Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments

Geoff Hollis¹

Published online: 27 May 2017
© Psychonomic Society, Inc. 2017

Abstract Best-worst scaling is a judgment format in which participants are presented with a set of items and have to choose the superior and inferior items in the set. Best-worst scaling generates a large quantity of information per judgment because each judgment allows for inferences about the rank value of all unjudged items. This property of best-worst scaling makes it a promising judgment format for research in psychology and natural language processing concerned with estimating the semantic properties of tens of thousands of words. A variety of different scoring algorithms have been devised in the previous literature on best-worst scaling. However, due to problems of computational efficiency, these scoring algorithms cannot be applied efficiently to cases in which thousands of items need to be scored. New algorithms are presented here for converting responses from best-worst scaling into item scores for thousands of items (*many-item scoring problems*). These scoring algorithms are validated through simulation and empirical experiments, and considerations related to noise, the underlying distribution of true values, and trial design are identified that can affect the relative quality of the derived item scores. The newly introduced scoring algorithms consistently outperformed scoring algorithms used in the previous literature on scoring many-item best-worst data.

Keywords Best-worst scaling · Tournament scoring · Rank judgment · Semantics · Human judgment

✉ Geoff Hollis
hollis@ualberta.ca

¹ Department of Psychology, University of Alberta, P217 Biological Sciences Building, Edmonton, Alberta T6G 2E9, Canada

Semantic judgments such as concreteness (to what extent is the referent something detectable by the senses), valence (how pleasant or unpleasant is the referent), and arousal (how calming or relaxing is the referent) are of wide and varied use within psycholinguistics and natural language processing (NLP). They are helpful for building classification models of text sentiment (e.g., Hollis, Westbury, & Lefsrud, 2017; Kiritchenko & Mohammad, 2016a; Mohammad, Kiritchenko, & Zhu, 2013; Zhu, Kiritchenko, & Mohammad, 2014), are used in experimental research studying cognitive processes (Abercrombie, Kalin, Thurow, Rosenkranz, & Davidson, 2003; Hamann & Mao, 2002; Lodge & Taber, 2005), as well as larger-scale statistical and modeling approaches (e.g., Baayen, Milin, & Ramscar, 2016; Hollis & Westbury, 2016; Kuperman, Estes, Brysbaert, & Warriner, 2014; Pexman, Heard, Lloyd, & Yap, 2016; Westbury et al., 2013). Historically these judgments have been collected under laboratory conditions (e.g., Bradley & Lang, 1999). However, it has now been demonstrated numerous times that high quality judgments can be collected online via crowdsourcing platforms (e.g., Brysbaert, Warriner, & Kuperman, 2014; Kiritchenko & Mohammad, 2016b; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Warriner, Kuperman, & Brysbaert, 2013). This is an important methodological step forward, as it allows for the collection of more data in less time; whereas lab-based data collection may take years to collect judgments for thousands of words, crowdsourcing provides judgments for tens of thousands of words in months. Having judgments available for orders-of-magnitude more words both opens up qualitative new lines of research (see Hollis et al., 2017) and improves outcomes in applied research settings (e.g., Mohammad et al., 2013; Zhu et al., 2014).

The ease of using crowdsourcing to collect human judgments, combined with the high quality of these judgments, has

led to a small data collection boom within psychology and NLP (e.g., Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014; Brysbaert, Warriner, & Kuperman, 2014; Imbir, 2015; M. Keuleers, Stevens, Mandera, & Brysbaert, 2015; Kiritchenko, Zhu, & Mohammad, 2014; Kuperman et al., 2012; Montefinese, Ambrosini, Fairfield, & Mammarella, 2014; Stadthagen-Gonzalez, Imbault, Pérez Sánchez, & Brysbaert, 2017; Warriner et al., 2013). However, unlike many participants from voluntary introductory psychology research pools, crowdsourcing workers need to be financially compensated for their time. This creates a need to innovate more economical ways to collect data.

Nearly all crowdsourcing efforts have used the conventional response format of a rating scale when possible, and numeric estimation when otherwise appropriate (e.g., Kuperman et al., 2012). Thus far, very little attention has been paid to the efficiency of these response formats, nor have researchers entertained the possibility that alternative response formats might be more suitable for data collection on a large scale (cf. Kiritchenko & Mohammad, 2016a, 2016b; Kiritchenko et al., 2014). This study introduces a response format, called *best–worst scaling*, that has multiple desirable properties for collecting large volumes of judgment data. Challenges are also described of scoring data generated from this response format when many items are being considered. Scoring is defined as the aggregation of individual data to produce a value that is expected to have a relationship with a latent variable of interest (as in, a student's exam is *scored* by summing the number of questions answered correctly and dividing by the total number of questions, which is expected to be related to the student's course knowledge). Finally, this article introduces a new set of scoring methods that overcome the computational limitations of extant methods, and then demonstrates their effectiveness in a series of simulation experiments.

Best-worst scaling

Best-worst scaling is a judgment format that has seen much use within economics contexts (for an introduction, see Louviere, Flynn, & Marley, 2015). Participants are presented with a set of N items and need to choose the superior and inferior item along some latent evaluative dimension. For example, participants might be presented with a list of four hotel amenities and be asked to choose the two from the list that patrons would consider the most and least essential. Not all items of interest are presented simultaneously. A hotel may be interested in evaluating 20 amenities but only ask a particular individual about four. Principles of block design can be applied to create sequences of subgrouped items that allow for accurate inferences about the full set of items.

Best-worst scaling has multiple desirable properties. Interval scales of judgment like the rating scale may fail to differentiate between items that are all of similar value along the underlying dimension of evaluation (e.g., *amenity essentialness*). Because best-worst scaling requires judges to make an ordinal decision, discrimination is forced even for items that are of similar value. The second, and more notable, property of best-worst scaling is that it generates a great deal of implied rank information. Over a set of four presented items, best-worst judgments provide ordinal information about five of the six implied relationships between the items. Suppose that items A, B, C, and D are presented. The person chooses A as best and D as worst. From these two judgments, we now know that $A > B$, $A > C$, $A > D$, $B > D$, and $C > D$. The only relationship we know nothing about is that between B and C. Note that this property is not unique to best-worst scaling. Rather, it is a consequence of sequential rank judgments. If participants were instead asked to choose the best and second-best items, the same amount of rank information would be provided, but about different pairs of items.

With four items per trial, best-worst judgments generate a maximum expected information of 3.59 bits. In contrast, judgments over a 7-point rating scale generate a maximum expected information of 2.81 bits. These values can be calculated from (1) the Shannon (2001) definition of expected information as the sum of $-p(x) * \log[p(x)]$ for each outcome, x ; (2) the fact that the information per decision is maximized when each outcome is equiprobable (Shannon, 2001); and (3) the fact that there are seven possible outcomes for a 7-point decision task and 12 possible outcomes for a best-worst decision over four items.

Within best-worst scaling, the number of items presented per trial can be increased. This in turn increases the amount of information generated by best-worst judgments. One concern is that beyond some N , participants may display difficulties in making such judgments. However, even in the case of an N of four, the fact that best-worst scaling generates such high quantities of information per judgment suggests it may be of value for reducing the costs in crowdsourcing human judgments; relative to rating scales, fewer judgments are needed to generate an equal amount of information. It remains to be demonstrated whether the information generated by best-worst scaling is equally useful for quantifying words along latent dimensions as the more common response format, the rating scale.

There are practical considerations for scoring best-worst data. With rating scales, a researcher can average responses to a particular item to derive an unbiased estimate of the true value, under the assumption that variation over judgments is error variation and that errors have a symmetric distribution. This assumption is consistently applied to rating data (e.g., Bradley & Lang, 1999; Stadthagen-Gonzalez et al., 2017; Warriner et al., 2013). However, best-worst judgments

produce a more complex datum. Each trial provides two nonnumeric values (the best item and the worst item), so averaging the outcomes of trials is not applicable. An alternate is to employ counting methods, for example, tallying how often each item was chosen as the best item. Even so, there are multiple ways to count best-worst data, each producing scores of different quality (e.g., Marley, Islam, & Hawkins, 2016).

Previous applications of best-worst scaling have almost all exclusively been in domains where a couple of dozen items need to be scored. In such cases, there are a variety of different scoring algorithms. One common method for scoring best-worst judgments is to subtract the proportion of “worst” responses from the proportion of “best” responses for a given item (Kiritchenko & Mohammad, 2016a, 2016b; Kiritchenko et al., 2014; Louviere et al., 2015). This scoring algorithm is referred to as *best-worst counting* (the dash indicating a minus sign, not a hyphen). There are two main shortcomings of best-worst counting. First, it does not leverage the fact that making “best” and “worst” judgments provides information about the relative rankings of unjudged items. Second, it ignores the degree of competition between alternatives; a close win for a status as “best” is treated exactly the same as a landslide victory. This latter point can be addressed by employing balanced incomplete block designs (BIBDs) to construct subsets of items to be judged on particular trials (Louviere et al., 2015). However, due to combinatorial growth, BIBDs cannot be used in situations where thousands, or tens of thousands, of items need to be scored (henceforth referred to as *many-item experiments*).

