Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Can J Exp Psychol. Author manuscript; available in PMC 2018 March 01.

Published in final edited form as:

Can J Exp Psychol. 2017 March ; 71(1): 71–88. doi:10.1037/cep0000112.

Sliding into happiness: A new tool for measuring affective responses to words

Amy Beth Warriner[1,2], **David I. Shore**[1,2], **Louis A. Schmidt**[1], **Constance L. Imbault**[2], and **Victor Kuperman**[1,2]

[1]Department of Psychology, Neuroscience & Behaviour, McMaster University

[2]Department of Linguistics and Languages, McMaster University

Abstract

Reliable measurement of affective responses is critical for research into human emotion. Affective evaluation of words is most commonly gauged on multiple dimensions—including valence (positivity) and arousal—using a rating scale. Despite its popularity, this scale is open to criticism: it generates ordinal data that is often misinterpreted as interval, it does not provide the fine resolution that is essential by recent theoretical accounts of emotion, and its extremes may not be properly calibrated. In five experiments, we introduce a new *slider* tool for affective evaluation of words on a continuous, well-calibrated and high-resolution scale. In Experiment 1, participants were shown a word and asked to move a manikin representing themselves closer to or farther away from the word. The manikin's distance from the word strongly correlated with the word's valence. In Experiment 2, individual differences in shyness and sociability elicited reliable differences in distance from the words. Experiment 3 validated the results of Experiments 1 and 2 using a demographically more diverse population of responders. Finally, Experiment 4 (along with Experiment 2) suggested that task demand is not a potential cause for scale recalibration. In Experiment 5, men and women placed a manikin closer or farther from words that showed sex differences in valence, highlighting the sensitivity of this measure to group differences. These findings shed a new light on interactions among affect, language, and individual differences, and demonstrate the utility of a new tool for measuring word affect.

Keywords

valence; arousal; sex differences; individual differences

Given the fundamental contribution of emotion to human psychology and behavior, it is critical for psychological research to have reliable ways of evaluating emotional responses. Ratings scales applied to individual words provide on commonly used measure of emotional response. These scales assume that emotions are multi-dimensional psychological states (e.g., Osgood, Suci, & Tannenbaum, 1957) that can be decomposed into core dimensions (e.g., valence, arousal, dominance); each word represents a point in a multi-dimensional

Address correspondence to: Victor Kuperman, Department of Linguistics and Languages, McMaster University, Togo Salmon Hall 626, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4M2, vickup@mcmaster.ca, Phone: 905-525-9140, x. 20384, Fax: 905-525-9140, x. 20384.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

space that maps out emotion as distinct psychological states. Until recently, most studies of emotion and language in English relied on Bradley and Lang's (1999) early collection of ratings for 1,034 words on separate scales relating to three difference dimensions of emotion, a collection known as Affective Norms for English Words (ANEW): for important alternatives, see work based on Pennebaker, Francis, and Booth (2001). One dimension concerns the valence or positivity of the emotions invoked by a word, going from unhappy to happy. The second addresses the degree of arousal evoked by a word, and the third refers to the dominance/power of the word—the extent to which the word denotes something that is weak/submissive or strong/dominant. The last decade saw an expansion of this dataset to about 14,000 English words (Adelman & Estes, 2013; Adelman, Marquis, Sabatos-DeVito & Estes, 2013; Kloumann, Danforth, Harris, Bliss, & Dodds, 2012; Mohammad & Turney, 2010; Warriner, Kuperman, & Brysbaert, 2013), as well as the creation of similar datasets in other languages (see Warriner et al., 2013, for an overview).

Importantly, the basic procedure utilized by Bradley and Lang (1999) remains virtually unchanged across all studies mentioned above. Bradley and Lang (1999) used a series of self-assessment manikins that represented a scale from 1 to 9 to measure valence, arousal and dominance, with 1 corresponding to positive/calm/weak and 9 corresponding to negative/aroused/strong respectively. Further studies have employed the same procedure with minor modifications (e.g., manikins were replaced with a numeric rating scale in Adelman et al. (2013), Kloumann et al. (2012), and Warriner et al. (2013); the scale for positivity was reversed in direction and reduced to 7 instead of 9 points in Adelman et al. (2013), such that 1 denotes "negative" and 7 "positive"; and the wording of instructions has been slightly altered across papers): for a detailed discussion of instructions and the scale, see below.

