, best-worst norms and rating scale norms for valence, arousal, and concreteness measures do not differ substantially.

Response times from the Calgary semantic decision project, where participants made concrete/abstract decisions to 10,000 words, are more sensitive to differences in word concreteness than are lexical decision times (Pexman, Heard, Lloyd, & Yap, 2017). If there are differences in the predictive validities of concreteness norms, those differences might be made more apparent when predicting response times from concrete/abstract decision data rather than lexical decision data. Using only linear terms, model comparison revealed that best-worst norms accounted for 2.92 % of the variance in response times beyond rating scale norms (F(1,361) = 17.934, $p < 2.9e-5$). Rating scale norms accounted for no unique variance beyond best-worst norms (F(1,361) = 0.265,

p = 0.61). Likewise, when restricted cubic splines were used to fit for nonlinear effects of concreteness, the model using only rating scale norms minimized information loss less (AIC = 4,446.64) than did the model using only best-worst norms (AIC = 4,419.25) or both norms (AIC = 4,418.73). Comparing relative likelihoods, the best-worst model is 885,139 times more likely to reduce information loss than the rating scale model and the best-worst + rating scale model is over a million times more likely to reduce information loss than the rating scale model. However, adding rating scale norms to the best-worst model does not further reduce information loss by an appreciable amount (1.29 times). These analyses provide evidence that best-worst scaling can produce concreteness norms with higher predictive validity than rating scales, when abstract/concrete semantic decision latencies are considered.

Measurement efficiency

The subsampling analysis presented for age of acquisition was repeated for each of valence, arousal, and concreteness. Because arousal and concreteness norms were collected online via CrowdFlower and not every participant saw every word once, an additional step was needed to perform the analysis. Trials were assigned to groups of 260, corresponding to a complete list of every word appearing once. Data from these complete lists were then sampled identically to how data from participants were sampled in the age of acquisition section. Since valence data were collected in a laboratory environment where each participant saw 260 trials and each word once, valence data were subsampled identically to age of acquisition.

Results from the subsampling analyses can be found in Fig. 3. Similar to age-of-acquisition results, best-worst scaling consistently results in more efficient data collection than rating scales when collecting arousal and concreteness data. This is the case for every trial-to-word ratio. This was statistically verified by performing t-tests on the r-squared values for the 100 subsampled datasets of best-worst norms and comparison norms. In each case, best-worst r-squares with LDRTs were reliably higher than those for comparison norms ($p < 2.26e\text{-}16$ in every case). For valence, best-worst scaling provided better fitting norms when subsamples contained one, two, or four trials per word (Welch's t[181.86] = 7.05, $p < 3.63e\text{-}11$; Welch's t[176.85] = 6.93, $p < 7.43e\text{-}11$; Welch's t[192.25] = 2.10, p = 0.04). No difference in fit was present when eight trials per word were subsampled (Welch's t[176.55] = 1.17, p = 0.24). Rating scales provided a superior fit when 16 trials per word were subsampled (Welch's t[121.04] = -5.97, $p <$ 2.43e-08).

As with age of acquisition, Fig. 3 suggests different asymptotic predictive validity with LDRTs for concreteness and arousal, depending on whether best-worst scaling or rating scale data are used. In both cases, it appears that best-worst scaling data has higher asymptotic predictive validity than rating scale data. Again, these results suggest that differences between the response formats are not merely about measurement error. Each instrument is producing qualitatively, not quantitatively, different data.

We note a peculiar finding from the arousal subsampling analysis. Arousal estimates from rating scales do not improve as more trials per word are subsampled (Fig. 3c). We should expect norms to have higher predictive validity when more data are used to construct those norms, simply because more data usually implies lower measurement error. This finding suggests that rating scales have no appreciable predictive validity of lexical decision times or, alternatively, rating scales do not elicit stable estimates of arousal. Notably, however, we do see the expected increase in predictive validity as sample size increases for best-worst estimates of arousal, along with both response formats for each of the other variables. This difference between arousal norms will be revisited in later analyses and the discussion.

Further comparison of measurement instrument validity

Whereas objective estimates of age of acquisition are available via test-based norms, no such objective estimates exist (to our knowledge) for valence, arousal, or concreteness; only human judgment norms are available. We are unable to provide further validation for best-worst norms of these three variables in the same way as was done for age of acquisition.

Recently, Hollis and Westbury (2016) have demonstrated that principal component analysis of the skip-gram distributional semantic model produces principal components that are interpretable as human-relevant semantic dimensions. Among them are dimensions interpretable as valence and concreteness, along with other dimensions that also suggest themes of agency, edibility, gender, among others. Semantic norms play a critical role in being able to objectively determine that a learned dimension in a computational model maps onto a semantic concept that a human can articulate.

More generally, semantic dimensions learned by the skip-gram model can be used to estimate human semantic judgments for valence, arousal, dominance, and concreteness with accuracies that approach norm split-half reliabilities (Hollis, Westbury, and Lefsrud, 2017). This implies that the model is indeed identifying variability in word use that maps onto semantic concepts that guide human selection over word use.

Often within cognitive psychology, there is a degree of separation between the thing that is measured (e.g., human judgments of arousal) and the object of explanation (e.g., the construct of arousal). Sound explanation requires us to map the object of explanation into a simpler domain (e.g., a co-varying feature of a simple computational model). Depending on how an object of explanation is operationalized (e.g., with best-worst scaling or rating scales), the window that operationalization provides into the object of explanation may be more or less clear.