A more sophisticated way to score best-worst data is to construct statistical models that predict choice outcomes from ranking data. Scores for individual items can be derived from model parameter values. Both hierarchical Bayes and multinomial logit models have been successfully applied to scoring best-worst data, with hierarchical Bayes currently being the gold standard scoring method (Orme, 2005). However, both of these models are computationally costly and do not scale well to many-item experiments.

A more recent best-worst scoring algorithm is the analytical closed-form solution to best-worst scaling (ABW; Lipovetsky & Conklin, 2014). The ABW is calculated as the logarithm of the ratio: one plus the unit-normalized best-worst score, over one minus the unit-normalized best-worst score. ABW has been demonstrated to have lower error than best-worst counting for estimating the latent values of items. ABW also results in value estimates of comparable quality to multinomial logit models while being computationally more efficient to calculate (Marley et al., 2016). However, like best-worst counting, ABW does not take into account the strength of competition between items in a best-worst set and so depends on BIBDs for unbiased estimates. Thus, ABW as a scoring algorithm may have limitations in many-item experiments. If best-worst scaling is to be leveraged to

compile high-coverage datasets of semantic values for words, new scoring algorithms need to be developed.

Tournament scoring

Consider a sequence of best-worst trials over a set of S items. Suppose each trial poses a subset of $N = 4$ of the possible S items for evaluation. At the end of the experiment, the researcher is left with a sequence of trials $t_0, t_1, t_2, \dots, t_i$. The question is: How can the researcher derive scores for each item from this sequence of trials?

Think of the S items as being contestants in a tournament. A tournament is defined as a collection of sequential confrontations (matches) between pairs of contestants. Match resolutions assign statuses of winner or loser to each contestant that was involved, or alternately a resolution of draw if no winner or loser emerged. When a tournament terminates, it typically produces a list of contestants, ranked by their performance.

Each best-worst trial can be thought of as containing information about the outcomes of five hypothetical matches in a tournament. If A was judged best and D was judged worst, we know that A would win a match against B, C, and D. We also know that B and C would each win a match against D. We can use the information contained in best-worst data to simulate a tournament that will assign ranks to each of the S items based on hypothetical matches derived from trial data.

Thus, we can frame the problem of scoring items from best-worst data as being analogous to the problem of finding competitor ranks in a tournament with a sequence of match outcomes. There are many known computationally efficient solutions to this problem, chess-scoring algorithms (Elo, 1973) being one of the most well-known examples. Expanded-rank approaches have been taken to scoring best-worst data previously (e.g., Marley & Islam, 2012) but, to our knowledge, the problem of scoring best-worst data has never previously been explicitly framed as a tournament problem, nor has the effectiveness of tournament algorithms been measured for scoring best-worst data. The tournament framework may be a productive way to think about scoring best-worst data in many-item unbalanced designs; common tournament-scoring algorithms have mechanisms that accommodate for unbalanced paired comparisons in many-item contexts.

Here I introduce two simple tournament-scoring algorithms and point out their relationship to best-worst counting, as well as their limitations. The purpose of this exposition is to illustrate the most relevant feature of Elo scoring: It takes the quality of competition into account when assigning scores. Scoring algorithms that take into account the quality of competition will be essential to scoring many-item problems with unbalanced designs. Two other (nontournament) scoring algorithms that also have this feature will be introduced after.

The simplest tournament scoring algorithm would be to assign a point to each winner (win scoring). Win scoring is unsatisfactory because it ignores the quality of competition; a win against a strong opponent is weighted equally to a win against a weak opponent. This would be sufficient in the situation where each item faces off against each other item an equal number of times (e.g., a round-robin tournament). However, this situation is impractical when the number of items to score increases. A variant of win scoring is to additionally subtract a point for each loser. Win–loss scoring calculates a score over all *implied* wins and losses within a best-worst trial. Win–loss scoring results in scores proportional to best-worst counting; in best-worst counting, the best option gets a value of +1, the worst option gets a value of −1, and the two unchosen options get values of 0. In win–loss scoring, the best option gets a value of +3 (three wins), the worst option gets a value of −3 (three losses), and the two unchosen options get a value of 0 (one win, one loss).

As compared to win scoring, best-worst counting (and win–loss scoring) creates an extra source of differentiation among competitors (e.g., now players who always draw have different scores from players who always lose). However, best-worst counting still does not weight scores by the quality of competition.

A well-established tournament-scoring algorithm is the one developed by Arpad Elo (Eq. 1; see Elo, 1973). After a match, each player's score is updated according to the disparity in their scores prior to the match. If a player with a *high* score, H , beats a player with a *low* score, L , then H 's score increases a little and L 's score decreases a little. If, instead, L beats H , L 's score increases a large amount and H 's score decreases a large amount. Specifically, for some player A , their rank R_A is updated after a match based on the discrepancy between their actual score in a match, S_A , and their expected score, E_A , scaled by a constant factor K . The calculation for player rank is presented in Eq. 1.

$$R'_A = R_A + K(S_A - E_A) \quad (1)$$

$$E_A = 10^{R_A/400} / (10^{R_A/400} + 10^{B_A/400}) \quad (2)$$

The Elo scoring system assumes that player skill is normally distributed and has a standard deviation of 100 along the Elo scale, and that wins and losses are primarily determined by player skill. Consequently, it is expected that two players of equal skill will win against each other equal numbers of times, and that in cases in which two players have different scores, the player with the higher score should be more likely to win and that the win chance is a function of skill disparity. From these assumptions, it can be proven (Elo, 1973) that the expected chance that a player A will win against another player B is an exponential of A 's score, divided by the same number

plus an exponential of B 's score. This ratio provides the basis for calculating the expected score term in Eq. 1, and is detailed in Eq. 2.

Elo's update rule is sensitive to the relative disparity of rank between the two competitors. Thus, it addresses the central shortcoming of both best-worst counting and win–loss scoring. Also unlike both logit and hierarchical Bayes scoring methods, as well as BIBDs with win–loss counting, Elo scoring is not limited by problems of computational efficiency. It is a very simple calculation whose time complexity for score updates after a match is constant. Thus, Elo presents itself as an ideal method for scoring best-worst data when large numbers of items need to be scored.

Discriminative learning

Elo is a scoring mechanism that learns to discriminate players based on relative skill. It is possible that other discriminative-learning algorithms may likewise be suitable for scoring best-worst data. One of the most well-known discriminative-learning algorithms comes from Rescorla and Wagner (1972). The Rescorla–Wagner model is a model of classical conditioning in which, rather than learning a relationship between two stimuli by association, learners learn via discrepancies between what does happen and what is expected to happen.

When a cue occurs (e.g., a particular player in a tournament) and is followed by a specific event (e.g., that player winning or that player losing), the association strength between the cue and the event is updated to reflect that observed relationship (Eq. 3). The degree of change due to each cue–event pair (Eq. 4) depends on the salience of the observed association, α , a learning rate parameter for the cue, β , and the difference between the maximum association strength for the event, λ , and the total association strength between the cue and all events, V_{tot} .

$$V_x^{n+1} = V_x^n + \Delta V_x^{n+1}, \quad (3)$$

$$\Delta V_x^{n+1} = \alpha_x \beta (\lambda - V_{\text{tot}}). \quad (4)$$

In this particular case, the cue (an item to be scored or, equivalently, a participant in a tournament) can be associated with two different events: a win or a loss.

The Rescorla–Wagner update rule can incorporate information about the quality of competition via the salience term. When an item wins against a much worse opponent, the outcome is less salient than when an item wins against a much better opponent. We can quantify the salience of the event as 1—the probability that the actual winner was expected to win (i.e., unexpected wins are salient; Eq. 5). The probability that the actual winner would have won can be estimated as the association strength between the winning item and a win, divided by the summed association strength between each

competitor in the match and a win. In the case in which neither competitor has a prior association with winning, the salience term is instead set to .5.

$$\alpha = 1.0 - V_w(\text{winner}) / [V_w(\text{winner}) + V_w(\text{loser})] \quad (5)$$

There is an important difference between the Rescorla–Wagner update rule and Elo scoring, as it pertains to scoring best-worst data. Elo scores have no upper or lower boundaries: As long as a particular player is winning and losing consistently, larger and larger discriminations can be made between him and the rest of the field. In comparison, Rescorla–Wagner association strengths are bounded by a theoretical maximum conditioning strength possible between a cue and event, specified by λ . We should expect the two models to produce different predictions, and the differences in predictions should be most pronounced for players that consistently win or lose, due to the fact that the Rescorla–Wagner update rule has boundaries on association strength, but Elo does not have boundaries on the differences between players' skills.

Value learning

Now a third scoring algorithm will be introduced, referred to as *value learning*. Value learning is an algorithm that learns the expected value of a match for each player. Tournament ranking can then be determined by ordering players according to their expected value of being involved in a match. Player A's value, V_A , is calculated by updating the player's estimated value towards the observed value, γ , from each encountered match. The update rule for value learning is presented in Eq. 6.

$$V_A = V_A + \alpha\beta(\gamma - V_x) \quad (6)$$

Like with the Rescorla–Wagner update rule, value learning learns from a discrepancy between an expectation and an observation. However, value learning learns from the observed *value*, γ , (e.g., win = 1, loss = 0) of an event rather than learning an association between a cue and an event.