Insofar as a domain of explanation is valid to use, a construct is explainable, and different operationalizations of that construct differ only in measurement error, the extent to which an operationalization can be mapped into the domain of explanation is a comparative indicator of the appropriateness of the method of operationalization.

Assuming that (1) distributional semantic models provide an objective but incomplete basis of semantics, (2) constructs of concreteness, valence, and arousal are themselves aspects of semantics, and (3) best-worst scaling, rating scales, and numeric estimation are all reasonable means of operationalizing the underlying construct they reference, then the extent to which such operationalizations can be accurately predicted by a distributional semantic model is an indicator of the appropriateness of the particular operationalization.

Hollis and Westbury (2016) found that valence estimates derived using rating scales correlated maximally with principal component 5 (PC5) in their skip-gram model. Concreteness and arousal both correlated maximally with principal component 2 (PC2). Within the skip-gram model, these principal components appear to provide the basis for constructs of valence (PC5), concreteness, and arousal (PC2). Each of the three variables correlated broadly and weakly with a large number of other principal components, but never as strongly as with the principal component they correlated maximally with.

We performed two tests of rating scale norms and best-worst scaling norms. The first test compared correlation strengths of the norms with the principal component with which they shared the most variance, as reported by Hollis and Westbury (2016). To the extent that a response format provides a reasonable estimate of an intended underlying construct, norms collected using that response format should be predictive of the corresponding principal component in the model of Hollis and Westbury (2016). The second test compared the distributions of correlation strengths across all 300 principal components from Hollis and Westbury. To the extent that a response format provides measurements that are carrying semantic information, norms should correlate with the range of principal components extracted by Hollis and Westbury (2016) from their model.

Best-worst norms for arousal more strongly correlated with PC2 than did the rating scale norms reported by Warriner et al. (2013): $r[1,034] = -0.33$ vs. $r[1,034] = -0.19$ ($p = 0.0003$). On average [SE], r-squared values were higher across the range of all 300 PCs for best-worst norms than rating scale norms: average [SD] $r^2 = 6.78\text{e-}03$ [7.84e-04] vs. average [SD] $r^2 = 4.99\text{e-}03$ [5.31e-04] (paired $t[299] = 2.75$, $p = 0.006$). No differences were observed between best-worst and rating scale norms for concreteness or valence ($p > 0.05$ in every case). Best-worst scaling provides measurements of arousal that are better explainable with a computational semantic model than are measurements taken using a rating scale.

Differences in norm prediction errors

With the exception of valence, best-worst and comparison norms were differentiated from each other based on predictive validity. Best-worst norms for age of acquisition and concreteness both had higher predictive validities with a behavioral measure of lexical processing than did comparison norms after accounting for variance due to other factors (LDRTs for age of acquisition, abstract/concrete semantic decision times for concreteness). Best-worst norms for arousal better predicted LDRTs than did rating scale norms for arousal, although this difference disappeared after variance due to other factors was accounted for. Arousal norms were also differentiated by their ability to have their variance explained by a computational semantic model. Best-worst norms were better predicted than rating scale norms, suggesting that best-worst arousal norms are more semantically rich than are rating scale arousal norms. We next report additional analyses that were conducted to examine differences in predictive validities of norms. These analyses compared the magnitude of prediction errors made by different types of norms for relevant behavioral measures of lexical processing.

Both types of age of acquisition, valence, and arousal norms were individually fitted to ELP LDRTs using restricted cubic splines. Both types of concreteness norms were fitted to concrete/abstract decision times using restricted cubic splines. Differences in the magnitude of errors from either type of norm were calculated, as a function of response time. Data are presented in Fig. 4.

For age of acquisition, the average prediction error of best-worst scaling norms is lower than the average prediction error for numeric estimation across the range of ELP LDRTs. The more extreme the response time (either fast or slow), the proportionally better the fit provided by best-worst norms. The same pattern is observed for best-worst versus rating scale concreteness norms regressed on concrete/abstract decision times. However, the reverse pattern is seen for valence norms; rating scale norms have lower prediction error of ELP LDRTs than do best-worst norms for extreme decision times. We performed a t-test to test the hypothesis that difference in prediction error for the two valence norms was not different from zero. The null hypothesis was rejected ($t[1025] = 2.43$, $p = 0.015$), suggesting that rating scale norms have lower prediction error of ELP LDRTs than do best-worst norms. A similar test was conducted for each of the other variables. Each favored the interpretation that best-worst norms have lower prediction error than comparison norms: age of acquisition t(987)