Value learning also includes a learning rate, β , and a salience parameter, α . For value learning, salience is defined in terms of the relative odds that either player is expected to win a matchup against the entire field of opponents. The odds of a player winning, O_A , are calculated by treating the expected value of the player as a win probability, and converting it to an odds value (Eq. 7). Outcome salience is then calculated as 1.0 minus the odds of the actual winner winning a match, divided by the sum of the odds of each player winning (i.e., unexpected wins are salient; Eq. 8).

$$O_A = V_A / (1.0 - V_A), \quad (7)$$

$$\alpha = 1.0 - O_A / (O_A + O_B) \quad (8)$$

Equation 8 ensures that salience is bounded between 0 and 1 but grows as the observed outcome becomes less expected. In the case in which neither player has any odds of winning a match (e.g., because neither has ever won a match), salience is set to .5.

A limitation to value learning is that values need to be specified for the varieties of outcomes, but such values may not be known ahead of time. Critically, the *absolute* value is not relevant; the *relative* value is what matters. It does not matter that a win is worth 1, a tie is .5, or a loss is 0. The learned values would be proportional if instead a win were 10, a tie were 5, and a loss were 0. What matters is the values of the outcomes relative to each other. When a win is worth 1 and a loss is worth 0, should a tie be valued at .5, .3, or other? Consequently, in the case of only two outcome events, deciding on the values is inconsequential; one will always be lower than the other, and the final learned value can be scaled as desired. Thus, it is reasonable to arbitrarily assign wins a value of 1 and losses a value of 0 if those are the only two outcomes from which values are being learned.

Experiment 1

The purpose of this simulation was to test the relative quality of the various scoring algorithms for deriving item scores from best-worst data. Quality was assessed on the basis of the abilities of scored values to be used to predict true values along a latent dimension that best-worst judgments were made along. For instance, we might want to know whether best-worst scaling can be used to derive useful measures of word valence. Participants would be presented with a series of trials, where on each trial they see a selection of four words. Participants would then be instructed to choose the item that has the highest valence and choose the item that has the lowest valence.

The challenge is that with real-world examples such as the one above, we do not know the true values along the underlying latent dimension; this is why we need to measure them in the first place. Rather than using a real-world application of best-worst scaling, here simulated best-worst responses are used instead, over a latent dimension where the true values were known ahead of time. This will allow us to assess the abilities of the various learning algorithms to predict true values from observations.

Method

Judgments over a simulated latent dimension Simulations were based on a sequence of best-worst judgments over a set of 1,000 items to be evaluated. Each item was assigned a true value along the latent dimension that judgments were made over. These values were generated by sampling randomly

from a normal distribution with a mean of 0 and a standard deviation of 1. Among a set of items present on a particular trial, the best choice was determined as the item with the highest value along the latent dimension. The worst choice was determined as the item with the lowest value along the latent dimension. Each simulation was repeated 100 times, with randomized trials each time.

Trial structure The items from each trial were chosen with random uniform sampling over the set of all possible 1,000 items, replacing items after the creation of each trial. Each trial contained four items. Within the simulation, some items might occur in more trials than other items. Simulations were run using 1,000, 2,000, 4,000, 8,000, 16,000, and 32,000 trials. The expected numbers of occurrences of each item are 4, 8, 16, 32, 64, and 128, respectively. The impact of sampling equality was tested in Experiment 4.

Dummy players and conversion to match format After best-worst trials were simulated, data were converted to pairwise match format so scoring algorithms could be applied. Two additional matches were included for each item being scored. One match was against a dummy player who always won. The other was against a different dummy player who always lost. Thus, with 1,000 items to be scored, 2,000 dummy trials were added to each simulation. Inclusion of dummy players ensured that none of the items to be scored had perfect win or perfect loss records. During early exploratory work, introducing dummy players produced scores that were more strongly correlated with the true values.

Scoring Once best-worst judgments were simulated, they were scored using best-worst counting, ABW, Elo scoring, Rescorla–Wagner scoring, and value scoring. The first two algorithms will henceforth be collectively referred to as count algorithms, whereas the remaining three will henceforth be collectively referred to as predict algorithms.

In the cases of predict algorithms, 100 iterations were run over the match data. A large number of iterations were used to ensure learning convergence. On each iteration, the order of match data was randomized. On the basis of early exploratory work, the initial learning rate for Rescorla–Wagner and value scoring was set to .05. On each successive full iteration over the match data, the learning rate was divided by the iteration number. Reductions in learning rate over epochs generally lead to better performance for learning algorithms (Raschka, 2015). For Elo, the K factor was set to 30. The K factor is related to the learning rate; it determines gains and losses in the scores of the winners and losers after a matchup. This value was determined by exploratory work prior to the experiment. Learning rate was not explicitly reduced over iterations for Elo scoring, since this feature

is an implicit property of Elo scoring; scores become more differentiated due to more matches being played, but the change in value for two evenly matched opponents stays constant, as determined by K .

Scores assigned by the count algorithms do not change as a result of multiple iterations over the match data because in each case, scores are learned from the differences of two fixed probabilities. ABW and best-worst scores were calculated from only a single iteration over the match data.

Comparison of scoring methods The scoring algorithms were assessed by calculating R^2 values (from the Pearson correlation coefficient) between the derived scores and the true values along the latent dimension that best-worst decisions were made over.

Results

It was evident from the data that a log-odds transformation to the calculated scores fit the true values better than did the raw scores themselves (see, e.g., Fig. 1). All results reported are thus for log-odds transformations of the data, with the exception of ABW scoring, in which a log-ratio transformation is already applied. The odds ratio of a probability, p , is equal to $p/(1 - p)$. If item A has a .75 chance of winning a match, its odds are $.75/(1 - .75) = 3$ (or, three-to-one). The log-odds are the logarithm of this ratio (also referred to as the *logit function*).

In applying a log-odds transformation, we assume that the calculated scores are proportional to the probabilities that items will win matches. Empirically, this is clearly a reasonable step; log-odds transformations of scores demonstrably provides a substantially better fit to the true values than do the raw scores themselves (Fig. 1). Conceptually, this also seems like a reasonable step: Items with higher scores win more of their matches than do items with lower scores.

In cases in which the scores did not range between 0 and 1 naturally (Elo, best-worst counting, win–loss scoring), the data were first scaled to be within the range of 0 and 1 (exclusive). For Elo scores, which theoretically can range between negative and positive infinity, data were scaled so that two dummy players (always-lose, always-win) had rescaled values of 0 and 1, respectively. The values produced by best-worst counting and win–loss scoring range between –1 and 1. The scaling was accomplished by adding 1.0001 to the scores and dividing by 2.0002. The small fractional values were included to ensure that no scaled value was exactly 0 or 1, at which points the log-odds transformation is not defined.

A 5 (Type of Scoring) $\times$ 6 (Number of Trials) analysis of variance (ANOVA) was computed over the R^2 values between the true values and derived scores for each simulation to compare the relative performances of the five scoring algorithms.

Fig. 1 Comparison of fits between the true values of the latent dimension and (a) scores derived from the Rescorla–Wagner update rule and (b) log-odds-transformed Rescorla–Wagner scores. The data come from Iteration 100 on the 8,000-trial simulation. Log-odds transformation provides a

better fit to the underlying true values. This was true for each of the three new scoring algorithms investigated (Elo, Rescorla–Wagner, and value)

The mean data are presented in Fig. 2. As anticipated, there were main effects for both scoring algorithm [$F(4, 2970) = 9,551, p < 2.2e-16$] and the number of trials in a simulation [$F(5, 2970) = 100,072, p < 2.2e-16$], as well as an interaction between the two factors [$F(20, 2970) = 600, p < 2.2e-16$]. A planned contrast revealed that predict algorithms consistently performed better than count algorithms [$F(1, 2970) = 24,171, p < 2.2e-16$]. All three of the predict algorithms produced scores that were nearly perfectly correlated with the true values, given enough data (Elo mean $R^2 = .996$, R–W mean $R^2 = .992$, value mean $R^2 = .994$). Although the two count algorithms do perform well with enough data, neither quite reaches the same performance level as the predict algorithms. Looking at just the predict algorithms, Elo scoring was consistently the best of the three [$F(1, 2970) = 11,757, p < 2.2e-16$]. Indeed, Elo's R^2 first exceeded .99 at 8,000 simulated

judgments; value and Rescorla–Wagner scoring did not reach this mark until 32,000 simulated judgments. Between the two count algorithms, ABW consistently performed better than best-worst counting [$F(1, 2970) = 11,760, p < 2.2e-16$].

Discussion

This simulation provides evidence that best-worst scaling can be used as a judgment format to derive scores that correlate nearly perfectly with the true values along a latent dimension. However, this simulation also demonstrates that the scoring algorithm matters. Up to this point, the only scoring algorithm used in many-item best-worst experiments has been best-worst counting (Kiritchenko & Mohammad, 2016a, 2016b; Kiritchenko et al., 2014). However, this was the worst algorithm tested in the present simulation. Best-worst counting does not take into consideration the strength of competition. This is also true of analytic best-worst scoring. When the strength of competition is incorporated into scoring—for instance, in the predict algorithms—effectively perfect scores can be obtained. These simulations indicate that both best-worst counting and analytic best-worst scoring will converge to perfect performance but require substantially more data to do so.

The most promising scoring algorithm thus far is Elo. We find that Elo can estimate the underlying true value along a latent dimension for 1,000 items with as few as 8,000 best-worst judgments (an average of eight decisions per item). Most large-scale research with the goal of collecting human estimates of semantic properties uses, as a convention, about 20–25 judgments per item (e.g., Brysbaert et al., 2014; Kuperman et al., 2012; Warriner et al., 2013). These studies have primarily relied on rating-scale responses. If best-worst scaling provides scores of equal quality, this means that switching to best-worst scaling for data collection might result in requiring 60%–68% fewer judgments to derive

Fig. 2 Comparison of the quality of estimates derived from each of the five scoring algorithms over six logarithmically spaced dataset sizes. Elo scoring consistently outperforms all other scoring algorithms

comparably useful scores. That is a substantial savings of both time and money, particularly when experiments may cost thousands of dollars to run.

The applicability of best-worst scaling to collecting judgment data in high volumes is predicated on a few points. First, it assumes that best-worst scaling and rating scales provide data of equal quality. I am aware of no extant literature comparing the quality of the data generated by these two response formats. An empirical study comparing the quality of data generated by these two response formats needs to be carried out. However, a few things that we know about both best-worst scaling and rating scales allow us to make predictions of relative quality. It is known that rating scales have numerous sources of bias, including tendencies of the judges to avoid using extreme values on scales, individual judge tendencies toward one end or the other of the scale (Saal, Downey, & Lahey, 1980), as well as interparticipant variation in interpretations of what the scale values mean (Ogden & Lo, 2012). Furthermore, small changes to scale features can result in notable differences in value distributions (e.g., Weijters, Cabooter, & Schillewaert, 2010). By the very nature of the judgment, it is unlikely that best-worst scaling suffers any of these three issues. However, rating scales produce interval information, whereas best-worst scaling produces ordinal information. It may be that interval information is a fundamental necessity for talking about the organization of semantics (e.g., Osgood, Suci, & Tannenbaum, 1957), and that best-worst scaling ultimately does not provide the right type of information.

Although the reported results allow us to be optimistic, they were made under highly idealized assumptions. This simulation assumed a perfect judge that always made best-worst choices according to the rank value of items. It is possible that this assumption does not hold in practice, though results from Kiritchenko and Mohammad (2016b) suggest that responses to particular best-worst trials are highly reliable between participants. Regardless, the scoring algorithms need to be reassessed in the presence of noisy judgments. Also, Simulation 1 assumed that the underlying distribution of true values was normally distributed. It is possible that these scoring algorithms would break down when assumptions of normality are violated. These two points will be the focuses of the following two simulations.

Experiment 2

Experiment 1 made the assumption of a perfect judge: Best-worst judgments always corresponded to the items with the highest and lowest true value along the latent dimension that best-worst judgments were being made over. In reality, human judgments involve both inter- and intrajudge variability. This noise may interfere with the ability of various scoring

algorithms to produce scores that are proportionally related to true values.

According to the present hypothesis, the differences in performance between the three predict algorithms and the two count algorithms would diminish as noise is added to judgments. One main shortcoming of count algorithms is that they do not take into consideration the quality of competition among any particular set of items to be judged; all wins are treated the same, regardless of who was beat, and so are all losses. In contrast, the three predict algorithms do take into account such information, and thus are better able to estimate the underlying latent values that are driving best-worst judgments. However, as judgments become increasingly noisy, the presented items become more homogeneous in their competitiveness; with enough noise present, the least-valued item has a fair chance of being judged best.

In tournament terms, noise shifts the context away from a “game of skill,” whose outcome is primarily determined by latent ability, and more toward a “game of chance,” where randomness can lead to the worse player winning. Predict algorithms should only show an advantage over count algorithms if there is, indeed, a stable basis on which to predict the outcome. There is less basis for prediction in games of chance, and hence we should expect all predict algorithms to have higher losses in score quality under noise than count algorithms, and eventually to converge with the performance of count algorithms under conditions of high noise in judgments.

Method

For Experiment 1, four items were chosen to be judged on each trial. Best-worst judgments were made on the basis of which of the two items had the highest and lowest value along the latent dimension, respectively. In the present simulation, a noise component was added to each true value prior to choosing the best-worst items of a trial. Thus, there was the possibility that the chosen best-worst items did not correspond to the actual best-worst items. This simulation was identical to Experiment 1, with the exception that noise was added to judgments.

The noise was drawn from a normal distribution with a mean of 0.0 and a standard deviation of 0.5, 1.0, or 2.0 (low-, medium-, and high-noise conditions, respectively). Note that the latent dimension was simulated by pulling values from a normal distribution with a mean of 0.0 and a standard deviation of 1.0.

Results

As in Experiment 1, log-odds transformation of the scores for the predict algorithms provided a better fit to the true values than did the raw scores. Only the results based on log-odds

scores are reported in these cases. However, note that under the presence of noise, the raw scores did not show the same degree of distortion in their estimates of extreme values (Fig. 3) that were observed in Experiment 1, in which no noise was present (Fig. 1).

The experimental results are presented in Fig. 4. A 5 (Scoring Method) $\times$ 6 (Number of Trials) $\times$ 3 (Noise Level) ANOVA was conducted to assess the relative performance of the scoring algorithms. All main effects were highly reliable ($p < 2.2\text{e-}16$), as were all interactions ($p < 2.2\text{e-}16$). We now turn to planned contrasts, from which three findings are of interest. First, under conditions of noise, Elo scoring no longer stands out as the clear best scoring algorithm. In fact, Elo provided the worst scores by a large margin [$F(1, 8910) = 6,102, p < 2.2\text{e-}16$], and this effect interacted with noise levels: As more noise was present in judgments, Elo's performance relative to the other scoring algorithms got worse [$F(2, 8910) = 4,365, p < 2.2\text{e-}16$]. Second, as expected, the predict algorithms performed better than the count algorithms, but this effect was diminished as more noise was added to judgments [$F(2, 8910) = 453, p < 2.2\text{e-}16$]. Third, under conditions of noise, value scoring stood out as the best overall scoring algorithm [$F(1, 8910) = 2,434, p < 2.2\text{e-}16$]. However, its advantage did depend on the number of trials present in a simulation [$F(5, 8910) = 204, p < 2.2\text{e-}16$]. Elo scores were omitted from the last four contrasts, due to Elo's evidently unique performance attributes.

Discussion

The various scoring algorithms are differentially sensitive to judgment noise. Experiment 1 demonstrated that Elo is the best-performing scoring algorithm under conditions of no noise in judgments. However, Experiment 2 further revealed that Elo is also the scoring algorithm most impaired by the presence of noise in judgments. Value scoring appears to be

the most robust scoring algorithm across all conditions from Experiments 1 and 2 and is consistently the best scoring algorithm in the presence of any noise in judgments. As predicted, the relative differences between count algorithms and predict algorithms was minimized as judgment noise increased.

It is a surprise to observe that Elo suffers so markedly under the presence of high levels of noise. Perhaps this has something to do with differing learning objectives. Elo is learning to discriminate players on the basis of relative performance, whereas Rescorla–Wagner is instead learning to discriminate between players on the basis of match outcomes, and value scoring is learning the expected value of a match for a particular player. These different learning objectives may be more or less sensitive to noise in the outcomes. However, currently this is mere speculation.

Now we can contextualize the magnitude of noise examined into extant data from human judgment experiments. Our high-noise condition involved the addition of noise pulled from a normal distribution with a standard deviation of 2.0. In comparison, values for our simulated latent dimension were drawn from a normal distribution with a standard deviation of 1.0. This leads to a situation in which the relative ordering of items, according to calculated scores, can be opposite that specified by the underlying latent dimension; under the high-noise condition, it is reasonably possible for the scores of items to not properly capture the relative ordering of items separated by two standard deviations along the underlying latent dimension. For context, that is like confusing the ordering of a cactus and a princess along a measure of valence ($z\text{Valence} = 0.03$ and $z\text{Valence} = 2.01$, respectively; data from Warriner et al., 2013). It seems unlikely that humans would confuse the relative valences of two such words. A much more psychologically realistic level of noise is the 0.5-SD condition, which allows for the misordering of words like *admiration* ($z\text{Valence} = 1.97$) and *fantasize* ($z\text{Valence} = 1.51$).