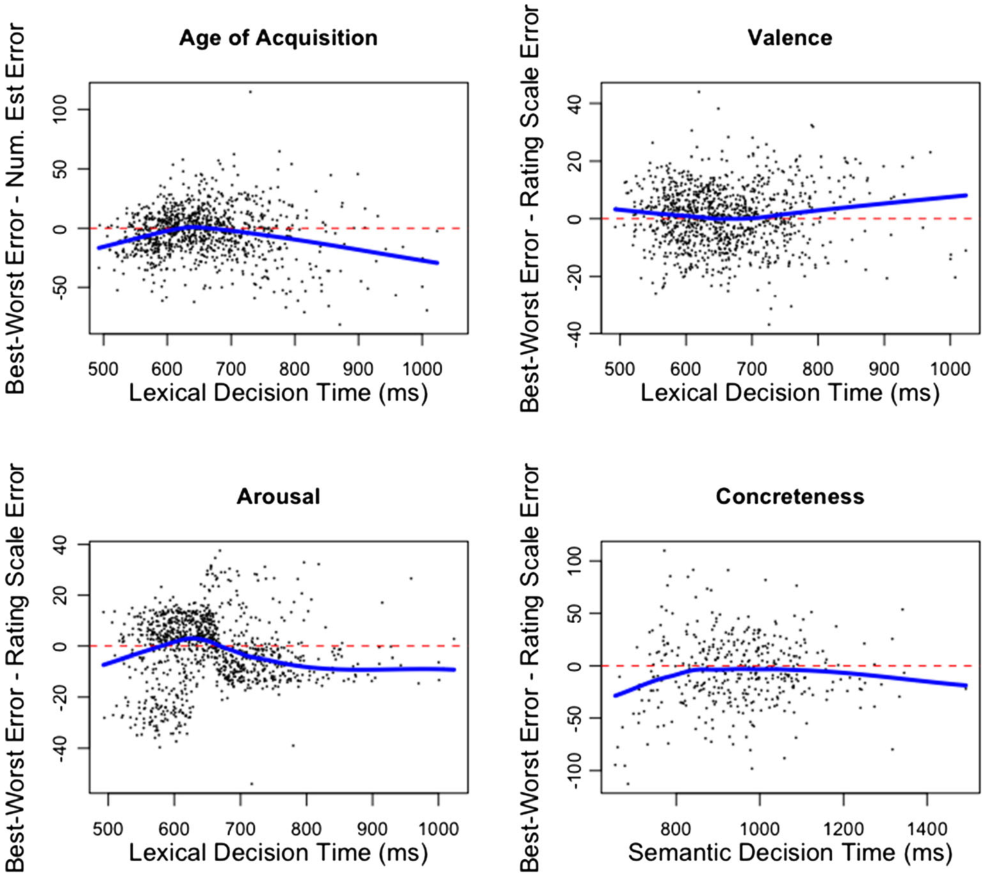

Fig. 4 Difference in absolute residuals after regressing either best-worst norms or rating scale norms (valence, arousal, concreteness) / numeric estimation norms (age of acquisition) on a behavioral measure of lexical processing. Norms for (**a**) age of acquisition, (**b**) valence, and (**c**) arousal were regressed on English Lexicon Project lexical decision times (Balota et al., 2007). (**d**) Concreteness norms were regressed on abstract/concrete semantic decision response times (Pexman et al., 2017). In all cases, restricted cubic splines were used to fit data. Solid blue lines are loess-smoothed fits between behavioral measure (x-axis) and difference in prediction error (y-axis). Dotted red lines indicate the point where there is no difference in prediction error

= -4.88 (p < 1.25e-6), arousal t(1025) = -3.71(p < 2.2e-4), concreteness t(380) = -2.77 (p < 5.8e-3).

An unusual pattern of differences in prediction error was observed for arousal norms. Descriptively, there is a sharp boundary at ~660 ms, below which rating scale norms have lower prediction error for ELP LDRTs, but above which best-worst norms have lower prediction error. Furthermore, there appears to be a smaller subset of words for which the opposite pattern is observed.

We examined words with LDRTs below 640 ms or above 680 ms as a function of whether best-worst scaling or rating scale norms had lower prediction error of LDRTs. First, we note that the difference in prediction error is bimodal both for words with high response latencies (> 680 ms) and low response latencies (< 660 ms). Histograms are presented in Fig. 5. This bimodality suggests that there are different groups of words, for which best-worst scaling and rating scales are differentially sensitive to variability in arousal, as it pertains to lexical decision.

We tested the possibility that words could be clustered by their semantic properties. Words were grouped using k-means clustering over semantic representations derived from a distributional semantic model (Hollis & Westbury, 2016). Fast response latency and slow response latency words were clustered separately. K values were determined according to the elbow method. A k of 8 was chosen for fast response latency words, and a k of 12 was chosen for slow response latency words. We then examined clusters where, among the words contained within them, LDRT prediction errors were more likely to be minimized when using best-worst arousal norms instead of rating scale arousal norms, or vice versa. The rule for determining whether a cluster was examined was: (C_{bw} / N_{bw}) / (C_r / N_r) is greater than 2.0 or less than 0.50. C_{bw} was the number of words within in cluster, C, whose prediction error was best minimized by best-worst scaling and N_{bw} was the number of words across all clusters whose prediction error was best minimized by best-worst scaling. C_r and N_r correspond to similar values for rating scales. Conceptually, this rule is identifying clusters within which words have their prediction error consistently minimized more by one of the two arousal norms. Six clusters were examined for words with fast

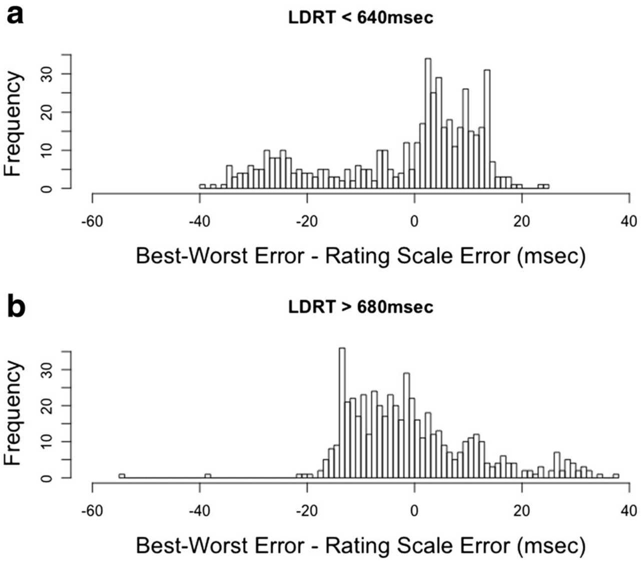

Fig. 5 Difference in prediction error of ELP LDRTs between arousal norms derived from best-worst scaling or rating scales. Positive values indicate lower prediction error for best-worst scaling and negative values indicate lower prediction error for rating scales. Data are broken up by words with response latencies a) less than 640msec or b) greater than 680msec

response latencies and four clusters were examined for words with slow response latencies.

A summary of results is presented in Table 3. The content of each cluster was highly interpretable as organizing around a semantic theme. For example, harm, concrete objects, affective states, or food and animals. Within each cluster, words tended towards being either concrete or abstract. For fast response latency words, rating scale arousal norms of abstract clusters tended towards lower prediction error of LDRTs than did best-worst norms, and best-worst norms tended towards lower prediction error of concrete clusters. For slow response latency words, this pattern reversed: best-worst norms tended towards lower prediction error for abstract clusters and rating scale norms tended towards lower prediction error for concrete clusters. In line with these observations, differences in prediction error between best-worst norms and rating scale norms are correlated with concreteness, but the direction of relationship differs based on whether words are responded to fast or slow. For fast response words, concreteness is negatively correlated with the difference in prediction error between best-worst norms and rating scale norms (r[490] = -0.434, p < 2.2e-16; i.e., best-worst norms have less prediction error for concrete words). However, for slow response words, concreteness is positively correlated with the difference in prediction error between best-worst norms and rating scale norms (r[332] = 0.387, p < 2.3e-13l; i.e., rating scale norms have less prediction error for concrete words). These results may be reflecting different *types* of arousal, for which best-worst scaling and rating scales are differentially sensitive.

Distinguishing between scoring methods and response formats

The reported comparisons of best-worst scaling norms to rating scale or numeric estimation norms have conflated two aspects of the norms creation process: the response format and the scoring method. For example, data can be collected using the rating scale response format. Those data are then typically scored by averaging all values together for a specific item. However, there is no reason why other scoring methods could not be used. For example, taking the median. Hollis (2017) has introduced numerous scoring methods for the best-worst scaling response format. The *value scoring* method typically produces norms that best capture variability along the latent dimension judgments are being made over (Hollis, 2017) and produces norms with the highest predictive validities of lexical processing (here). However, the most appropriate scoring method ultimately depends on idiosyncratic details specific to the latent dimension judgments are being made over, and the reliability of the individual producing judgments (Hollis, 2017). Because the current comparisons conflate response format and scoring method, we cannot be certain whether observed differences of norms are due to the response format used (e.g., best-worst scaling vs. rating scale), the scoring method used (e.g., value scoring vs. averaging), or a combination of the two.

To address the issue of conflating response format with scoring method, it would be informative to compare norms produced by different response formats when the same scoring method is used, or different scoring methods with the same response format is used. Best-worst scaling produces a data type that makes little sense to score with averaging or median values; data are ordinal, which excludes averaging, and the number of unique positions an item can take on is arguably too small (3; best, unchosen, worst) to be able to provide sufficient differentiation of items along the latent dimension that is being judged over. It makes more sense to score rating scale and numeric estimation data using the methods typical to best-worst scaling; items can be grouped into N-tuples and ordered from best to worst based on individual rating scale judgments or numeric estimations provided for those items. Once items are grouped and ordered as such, it is possible to apply the scoring methods of best-worst scaling to such data. Applying such a transformation creates an ordinal data type, which results in a loss of interval information that is available through numeric estimation and rating scale response formats. However, the scoring methods of best-worst scaling accurately reconstruct interval information as long as individual judgments are sufficiently noisy (see Hollis, 2017, for simulation results and discussion); one of the important properties of the scoring methods presented in Hollis (2017) is that they can use ordinal information to appropriately space items along an interval scale.

Table 3 Words with fast (< 640 ms) or slow (> 680 ms) lexical decision reaction times (LDRTs) are differentially predicted by best-worst arousal norms or rating scale arousal norms. This table provides a qualitative description of two cluster analyses, conducted either on words with fast LDRTs or slow LDRTs. Verbal descriptions of cluster content, example members, mean (SD) abstract/concrete values for words in the cluster, and the number of words in the cluster whose LDRTs were best predicted by best-worst arousal norms or rating scale arousal norms. Abstract/concrete values are from the best-worst data presented in this study

Cluster description/examples	Concreteness (SD)	Cluster size	# Words where norm type minimized prediction error	
			Best-worst	Rating scale
Fast LDRT words				
Self; persons, actions, traits *vampire, king, alien, bastard; evil, bored, noisy, cute; kiss, lie, joke, scream*	-0.455 (0.384)	68	10	58
Harm *sickness, toxic, cancer, germs, foul, cut, injury, broken, panic, anxious, fear, dismay*	-0.217 (0.409)	65	10	55
Pleasant concepts *pride, bless, loved, confident, charm, friendly, love, pleasure, victory, loyal dream*	-0.505 (0.417)	73	11	62
Concrete objects; furniture, weapons, bodyparts *chair, door, pencil, lamp; gun, tank, hammer, pistol; hand, chin, breast, elbow*	0.863 (0.381)	67	47	20
Environmental features *swamp, cliff, ocean, sun, farm, lighthouse, hotel, house, taxi, ship, bus, truck*	0.785 (0.388)	52	31	21
Food and animals *wine, eat, muffin, pizza, sugar, dinner, puppy, horse, bird, cow, pet, kitten*	1.105 (0.444)	45	31	14
Slow LDRT words				
Sexuality/intimacy *cuddle, vagina, orgasm, loneliness, spouse, masturbate, ecstasy, alimony*	-0.222 (0.588)	25	22	3
Affective states *weary, overwhelmed, awed, blase, timid, frustrated, upset, satisfied, elated*	-0.783	32	26	6
Negative social affect *scum, shamed, traitor, snob, humiliate, despise, insult, maniac, obnoxious*	-0.612 (0.385)	48	37	11
Physical space: places, vessels, spatial projection *penthouse, corridor, highway; locker, barrel, jug; revolver, lightbulb, whistle*	0.798 (0.371)	82	32	50