Fig. 3 Comparison of fits between the true value of the latent dimension and (a) estimates derived from the Rescorla–Wagner update rule and (b) log-odds-transformed Rescorla–Wagner estimates under conditions of noise ($SD = 0.5$). The data are from Iteration 500 of the 8,000-trial

simulation. Unlike in Experiment 1, under conditions of noise a log-odds transformation only marginally improved the fit between predicted values and the true underlying values

Fig. 4 Relative performance of the five scoring algorithms under the presence of various levels of noise. Performance is measured in terms of R^2 between the true values and the derived scores

Empirical results from Kiritchenko and Mohammad (2016b) suggest that humans can reliably make very fine-grained distinctions in the semantic properties of words using best-worst scaling, supporting the argument that any sort of noise that is present in human judgments using best-worst scaling will be minimal.

Under the 0.5-SD noise condition, value scoring stands out as clearly the best scoring algorithm, followed by Rescorla-Wagner and win-loss scoring. All three algorithms produce scores that correlate near-perfectly with the true values, given sufficient data. However, even at the smaller scale of 4,000–32,000 judgments, each of these three algorithms converges to scores that strongly correlate with the true values (R^2 ranging from .91 to .98). Note that best-worst counting and ABW scoring are the worst-performing algorithms in this range by a large margin, yet best-worst counting is the only algorithm that has been employed thus far for scoring best-worst data over large item sets (Kiritchenko & Mohammad, 2016a, 2016b; Kiritchenko et al., 2014). We can conclude that under the presence of noisy judgments, best-worst scaling displays evidence of being a useful format for capturing the variation that exists along latent dimensions of judgment, as long as an appropriate scoring algorithm is used. Thus far, value scoring appears to be the most robust scoring algorithm. Elo is the least robust scoring algorithm, but also has the best performance when noise is absent from the data.

Experiment 3

We have assumed thus far that the values along the underlying latent dimension are normally distributed. The importance of this assumption needs to be tested, because existing estimates of semantic dimensions (derived from rating scales) are

demonstrably nonnormal, displaying skewness in the case of affect judgments (e.g., Warriner et al., 2013) and bimodality in the case of concreteness judgments (e.g., Brysbaert et al., 2014). It is possible that these distributions reflect an artifact of responding with rating scales rather than something about the latent dimension underlying the judgments. Alternately, but not exclusively, these distributional properties could reflect something important about the underlying semantic dimensions themselves. If the latter case is true, it becomes important to assess the relative performance of the proposed scoring algorithms when dealing with items that are not normally distributed along the latent dimension.

The prediction in this experiment was that all five scoring algorithms would display difficulties accurately scoring nonnormal data. This is because best-worst scaling provides only ordinal information; judgments carry information that A is greater than B along some dimension, but not by how much. When a distribution changes shape such that A and B get closer or farther from each other in absolute terms, best-worst scaling will be insensitive to those shifts in distance.

A further prediction was that the presence of noise in judgments would improve the ability to infer latent values from best-worst data. Noise introduces interval information across large numbers of best-worst judgments. Suppose that two items, A and B, are separated by one unit distance along the latent dimension underlying judgments. Then suppose that A and C are separated by two unit distance. If any noise is present in the judgments, A and B will reverse their ranking order more often than A and C. Thus, ratios of best judgments and ratios of worst judgments between pairs of items when both items appear in the same trial will carry information about their distance along the underlying latent dimension.

Method

This simulation was identical to Experiment 1, with the exception that true values were drawn from three new distributions: uniform, exponential, and F . Furthermore, simulations were run under conditions of no or of some noise in judgments (Gaussian noise with mean = 0.0 and $SD = 0.5, 1.0$, or 2.0). For the uniform distribution, values were bounded between 0.0 and 6.0. For the exponential distribution, the decay rate was set to 1.0. The result was a one-tailed distribution in which nearly all values ranged between 0.0 and 6.0. For the F distribution, df_1 was set to 100, and df_2 was set to 10. The result was a two-tailed distribution with a positive skew in which nearly all values ranged between 0.0 and 6.0. These parameters were chosen so that the ranges of values would roughly match the range of values from Experiments 1 and 2; the large majority of normally distributed data with an SD of 1.0 fall within a range of 6 (99.8% of the data fall within ± 3 SD s).

Results

As with the previous simulations, log-odds transformation of the scores fit the true values better than did the raw scores. Only results for log-odds data are reported.

For ease of comprehension, only the results for simulations containing 32,000 best-worst judgments are described here. The results from simulations containing fewer judgments were as predicted on the basis of the results from Simulations 1 and 2. To recapitulate, Elo generally performed best in cases of no noise and when few simulated judgments were made. Elo performance dropped off in the presence of noisy judgments. Value learning was the most robust scoring algorithm across all levels of noise and also was generally the best-performing algorithm in the presence of noise. Predict algorithms generally outperformed count algorithms, and the relative performance differences between the two sets of algorithms was attenuated as more noise was present in judgments.

The two predictions were supported by the data (presented in Table 1). First, all scoring algorithms showed deficiencies in producing scores that matched the underlying true values across all nonnormal distributions in the absence of judgment noise. The most pronounced deficiencies were observed for best-worst counting, which performed decently when the true values were normally distributed ($R^2 = .964$), and substantially worse when the values had an $F(100, 10)$ distribution ($R^2 = .745$).

The second expected and confirmed finding was that these scoring deficiencies would be mitigated under the presence of judgment noise. The mean performances of the scoring algorithms for normally distributed latent values were $R^2 = .984$ and $.981$ under the 0.0- and 0.5- SD noise conditions, respectively. Across all nonnormal distributions, the averages were

$R^2 = .850$ and $.961$, respectively. Thus, a small amount of noise actually *increased* the scoring performance for nonnormal distributions, but reduced it for normal distributions. Figure 5 provides an example of how the Rescorla–Wagner algorithm performed at scoring items under varying levels of noise when the true values had an $F(100, 10)$ distribution. The presence of a small amount of noise in the judgments ($SD = 0.5$) was sufficient to linearize the relationship between the derived scores and the true values. Complete summary data are presented in Table 1.

A few interesting findings are visible in Table 1. First, it is apparent that under high-noise conditions, value scoring was consistently the best algorithm for all distributions. Under no- or low-noise conditions, the best scoring algorithm was

Table 1 Performance of the five scoring algorithms as a function of judgment noise and the underlying distribution of the true values

Latent Value Distribution	Noise Condition			
	0.0	0.5	1.0	2.0
Normal				
Elo	.996	.979	.945	.807
Rescorla–Wagner	.992	.988	.978	.929
Value	.994	.989	.980	.940
Best–Worst	.964	.972	.973	.937
ABW	.976	.978	.976	.938
F				
Elo	.821	.967	.895	.702
Rescorla–Wagner	.823	.968	.962	.882
Value	.814	.963	.968	.901
Best–Worst	.745	.846	.896	.872
ABW	.821	.966	.966	.901
Exponential				
Elo	.823	.975	.925	.784
Rescorla–Wagner	.827	.975	.971	.919
Value	.819	.974	.976	.931
Best–Worst	.774	.875	.922	.913
ABW	.814	.975	.973	.930
Uniform				
Elo	.928	.994	.981	.929
Rescorla–Wagner	.935	.988	.989	.972
Value	.936	.988	.989	.976
Best–Worst	.989	.992	.989	.975
ABW	.884	.972	.983	.973

Performance is measured in terms of R^2 between the true values and derived scores. Assumes 32,000 simulated trials. The best-performing algorithms for each noise condition are bolded, and the worst-performing algorithms are italicized. Value scoring consistently performs best under the two highest-noise conditions. The results are more varied for the two lowest-noise conditions. ABW, analytical closed-form solution to best-worst scaling

distribution-dependent. Indeed, in one case, best-worst counting even performed better than all other algorithms by a large margin (uniform distribution, no noise), with a mean R^2 of .989, as compared to 0.921 for the other algorithms. Generally, however, the results were split between either Elo or Rescorla–Wagner providing the highest-quality scores under conditions of minimal noise.

The findings replicated those from Experiment 2, that Elo scoring performed well in the absence of judgment noise and was also the algorithm most negatively affected by the presence of any judgment noise.

It is worth noting that in the noise-absent conditions, even though all the scoring algorithms showed deficiencies in capturing the distributional shapes of the underlying latent values, there is clear evidence that rank information was still being derived. For instance, although the strength of the linear correlation between Rescorla–Wagner scores and F distributed true values was only mean $R^2 = .823$, the strength of the rank relationship was mean $\rho^2 = .996$.

Discussion

Experiment 3 demonstrated that the scoring algorithms all show some robustness, to varying degrees, when nonnormal distributions are being considered, as long as the judgments are noisy. This was the hypothesized result, explained by the presence of noise implicitly generating interval information from rank judgments, in the form of the proportions of times that any two items switched relative ranks across judgments. This is a promising finding, because it suggests that natural inconsistencies in human judgments may prove to be an important feature for using best-worst judgments to derive estimates of latent values.