We re-scored numeric estimation age of acquisition norms (Kuperman et al., 2012), rating scale valence and arousal norms (Warriner et al., 2013) and rating scale concreteness norms (Brysbaert et al., 2014a) using scoring methods introduced for best-worst scaling (Hollis, 2017). Items were organized into a series of 4-tuples according to the same rules used for constructing best-worst trials in the current experiment (see *Stimuli* subsection of the *Methods* section). Sufficient 4-tuples were created such that each item appeared 256 times (compared to 70 times when best-worst stimuli were constructed for concreteness, arousal, and age of acquisition data, and 110 times for best-worst valence stimuli). Items within each 4-tuple were then ordered from "best" to "worst" by sampling from the empirical distribution of values supplied by corresponding numeric estimation or rating scale judgments. Ties in value were resolved through random choice. The resulting data were then scored by applying each of the scoring methods: Elo, value, Rescorla-Wagner, best-worst, and analytic best-worst (Hollis, 2017).

The first three of these scoring methods frame scoring as a tournament ranking problem, and further break each 4-tuple into a series of 4-choose-2 = 6 paired matches, each with a "winner" and a "loser." Best-worst scaling only provides relational information about five of the six pairs of items in a 4-tuple, so item scores are only updated based on those five pairs (with best-worst scaling, no information is available for the pair of items that are chosen as neither best nor worst). When these scoring methods are applied to numeric estimation and rating scale data, there is no reason not to also update scores based on this sixth pair; relational information about the two items is available from their numeric estimation or rating scale judgments. We thus also update scores based on this sixth pair of items.

In none of the cases examined did re-scoring norms improve the fit of the norms with LDRTs. The fits between re-scored age of acquisition data and LDRTs ranged between r(969) = 0.565 (best-worst) and r(969) = 0.570 (analytic best-worst), compared to r(969) = 0.589 for averaged data. For valence, fits ranged between r(1024) = -0.262 (Rescorla-Wagner) and r(1024) = -0.272 (best-worst), compared to r(1024) = -0.265 for averaged data. For arousal, fits ranged between r(1024) = 0.044 (Rescorla-Wagner) and r(1024) =

0.059 (Elo), compared to r(1024) = 0.036 for averaged data. Finally, for concreteness, fits ranged between r(1010) = -0.202 (Elo) and r(1010) = -0.212 (value), compared to r(1010) = -0.205 for averaged data. We additionally compared concreteness norms on the semantic decision data of Pexman et al. (2017): fits ranged between r(378) = -0.526 (Elo) and r(378) = -0.531 (value), compared to r(378) = -0.512 for averaged data. Comparisons of fit were performed with the Fisher r-to-z transformation, and in all cases, the fit of averaged data was not reliably different than the fit provided by either the best-fit alternative scoring method or the worst-fit alternative scoring method (all $p > 0.05$). We repeated all comparisons, allowing for nonlinear fits with restricted cubic splines. Again, in all cases, insufficient evidence was available to prefer one model over alternative models using a relative likelihood test of model AIC values.

These analyses provide modest evidence that differences observed between best-worst norms and comparison norms are primarily due to the response formats, not the scoring methods.

Discussion

Availability of software and data

Software for creating best-worst stimuli and scoring best-worst accompany this article. The software is also made publicly available on the first author's website (http://www.ualberta.ca/~hollis). Best-worst norms for age of acquisition, valence, arousal, and concreteness reported in this research have also been made available, accompanying this article.

Predictive validity and measurement efficiency of semantic norms

By the criterion of being able to predict behavioral measures of lexical processing, best-worst scaling produces higher quality data than previously used methods for measures of age of acquisition, concreteness, and arousal. However, evidence was also presented that suggests rating scales may be more appropriate than best-worst scaling for the case of measuring word valence; LDRT prediction errors were statistically likely to be better minimized by rating scale norms than best-worst norms.

Best-worst scaling presented itself as a consistently more efficient measurement instrument than alternatives; with smaller sample sizes (one, two, and four trials per word), best-worst norms outperformed numeric estimation norms and rating scales norms at predicting lexical processing times. With larger sample sizes (eight, 16 trials per word), best-worst norms outperformed other methods in all cases except for valence. Rating scales were marginally more efficient at collecting valence data with 16 trials per word. The reported data suggest that eight–16 trials per word is sufficient for norms with asymptotically maximum predictive validity of lexical decision times (Fig. 3). This is lower than current conventions for collecting norms with rating scales and numeric estimation, which typically collect 20–25 trials per word. However, we stress that this is convention and not a demonstrated optimal methodological choice.

If we consider only issues of predictive validity and measurement efficiency, these results suggest that researchers interested in collecting semantic norms may wish to favor the use of best-worst scaling over other response formats in the general case. However, there are some cases (e.g., valence) where rating scales might be the more appropriate measurement instrument. We conclude that best-worst scaling presents itself as a promising and flexible tool for the collection of semantic norms. However, we caution that relying exclusively on best-worst scaling for collecting semantic norms would be shortsighted. Such a restriction would impede understanding of why different measurement instruments provide divergent results (discussed next).