Even in the absence of noise, the algorithms all demonstrated some degree of success at predicting rank values. This is likewise promising, because it allows for the option of applying transformations over the scored values to fit them to the underlying latent dimension. Techniques for function approximation can be applied to these types of problems when the actual underlying distribution is unknown (e.g., Hollis & Westbury, 2006; Hollis, Westbury, & Peterson, 2006; Westbury & Hollis, 2007). The problem, of course, is that the true values along the latent dimension underlying the judgments made are also unknown. Thus, transformations would need to be found by validating transformed scores against some behavioral measure that would depend on the latent dimension being judged. This approach has already been fruitfully applied to derive estimates of various semantic constructs from co-occurrence models of semantics (e.g., Hollis & Westbury, 2016; Mandera, Keuleers, & Brysbaert, 2015).

The results from this simulation bolster a few conclusions from the previous two simulations. First, best-worst scaling is a useful judgment format for estimating true values along

some underlying latent dimension when one is interested in thousands of items. Second, the quality of the estimates hinges on what scoring algorithm is being used. Third, the best scoring algorithm appears to be situation-dependent. Whereas value scoring consistently emerged as the best algorithm under high levels of noise in the judgments, Elo, Rescorla–Wagner, and best-worst counting all presented themselves as the best algorithm under specific conditions. In some cases, the difference in quality between the best algorithm and the other algorithms was large (e.g., best-worst counting with a uniform distribution and no noise).

Thus far we have ignored the sixth hypothetical match provided by best-worst data, for which we assume no information is available. When best-worst scoring is reframed as a tournament-ranking problem, it may prove useful to treat that sixth case as a draw. The ratio of wins plus losses to draws may likewise provide interval information in the same way that noise provides interval information: Two items that are closer to each other along a latent dimension will be more likely to draw with each other than two items that are more distant from each other. A key variable of interest is the number of items presented on any particular trial. For the previous reported simulations, four items per trial were always assumed. However, as the number of items per trial increases, more tie information is available. I hypothesize that increasing the tie information available per trial will work much like increasing the noise in judgments; it will help linearize the relationship between the calculated scores and true values when nonnormal distributions are being considered. Treating the “unknown” outcomes as draws will be a future direction for research on scoring best-worst judgments.

Experiment 4

Thus far, predict algorithms have generally proved to be superior to count algorithms, with a few exceptions (see Exp. 3). The validity of count algorithms is predicated on the use of balanced incomplete block designs to construct judgment sets (Louviere et al., 2015). However, BIBDs do not scale computationally to the many-item case; they pose a combinatorial problem.

Thus far we have only considered a completely random block design in which the items available in each block are generated with no assurances about how often a given item will occur or how often it will co-occur in the same block with other items. In previous many-item best-worst experiments, blocks have been structured such that no two items occur together in more than one trial (Kiritchenko & Mohammad, 2016b), which maximizes the relational information available from best-worst data. A second design feature that is relatively easy to introduce is ensuring that each item occurs equal numbers of times across all blocks. Combinatorial explosion likely

Fig. 5 Performance of the Rescorla–Wagner scoring algorithm at predicting true latent values, distributed according to $F(100, 10)$ under varying levels of noise in the best-worst judgments. In the absence of noise, the scoring algorithm displays deficiencies in predicting the

relative distances between latent values. The presence of judgment noise, however, improved its performance. Similar patterns were seen across all other scoring algorithms

makes further substantial refinements to the block structure impractical.

Method

This experiment replicated Experiment 1, with the exception of how blocks were generated. Two manipulations to the block structure were introduced.

Manipulation 1 was the equality of item occurrences: Each item was presented in exactly the same number of blocks (equal sampling) or was randomized, as in the previous experiments (random sampling). Blocks in the random-sampling condition were constructed by sampling four unique items (with replacement). Blocks in the equal-sampling condition were generated by randomizing the order of a list containing each response item and then creating a block out of each successive sequence of four items in the list. With 1,000 items to be scored, this procedure created 250 blocks in which each item was present in exactly one block. This procedure was repeated until the required number of blocks were generated for a given simulation. For example, if a simulation was using 2,000 blocks, this procedure was repeated $2,000/250 = 8$ times, to create a series of 2,000 blocks across which each item occurred exactly eight times.

Manipulation 2 was whether or not item pairs could occur multiple times. In the *repetitions-allowed* condition, no restrictions were placed on the number of times that item pairs could occur together across blocks. In the *no-repetitions* condition, a block was resampled until all combinations of two items within the block were unique combinations for that simulation. In the case of equal sampling, resampling occurred after rerandomizing the sequence of all items that had not yet been put into a block. For the 32,000-block simulations, it became challenging to organize the items into blocks such that no item pair occurred more than once. In this situation only, the no-repetitions constraint was relaxed to allow for blocks that contained repetitions if and only if a block with no repetitions was not found after 20 resamplings of the remaining items. This resulted in 2,240 blocks that contained repeated item pairs (0.04% of all the simulated blocks).

Results and discussion

A 2 (Sampling Method) $\times$ 2 (Repetitions Allowed) $\times$ 6 (Number of Blocks) ANOVA was run on the score accuracies. The two main effects of interest were both statistically reliable, in the expected directions. First, when each item was sampled to occur an equal number of times, better scores were

produced [$F(1, 11976) = 844.94, p < 2.2\text{e-}16$]. Unsurprisingly, this effect was most pronounced for simulations with fewer blocks; an interaction between sampling method and number of blocks was also observed [$F(5, 11976) = 444.98, p < 2.2\text{e-}16$]. Second, when all pairwise combinations of items were unique within a simulation, better scores were produced [$F(1, 11976) = 4.92, p = .03$]. The results are displayed in Fig. 6.

Although balanced incomplete block designs are infeasible with large numbers of items to be scored, it is relatively easy to ensure that each item occurs an equal number of times. This demonstrably helps produce better scores. These results further suggest that ensuring no pairs of items repeat is a helpful design consideration for many-item best-worst experiments.

Experiment 5

The previous experiments reported results from simulated data. They provide a clear case that it is possible to score best-worst data in unbalanced, many-item designs with a new suite of algorithms. However, it is difficult to conclude from simulations alone whether the reported results will generalize to empirical data. Validation of the results with empirical data is necessary.

One of the benefits of simulation is that we can know the true values of items. This allows us to validate scoring algorithms by comparing the produced scores to the true values along a latent dimension of judgment. However, when behavioral data are considered, we are not privy to true values (hence the need for measurement in the first place). A different validation criterion needs to be used.

Fig. 6 Score quality as influenced by two block design features manipulated: whether or not item pairs were allowed to repeat and whether items were sampled evenly or randomly. Values are averaged over the performance for all scoring algorithms

The semantic content of a word often affects how that word is processed. For instance, it has been established that the emotional content of a word affects how long it takes people to access the meaning of that word; high valence words are recognized more quickly than low valence words, and low arousal words are recognized more quickly than high arousal words (e.g., Kuperman et al., 2014; Warriner et al., 2013). Emotional content is a good example because it has been well studied within linguistic tasks. However, as a general rule, the multitude of other semantic dimensions along which words differ in meaning all affect lexical access times (e.g., Hollis & Westbury, 2016).

In the context of using best-worst scaling to measure the meanings of words along various latent dimensions (e.g., valence, arousal, concreteness), one possible validation criterion would be to test whether scores account for variation in a behavioral measure of lexical access. For instance, by using scores to predict lexical decision reaction times (LDRTs; i.e., how long it takes a person to decide whether or not a string of letters is a word), or word-naming times.

Multiple LDRT and word-naming databases are frequently used within psycholinguistic research to test statistical models of lexical access (e.g., Balota et al., 2007; E. Keuleers, Lacey, Rastle, & Brysbaert, 2012). Multiple sets of valence norms are also available (e.g., Bradley & Lang, 1999; Warriner et al., 2013). To date, the construction of norms sets within psycholinguistics has exclusively been performed using rating scales to estimate values. The existence of these norms allows us to address another central question of this research: Is best-worst scaling more or less appropriate than other measurement techniques for collecting human judgments about the meanings of words?

The present experiment was designed to compare the effectiveness of the various scoring algorithms detailed in previous experiments on empirical data generated by having participants make best-worst judgments about word valence. This experiment also allowed comparing the quality of the calculated scores to other norms that had been compiled using rating judgments.

Method

Participants A total of 32 students (23 female, nine male) enrolled in introductory psychology at the University of Alberta volunteered for this experiment to earn partial credit in their course. The mean [*SD*] age of participants in this study was 19.43 [2.14] years. All participants spoke English as a first language.

Procedure Participants made 260 best-worst judgments over the semantic dimension of valence. On each trial, participants were presented with four words and were prompted to choose the “most pleasant” and “least pleasant” of the set.

Participants could make their choices in any order. When participants were finished with their decision, they could click a button labeled “done” in order to progress to the next trial. Participants were given a chance to take a self-timed break after every 52 trials (four breaks total). The data from multiple participants were pooled and scored using the five scoring algorithms described in previous experiments.

Stimuli A total of 1,040 words were used as stimuli in this experiment. Of these, 1,034 words came from the Affective Norms for English Words database (ANEW; Bradley & Lang, 1999), plus six additional words randomly drawn from the Warriner et al. (2013) affective norms set. Words were grouped into sets of four to make a best-worst trial. Trials were constructed in line with findings from Experiment 4: Each word appeared an equal number of times across the experiment (specifically, each word was seen exactly one time by each participant), and no two words appeared together more than once. Thus, although each participant saw the same words, no participants saw identical trials.

Validation measures LDRTs from the English Lexicon Project (Balota et al., 2007) were used to assess the relative quality of the scoring algorithms compared in this experiment. Scores that correlated more strongly with LDRTs were assumed to better capture the variation in valence between words. Valence measures from ANEW and the Warriner et al. (2013) norm set were also used to assess the overall quality of the scores produced by the various scoring algorithms.

Noncompliant participants After data were scored, a compliance measure was calculated for each participant: What proportion of the participant’s “best” and “worst” choices were consistent with expectations from scores for each word (using value scoring). Compliance would be 50% if participants were making “best” and “worst” choices randomly. Two participants had notably low compliance (57.69%, 58.84%), as compared to the median rate (98.46%). The data from these participants were removed from the analysis, at which point the data were rescored.

Simulating results for fewer human judgments Data were available for 30 participants ($260 \times 30 = 7,800$ trials). The results were simulated for fewer numbers of trials by taking 100 random subsets each, for $n = 4, 8$, and 16 participants (1,040, 2,080, and 4,160 trials, respectively). All reported R^2 values are averages over the relevant randomized subsets.

Results

Experiments 1–4 had involved hundreds of independent simulations. This allowed for statistical comparisons to be made

between the relative performances of the five scoring algorithms. In contrast, there were many fewer empirical data—too few to report on the present data with the same statistical rigor as in Experiments 1–4. Thus, the reporting here is limited to qualitative observations about the general patterns contained within these data.

In the four previous simulated experiments, log-odds transformation of the scores provided a better fit to the true values than did the raw scores. However, this finding was not replicated for the behavioral data; raw scores provided a better fit to the LDRTs than did log-odds-transformed data. Thus, only the results for raw scores are reported. These results are presented in Fig. 7.

This experiment replicated the main findings observed in the previous simulation studies. First, predict algorithms produced scores that more strongly correlated with LDRTs than did count algorithms. The exception appears to be Elo, which performed more like count algorithms with few judgments (2,080 or fewer). However, it performed comparably to other predict algorithms when more judgments were scored (4,160 or more). Elo scoring consistently produced atypical behavior in the previously reported simulations, so it is perhaps no surprise that we again see Elo display condition-contingent behavior in the present experiment. Second, out of all the scoring methods, value scoring produced the scores most strongly correlated with LDRTs, on average, across the range of possible dataset sizes. However, Elo did have marginally superior performance for the one condition in which the most number of human judgments were used. One surprising result was that ABW appears to have produced worse scores than

Fig. 7 Amounts of variance accounted for in lexical decision reaction times (LDRTs) for valence scores, calculated with each of the five scoring algorithms. Scores were calculated over different numbers of best-worst trials. For comparison, R^2 values are also reported for the ANEW (Bradley & Lang, 1999) and Warriner et al. (2013) norms

best-worst counting on the present behavioral data; ABW scores correlated less strongly with LDRTs than did best-worst scores.

When considering all the data from this experiment, Elo produced the scores most strongly correlated with LDRTs [$r(1025) = -.266, p < 2.2e-16$], and ABW produced the scores least strongly correlated with LDRTs [$r(1025) = -.259, p < 2.2e-16$]. A comparison of these two correlation coefficients using the Fisher r -to- z transformation revealed no statistical difference in magnitude ($z = 0.16, p = .43$). However, the general pattern of results does suggest that predict algorithms provide scores more strongly correlated with LDRTs than do count algorithms, which is consistent with the previous experiments; perhaps the differences in correlation strength are too small in this particular case for the r -to- z test to detect.

Considered over the same set of words, the ANEW norms correlated with LDRTs at $r(1025) = -.235 (p < 2.2e-16)$, and the Warriner et al. (2013) norms correlated with LDRTs at $r(1025) = -.266 (p < 2.2e-16)$. Fisher r -to- z tests likewise revealed no statistical difference in magnitude between these correlations and the correlation between the best-performing scoring algorithm in this experiment (Elo) and LDRTs. However, the shapes of the functions relating the number of trial judgments to R^2 (Fig. 7) suggest that scores produced by the five scoring algorithms would continue to improve with more data collection. With more data, they would possibly produce scores that are statistically a better fit to LDRTs than are either the ANEW or the Warriner et al. norms.

Discussion

The results of this experiment corroborate the main finding from Experiments 1–4. Namely, predict algorithms produce scores more strongly correlated with a validation measure than do count algorithms. Each of the predict algorithms takes into account the quality of competition when computing scores, whereas count algorithms do not. Thus, the count algorithms ignore useful information when scoring data. Also in line with the previous experiments, value scoring appears to be a robustly applicable scoring algorithm for best-worst data. Also, Elo scoring appears to produce scores of inconsistent quality across varied parameter values.

The present experiment additionally allowed for a comparison between valence measures derived from rating scales and best-worst scaling. With 7,800 best-worst trials worth of data, the produced scores were as strongly correlated with LDRTs as were both the ANEW and the Warriner et al. (2013) valence estimates. In fact, with as few as 2,080 judgments, scores were also comparable to the ANEW and Warriner et al. (2013) valence estimates.

Both the ANEW and Warriner et al. (2013) norms sets estimated word valence by averaging 9-point rating scale responses. Although the ANEW norms do not provide data

on the exact number of times each word was rated, the Warriner et al. (2013) data do. For the words studied in the present experiment, Warriner et al. (2013) reported a mean of 41.97 judgments per word. That sums to 43,649 judgments for 1,040 words. In other words, scores of equal quality were produced here using 82.13%–95.23% fewer judgments. On a crowdsourcing platform like Mechanical Turk, it would be reasonable to pay participants US \$0.02 per decision. Thus, the difference can be monetarily quantified as US \$872.98 versus \$41.60–\$156.00. Recent crowdsourcing efforts to create norms sets have in fact collected data for many more words. For example, Warriner et al. (2013) collected valence, arousal, and dominance judgments for 14,000 words, and Brysbaert, Warriner, and Kuperman (2014) collected concreteness judgments for 40,000 words. At that scale, the savings would be closer to \$10,000 per experiment.

As a measurement tool, rating scales are subject to the problem that people adopt response strategies when using them, particularly when making decisions for extreme values (Saal et al., 1980). These response strategies add systematic variation to the estimates that is not reflective of the underlying value to be estimated. By virtue of being based on ordinal decisions, it is unlikely that best-worst scaling would suffer from this source of bias. It is possible that best-worst scaling might actually prove to be the superior measurement tool when having people evaluate items along a latent dimension related to word meaning. In line with this claim, Fig. 7 indicates that best-worst scaling might produce valence scores that more strongly correlate with behavioral measures of lexical access than do either the Warriner et al. (2013) or the ANEW norms, given sufficient data. Further testing will be required.

In the present experiment, participants were requested to choose the “most pleasant” and “least pleasant” option from a set of four options. This is a conventional word choice for eliciting human judgments of valence (e.g., Bradley & Lang, 1999; Warriner et al., 2013). However, some of the words participants have to make judgments of are more extreme in nature (e.g., *rape*, *murder*) than the provided anchors (*pleasant*, *unpleasant*). This may require participants to recalibrate their conceptualizations of the scale as more of these extreme words are encountered (Westbury, Keith, Briesemeister, Hofmann, & Jacobs, 2015). If such were the case, measurement error would be added to the collected judgments; within a data collection session for a single participant, middling words (such as *pleasant* and *unpleasant*) would start off being rated as extreme, but as more *actual* extreme words (e.g., *murder*) were encountered, the ratings assigned to these middling words would become more moderate. Thus, the use of the terms *pleasant* and *unpleasant* as part of the instructions in the present experiment may have been a source of error in the data.

There are at least two reasons why the above critique should not be a point of concern for the present experiment. First, the possibility of recalibration in valence decisions when using the dimensional anchors of pleasant and unpleasant was explicitly tested by Warriner, Shore, Schmidt, Imbault, and Kuperman (2017). They found no evidence that words with middling valences were judged to be more extreme at the beginning of a data collection session. Second, during a best-worst judgment, participants are not presented with a scale that possesses an absolute minimum and an absolute maximum that need to be anchored; since participants are always making judgment relative to other items, there is no possibility of anchoring effects, only the possibility of misconstruing the instructions. However, it would be a feat of exceptional pedantry for a participant to judge the word *pleasant* as being “more pleasant” than *nirvana*.

General discussion

The main contribution of this work has been to introduce and validate three new algorithms for scoring best-worst data. Through simulation studies, each of the three new algorithms was demonstrated to produce scores that reconstructed with high accuracy the rank and distance of true values along a latent dimension underlying the best-worst judgments. Importantly, these algorithms can be applied to derive unbiased scores in situations in which many items need to be scored and BIBDs cannot be used. Highly accurate scores could be derived here with surprisingly few data (1–4 judgments for each item; Exp. 1), and near-perfect scores were reliably produced with slightly more data (8–32 judgments per item).