Different response formats measure different aspects of a semantic construct

One of the conclusions that this research points to is that more than just predictive validity and measurement efficiency needs to be considered when assessing the quality of semantic norms. Presented results suggest that best-worst scaling and other response formats have different asymptotic predictive validities for lexical decision times, depending on the underlying latent dimension over which judgments are being made (Fig. 3). This suggests that the different measurement instruments are in fact measuring different things. If they were measuring the same thing, their asymptotic predictive validities would be the same, although their rates of convergence may differ due to differences in measurement error.

A possibility is that differences observed between best-worst scaling norms and alternative norms are not due to the measurement instruments but, instead, due to differences in the samples of participants that generated judgments. For instance, whereas our best-worst age of acquisition norms were generated by undergraduate university students in a laboratory setting, the numeric estimation norms reported by Kuperman et al. (2012) were generated by a demographically much broader sample of the population via crowdsourcing. Improved predictive validities of best-worst norms may be due to being collected from a more relevant sample of the population, as it pertains to psychological research on language processing.

Though not entirely impossible, the "participant differences" explanation seems unlikely. Previous research has already established that crowdsourced norms are comparable in quality to lab-based norms for psycholinguistic research (e.g., Brysbaert et al., 2014, b; Warriner et al., 2013), despite coming from a demographically broader sample of the population. If participant differences were in fact the cause, we should then also expect to see differences in the predictive validities of our lab based best-worst valence norms and the crowdsourced rating scale valence norms of Warriner et al. (2013), but we don't. Likewise, we should not expect to see differences in the predictive validities of our crowdsourced best-worst arousal norms and the crowdsourced rating scale norms of Warriner et al. (2013), but we do. It is unlikely that participant differences are the cause of effects seen in these analyses.

Comparison of arousal norms presented a rather complicated case. Best-worst norms and rating scale norms asymptote to different predictive validities of LDRTs. More strikingly, rating scale arousal norms show no appreciable gain in predictive validity as sample size is increased, contrary to the pattern observed for best-worst arousal norms.

Another finding reported in the above analyses may help illuminate what is happening with arousal rating scale norms. We found that arousal rating scale norms better predicted LDRTs for words that were quickly responded to, but that best-worst norms better predicted LDRTs for words that were slowly responded to. A clustering analysis further revealed that for words with fast LDRTs, best-worst scaling provided better estimates of arousal for concrete words, but rating scales provided better estimates of arousal for abstract words. To add complexity to the issue, this pattern was exactly reversed for words with slow LDRTs.

We put forth the following interpretation of why arousal rating scale norms lack predictive validity of LDRTs but best-worst norms do not: the concept of arousal has multiple aspects to it, existing within the cleaves of concrete or abstract words, and words that are recognized fast or slow. Arousal rating scale norms lack predictive validity because from trial to trial, people are estimating different aspects of arousal and these aspects have different and opposing involvements within lexical processing. In best-worst scaling, multiple words are presented on a single trial. This added context constrains people towards responding with common and shared aspects of arousal; words are more consistently rated along the same aspect of arousal.

The descriptive labels of clusters extracted from the analysis of arousal norms (Table 3) point to possible themes that may tap into different aspects of arousal: person-related arousal (clusters of persons/actions/traits, harm, sexuality/intimacy, affective states, negative social affect), and object-related arousal (concrete objects, environmental features, food and animals, physical space). Possibly, these might be differentially tied to flee or freeze responses, which suggests the possibility that increasing arousal for some types of words might lead to quicker responses (flee) whereas increasing arousal for other types of words might instead lead to slower responses (freeze). We stress that our use of cluster analysis is exploratory, qualitative, and entirely post hoc. We hope these data provide a useful pointer for further attempts to deconstruct the construct of arousal, but strong claims about the construct of arousal are best left to subsequent experimental research.

Two other possible interpretations of the low predictive validities of arousal rating scale norms are that (1) for some currently unknown reason, the construct of arousal cannot be reliably estimated with rating scales, and (2) whatever is being measured when people are asked to make arousal judgments with rating scales does not impinge strongly on word processing during lexical decision. These possibilities are not mutually exclusive with each other, nor with the first interpretation proposed.

The past few years has seen the release of numerous, large-scale semantic norms sets. Such data collection efforts are quite costly due to the volume of data required to acquire norms that have reasonable coverage over a language. Such efforts have almost exclusively relied on standard response formats seen in psychology (i.e., rating scales, or numeric estimation where rating scales are not applicable), with little consideration for alternative response formats that may be cheaper to use (but, in natural language processing see, Kiritchenko & Mohammad, 2016, 2017). The original motivation of introducing best-worst scaling for the collection of semantic norms was to test if alternative response formats might result in more efficient data collection. However, we now see that this work also raises important questions about the process of measurement itself. Most pressing is the issue that different measurement instruments appear to have different asymptotic predictive validities with lexical decision response times. This suggests that some, or all, of the response formats being used to collect semantic norms are introducing biases specific to the response format. This is a fact that would not be easily observable if only a single response format were employed. Thus, as we move forward in our efforts to compile large lexical semantic norms, we should be aware of the fact that some of our measurement instruments may be ill-suited for the task at hand (see also, Pollock, 2017) and, relatedly, that some of our semantic constructs may be poorly defined to participants (e.g., Connell & Lynott, 2012). To compensate for these issues, it is advised that we employ multiple measurement instruments and instruction sets to help avoid biases that may not be apparent when we consistently apply the same norm collection methodology. This will also help build a rich set of norms collected in different ways, against which we can begin to investigate why norms produced with different response formats in fact differ in their quantitative and qualitative properties.

Validating semantic norms against measures of lexical processing, caveat

The obvious and conventional metric by which to measure the quality of semantic norms is the ability for norms to predict performance on tasks that require lexical processing (e.g., Hollis, Westbury, & Lefsrud, 2017; Mandera, Keuleers, & Brysbaert, 2015). Insofar as semantics is involved in lexical processing, measures of the semantic properties of words should account for variance in behavioral measures of processing (e.g., Kuperman, Estes, Brysbaert, & Warriner, 2014). There are numerous theoretically motivated reasons to think that certain aspects of semantics should be involved in lexical processing (Barsalou, 1999; Paivio, 1990; Schwanenflugel, Harnishfeger, & Stowe, 1988; Vigliocco, Meteyard, Andrews, & Kousta, 2009).

However, there are possible conceptual problems with this approach. Consider the situation where a semantic property reflects a valid and stable part of a word's meaning, but simply has no bearing on lexical processing. Accurate estimation of the construct would result in a set of measures that have no predictive validity on measures of lexical processing. Consider an alternative set of values that have bias in them. For argument's sake, let us say that imageability judgments are biased by word frequency (because, by assumption, people confuse imageability with familiarity). In such a case, the norm with bias would come out as having more predictive validity of lexical processing than the unbiased norm, by virtue of being affected by frequency.

Something like the above situation may be at play with arousal norms. We demonstrated that rating scale arousal norms do not gain in predictive validity of lexical decision times as more judgments are sampled. One possible interpretation of this is rating scales allow for unbiased estimation of arousal, but that arousal simply does not play a role in aspects of lexical processing observable through lexical decision. In contrast, we found that best-worst arousal norms did increase in predictive validity as more data were sampled. However, best-worst norms accounted for no more unique variance than did rating scale norms after accounting for variance that could be attributed to other common factors including frequency, word length, neighborhood size, valence, concreteness, age of acquisition. Possibly best-worst arousal norms are gaining their predictive validities of lexical decision times by being biased by some other factor and, once that factor is controlled for, best-worst arousal norms lose their predictive validity. If this were the case, it would be erroneous to conclude that best-worst scaling provides the more accurate estimate of arousal; the apparent gain in quality could be due to the introduction of response bias. Alternatively, it could also be that rating scale norms are deficient in their sensitivity in detecting an actual relationship between arousal and one of these control variables.

Our analysis of arousal revealed other notable findings about patterns of prediction error for arousal norms across concrete and abstract words. These results suggest that concreteness and arousal are intertwined in a complicated way, and that different response formats are picking up on different aspects of the construct called *arousal*, as it pertains to concrete or abstract things. Thus, the true picture is likely something more complicated than one set of arousal norms being unilaterally better than the other, and this fact being detectable through tests of predicting lexical processing times.

Validating semantic norms against measures of lexical processing only makes sense if the particular semantic property in question plays a role in lexical processing. Even then, it would be advised to consider multiple validation measures. We observed in our analysis of concreteness that best-worst scaling norms and rating scale norms were not distinguished on their ability to predict lexical decision times, but they were distinguished on their ability to predict abstract/concrete decision times. Differences between norms may play out in some lexical processing tasks, but not others. Currently, very large databases are freely available for data on lexical decision, word naming, and concrete/abstract decision tasks. Similarly, large databases for other types of behavioral measures of lexical processing would help support a more robust and nuanced investigation of how semantics impinges on lexical processing.

The appropriateness of human-generated semantic norms

We should question whether or not it is even appropriate to employ semantic norms collected through human judgment in behavioral research. Our goal in behavioral research is often to explain some cognitive process vis-à-vis modeling a related behavior like, for example, explaining lexical processing via modeling the amount of time it takes a person to recognize a string of letters as a word. To the extent to which semantics plays a role in that process, estimates of the semantic properties of words should have an influence over that behavior. However, when our estimates of semantic properties themselves depend on a cognitive process, which semantic judgments most certainly do, we are attempting to explain one process with another process, or one behavior with another behavior. The circularity inherent in these types of accounts is anathema to sound scientific explanation (for discussion on this topic specific to the study of semantics see Westbury, 2016). It is possible we are simply chasing our own tails when we develop large semantic norms sets derived from human judgments and use them to try and explain psychological phenomena.

Such a stance is likely overly pessimistic. In cases where more objective measures of semantic constructs are available, there is a high degree of consistency between those measures and human judgments (e.g., for age of acquisition, Brysbaert & Biemiller, 2017). Second, even for very straightforward constructs like valence, it is not entirely clear how to come up with a reasonably valid operationalization that do not incorporate human judgment – semantics is a relational phenomenon between an observer and their environment. The observer's mind must play a role. Thus, even if we must ground our constructs in something other than human judgments eventually, human judgments do provide a useful, temporary placeholder for developing and testing psychological theories.

One avenue for breaking away from the circularity of relying on behaviorally elicited human judgments of semantics and the cognitive processes that underpin those judgments is to begin grounding our notions of semantics in computational models. Such models provide an objective basis for operationalizing semantics by mapping semantics into statistical properties of the environment (see, e.g., Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Recent models are exceptionally accurate at predicting human judgments of semantic relatedness (e.g., Baroni, Dinu, & Kruszewski, 2014), can replicate core psycholinguistic phenomena (Baayen et al., 2011), and make semantic distinctions thought to be important in psychological theory (Hollis & Westbury, 2016), without being explicitly training to do so.