This study has demonstrated that the three new scoring algorithms are generally robust in the presence of noisy judgments and nonnormal distributions of true values. Overall, no scoring algorithm emerged as universally the best. Elo scoring presents itself as a good option in situations in which few judgments are available and/or noise is absent in the judgments. However, value scoring performed consistently better than other algorithms in cases in which large amounts of noise were present in the judgments. There were also conditions in which either Rescorla–Wagner scoring or best-worst counting was clearly the most accurate scoring algorithm. Thus, the conclusion here is that some situations favor each scoring algorithm and that, in practice, each scoring algorithm should be used. The best scores for a particular instance should be selected on the basis of some set of validation criteria.

The results of Experiment 4 provide prescriptive advice for designing many-item best-worst experiments. The highest-quality scores were derived when each item occurred an equal number of times across blocks. As such, researchers should try to ensure that all items to be scored occur an equal number of times across blocks. Previous researchers have also tried to

ensure that no item pairs occurred together multiple times across blocks (Kiritchenko & Mohammad, 2016b), which maximizes the relational information available in best-worst data. Experiment 4 suggests that this design consideration has a measurable, albeit marginal, impact on the quality of the derived scores within the boundaries of our tested parameter settings.

One concern with applying best-worst scaling to collect human judgments is that it only produces ordinal information. In domains in which interval information is required, best-worst scaling may not be as applicable as response formats like, for instance, rating scales. That said, Experiment 2 does demonstrate that interval information can be inferred from best-worst data when sufficient noise is present in the judgments. Experiment 3 additionally points to ways that “tie” outcomes may be used to infer interval information, even in the absence of judgment noise. Finally, Experiment 5 provides empirical data on the applicability of best-worst scaling to a domain in which interval information is required. The results suggest that within the domain of estimating word valence, best-worst scaling produces scores that are at least as good as those produced by rating scales, while also requiring many fewer data to calculate those scores.

Both psychologists and NLP researchers have had recent success using crowdsourcing to estimate latent semantic values for tens of thousands of words. However, these types of experiments are costly. Any methodological advancement that can reduce these costs will be useful for the advancement of large-scale research on semantics. It will be valuable to pursue best-worst scaling as a judgment format in such tasks, if for no other reason than that it can potentially allow researchers to reduce the costs of such research. Experiments 1–4 all demonstrated through simulation that high-quality item scores can be derived from a minimal number of judgments per item. Experiment 5 demonstrated that these findings generalize to empirical data. Best-worst scaling thus should have broad utility in the research context of inferring the latent semantic properties of items.

Author note Thank you to Jordan Louviere for helpful discussion on scoring best-worst judgments. Thank you to Marc Brysbaert, Emmanuel Keuleers, Svetlana Kiritchenko, Pawel Mandera, Saif Mohammad, and Chris Westbury for feedback on earlier drafts of the manuscript.

References

- Abercrombie, H. C., Kalin, N. H., Thurow, M. E., Rosenkranz, M. A., & Davidson, R. J. (2003). Cortisol variation in humans affects memory for emotionally laden and neutral information. *Behavioral Neuroscience*, 117, 505.
- Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30, 1174–1220.
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English Lexicon Project.

- Behavior Research Methods*, 39, 445–459. doi:[10.3758/BF03193014](https://doi.org/10.3758/BF03193014)
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings (Technical Report C-1)* (pp. 1–45). Gainesville: University of Florida, Center for Research in Psychophysiology.
- 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.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904–911. doi:[10.3758/s13428-013-0403-5](https://doi.org/10.3758/s13428-013-0403-5)
- Elo, A. E. (1973). The international chess federation rating system. *Chess*, 38, 293–296. 38(August), 328–330; 39(October), 19–21.
- Hamann, S., & Mao, H. (2002). Positive and negative emotional verbal stimuli elicit activity in the left amygdala. *NeuroReport*, 13, 15–19.
- Hollis, G., & Westbury, C. (2006). NUANCE: Naturalistic University of Alberta nonlinear correlation explorer. *Behavior Research Methods*, 38, 8–23. doi:[10.3758/BF03192745](https://doi.org/10.3758/BF03192745)
- Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, 23, 1744–1756. doi:[10.3758/s13423-016-1053-2](https://doi.org/10.3758/s13423-016-1053-2)
- Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, 70, 1603–1619. doi:[10.1080/17470218.2016.1195417](https://doi.org/10.1080/17470218.2016.1195417)
- Hollis, G., Westbury, C. F., & Peterson, J. B. (2006). NUANCE 3.0: Using genetic programming to model variable relationships. *Behavior Research Methods*, 38, 218–228. doi:[10.3758/BF03192772](https://doi.org/10.3758/BF03192772)
- Imbir, K. K. (2015). Affective norms for 1,586 Polish words (ANPW): Duality-of-mind approach. *Behavior Research Methods*, 47, 860–870.
- 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, 287–304. doi:[10.3758/s13428-011-0118-4](https://doi.org/10.3758/s13428-011-0118-4)
- Keuleers, M., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68, 1665–1692.
- Kiritchenko, S., & Mohammad, S. M. (2016a). *Sentiment composition of words with opposing polarities*. San Diego: Paper presented at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).
- Kiritchenko, S., & Mohammad, S. M. (2016b). *Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling*. San Diego: Paper presented at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).
- Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50, 723–762.
- 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, 1065–1081. doi:[10.1037/a0035669](https://doi.org/10.1037/a0035669)
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44, 978–990. doi:[10.3758/s13428-012-0210-4](https://doi.org/10.3758/s13428-012-0210-4)
- Lipovetsky, S., & Conklin, M. (2014). Best-worst scaling in analytical closed-form solution. *Journal of Choice Modelling*, 10, 60–68.
- Lodge, M., & Taber, C. S. (2005). The automaticity of affect for political leaders, groups, and issues: An experimental test of the hot cognition hypothesis. *Political Psychology*, 26, 455–482.
- Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge: Cambridge University Press.
- Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly Journal of Experimental Psychology*, 68, 1623–1642.
- Marley, A. A. J., & Islam, T. (2012). Conceptual relations between expanded rank data and models of the unexpanded rank data. *Journal of Choice Modelling*, 5, 38–80.
- Marley, A. A. J., Islam, T., & Hawkins, G. E. (2016). A formal and empirical comparison of two score measures for best-worst scaling. *Journal of Choice Modelling*. doi:[10.1016/j.jocm.2016.03.002](https://doi.org/10.1016/j.jocm.2016.03.002)
- Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. arXiv: 1308.6242.
- Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, 46, 887–903. doi:[10.3758/s13428-013-0405-3](https://doi.org/10.3758/s13428-013-0405-3)
- Ogden, J., & Lo, J. (2012). How meaningful are data from Likert scales? An evaluation of how ratings are made and the role of the response shift in the socially disadvantaged. *Journal of Health Psychology*, 17, 350–361.
- Orme, B. (2005). *Accuracy of HB estimation in MaxDiff experiments (Sawtooth Software Research Paper Series)*. Sequim: Sawtooth Software, Inc.. Retrieved from www.sawtoothsoftware.com/download/techpap/maxdacc.pdf
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.
- Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2016). The Calgary Semantic Decision Project: Concrete–abstract decision data for 10,000 English words. *Behavior Research Methods*. doi:[10.3758/s13428-016-0720-6](https://doi.org/10.3758/s13428-016-0720-6)
- Raschka, S. (2015). *Python machine learning: Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*. Birmingham: Packt.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413. doi:[10.1037/0033-2909.88.2.413](https://doi.org/10.1037/0033-2909.88.2.413)
- Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5, 3–55.
- Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49, 111–123. doi:[10.3758/s13428-015-0700-2](https://doi.org/10.3758/s13428-015-0700-2)
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45, 1191–1207. doi:[10.3758/s13428-012-0314-x](https://doi.org/10.3758/s13428-012-0314-x)
- Warriner, A. B., Shore, D. I., Schmidt, L. A., Imbault, C. L., & Kuperman, V. (2017). Sliding into happiness: A new tool for measuring affective responses to words. *Canadian Journal of Experimental Psychology*, 71, 71–88. doi:[10.1037/cep0000112](https://doi.org/10.1037/cep0000112)
- Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27, 236–247.

- Westbury, C. F., & Hollis, G. (2007). Putting Humpty together again: Synthetic approaches to nonlinear variable effects underlying lexical access. In G. Jarema & G. Libben (Eds.), *The mental lexicon: Core perspectives* (pp. 7–30). Bingley: Emerald.
- Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *Quarterly Journal of Experimental Psychology*, 68, 1599–1622.
- Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: On emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, 4, 991. doi:[10.3389/fpsyg.2013.00991](https://doi.org/10.3389/fpsyg.2013.00991)
- Zhu, X., Kiritchenko, S., & Mohammad, S. M. (2014). NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets. In P. Nakov & T. Zesch (Eds.), *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 443–447). New York: Association for Computational Linguistics.