Human judgments of semantic constructs currently have practical value for advancing and testing psycholinguistic theories. However, if we wish to avoid the problem of circular explanation, we will eventually have to move away from such norms and find alternative operationalizations of semantics that do not depend on the act of human judgment.

Acknowledgements This work was funded by a Discovery grant to the second author from the Natural Sciences and Engineering Research Council of Canada.

Appendix 1

This appendix contains instructions given to participants for each of the four semantic decisions data were collected for. Valence and age of acquisition data were collected in laboratory conditions, whereas concreteness and arousal data were collected online via crowdsourcing.

Valence

You are invited to take part in a study that is investigating emotion, and concerns how people respond to different types of words. You will be shown a set of words and choose the most and least pleasant from among the set. When a word is pleasant, it refers to something that makes you feel happy, pleased, satisfied, contented, hopeful. When a word is unpleasant, it makes you feel unhappy, annoyed, unsatisfied, melancholic, despaired, or bored.

Please work at a rapid pace and don't spend too much time thinking about each decision. Rather, make your ratings based on your first and immediate reaction as you read each word. Some words may be very close to each other in how pleasant or unpleasant they are. In such cases, go with your first impression.

Age of acquisition

You are invited to take part in a study that is investigating language, and concerns how people respond to different types of words. You will be shown a set of words and will have to decide which words you think you acquired when you were youngest or oldest. Examples of words that people typically acquire when they are young include 'mom', 'dad', 'yes', and 'no'. In comparison, words you have recently learned in University would be examples of words acquired later in life.

Please work at a rapid pace and don't spend too much time thinking about each decision. Rather, make your ratings based on your first and immediate reaction as you read each word. Some words may be very close to each other in when you likely learned them. In such cases, go with your first impression.

Arousal

You are invited to take part in a study that is investigating language, and concerns how people respond to different types of words. You will be shown a set of words and will have to decide which words you think are most and least emotionally arousing. Arousing words make you feel stimulated, excited, frenzied, jittery, wide-awake, or aroused. Non-arousing words make you feel relaxed, calm, sluggish, dull, sleepy, or unaroused.

Please work at a rapid pace and don't spend too much time thinking about each decision. Rather, make your ratings based on your first and immediate reaction as you read each word. Some words may be very close to each other in how arousing they are. In such cases, go with your first impression.

Concreteness

You are requested to make decisions about the meanings of words. You will be presented with a list of multiple words. For each list, you will need to choose the word that is MOST concrete in meaning and the word that is LEAST concrete in meaning. A concrete word is one whose meaning you can know directly from seeing, smelling, touching, or tasting things. For instance, 'dog' is a concrete word because you can see dogs. Words that are not concrete can only be known through other words. For instance, you cannot see, smell, touch, or taste 'justice'. Justice would not be a concrete word.

Words will sometimes be very close in how concrete they are. In such cases, we are only interested in your first impressions. Please do not spend too much time thinking about the answer.

References

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 11*8(3), 438.

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*(11), 1174-1220.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior research methods*, *39*(3), 445–459.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL* (1): 238–247.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, *22*(4), 637-660.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (pp. 1-45). Technical report C-1, the center for research in psychophysiology, University of Florida.

Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods*, 49(4), 1520-1523.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*(3), 904-911.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta psychologica*, *150*, 80-84.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*(3), 452-465.

Dale, E., & O'Rourke, J. (1981). The Living Word Vocabulary, the Words We Know: A National Vocabulary Inventory. Chicago: World book.

Estes, Z., & Adelman, J.S. (2008). Automatic vigilance for negative words in lexical decision and naming: Comment on Larsen, Mercer, and Balota (2006). *Emotion*, 8, 441-444.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, 35(3), 515-531.

Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How facebook and twitter provide the best frequency estimates for studying word recognition. *Cognitive science*, *41*(4), 976-995.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, *23*(6), 1744-1756.

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, *70*(8), 1603-1619.

Hollis, G. (2017). Soring best-worst data in unbalanced, many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, 1-19.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287-304.

Kiritchenko, S., & Mohammad, S. M. (2016). Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best-Worst Scaling. In *HLT-NAACL* (pp. 811–817) http://aclweb.org/anthology/N/N16/.

Kiritchenko, S., & Mohammad, S. M. (2017). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of The Annual Meeting of the Association for Computational Linguistics* (pp. 465-470). Vancouver, Canada. http://www.aclweb.org/anthology/P/P17/

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473-481.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables?. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1623-1642.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 3111–3119.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. The Quarterly Journal of Experimental Psychology: Section A, 50(3), 528-559.

Paivio, A. (1990). *Mental representations: A dual coding approach*. New York, Oxford University Press.

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *Behavior research methods*, *49*(2), 407-417.

Pollock, L. (2017). Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods*, 1–19. https://doi.org/10.3758/s13428-017-0938-y

Rescorla, R.A. & Wagner, A.R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning II, A.H. Black & W.F. Prokasy, Eds., pp. 64–99. New York: Appleton-Century-Crofts.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, *27*(5), 499-520.

Stadthagen-Gonzalez, H., Imbault, C., Sánchez, M. A. P., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior research methods*, *49*(1), 111-123.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(2), 219-247.

Vinson, D., Ponari, M., & Vigliocco, G. (2014). How does emotional content affect lexical processing?. *Cognition & emotion*, *28*(4), 737-746.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*(4), 1191-1207.

Westbury, C. (2016). Pay no attention to that man behind the curtain: Explaining semantics without semantics. *The Mental Lexicon*, *11*(3), 350-374.