The utility and popularity of affective ratings to words in psychological and computer science research is undeniable. As discussed in Warriner et al. (2013), at least four lines of research rely heavily on affective ratings to words: research on emotions themselves, research on the impact that emotion has on language production and comprehension, research on *sentiment analysis* (see review by Liu, 2012), and research on computational estimates of emotional variables for new words (e.g., Bestgen & Vincze, 2012; Recchia & Louwerse, 2015; Westbury, Keith, Briesemeister, Hofmann, & Jacobs, 2015; but see Mandera, Keuleers & Brysbaert (2015) for criticism of the method and alternative computational approaches). Citation patterns reflect this wide-spread use of affective ratings: for example, Bradley and Lang's (1999) dataset has accumulated 1723 citations, Mohammad and Turney (2013) 169 citations, and Warriner et al. (2013) 166 citations on Google Scholar (as of 30-May-2016).

In view of the massive body of research that relies on affective ratings derived using the ANEW dataset, it may be tempting to treat one (perhaps the most exhaustive) set of such ratings as a "gold standard", a benchmark that behavioral responses to emotion as well as their computational estimates might be compared against. However, we argue—along with earlier proposals (e.g., Westbury et al., 2015)—that the current procedure of collecting affective ratings does not fully warrant this privileged status. There are at least three reasons

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for this argument. In what follows, we outline these reasons and introduce a technique that eliminates or alleviates them.

Why is the current procedure of the affective evaluation task suboptimal? First, Bradley and Lang's (1999) method uses a rating-scale, as do virtually all norming studies developed since. Importantly, this is an ordinal scale, rather than an interval one. While intervals between positions on an ordinal scale are monotonic, they are generally not uniform, and so a difference in the underlying affective responses reflected by ratings 1 and 2 may not be equivalent to a difference reflected by ratings 5 and 6. The ordinal nature of a rating scale generally precludes the application of parametric statistics to analyze the obtained responses (Allen & Seaman, 2007; Jamieson, 2004), starting from reporting their mean values to applying variance or regression analyses to either the raw or aggregated ratings. Yet the use of parametric statistics on emotional norms characterizes both the datasets (Bradley & Lang, 1999) and the literature that uses the norms for the study of emotion (see for instance Kuperman, Estes, Brysbaert, & Warriner, 2014; Warriner et al., 2013). Severity of statistical violations that come from treating ordinal scales as interval is heavily debated (cf., Allen & Seaman, 2007; Jamieson, 2004; Knapp, 1990; Norman, 2010; Wang, Yu, Wang, & Huang, 1999) and is contingent on a number of factors (including the number of positions on the scale, as well as the distribution of values and size of the sample). A true interval scale to measure behavioral outcomes affective evaluation would resolve this issue.

Second, it is unclear whether even the most fine-grained rating scale using 9 points (as in Bradley & Lang, 1999 and several derived efforts, or a 7-point scale in Adelman et al., 2013) offers sufficient resolution to differentiate between emotional states that words may evoke. Recent regression studies of word recognition and word recognition memory (e.g., Adelman & Estes, 2013; Kousta, Vinson, & Vigliocco, 2009; Kuperman et al., 2014) reported results compatible with the notion that emotionality of words has a gradient effect on human language behavior, with even subtle differences in positivity or arousal translating into discernible behavior outcomes (see the hypothesis of gradient automatic vigilance in Kuperman et al., 2014). Thus, a substantial increase in the resolution of the scale from a maximum of 9 intervals is another desideratum for a refined task.

A third point of criticism against Bradley and Lang's (1999) method concerns how the scale for emotional response is anchored, and whether raters are able to adequately calibrate their responses using the task instructions. The point is well articulated by Westbury et al. (2015), who generated affective estimates for over 23,000 words by applying a computational lexical co-occurrence model HiDEx (Shaoul & Westbury, 2010) to natural texts and estimating the semantic distance from all target words to a small set of basic emotion labels. A function of those distances generates a "fixed" estimate of valence, arousal and dominance for about 23,000 words, an estimate that is not based on intuitions of human raters and is perfectly reliable as long as it is obtained with the same model and the same corpus[1]. Westbury et al.'s comparison of human ratings to their computationally derived measures of emotionality reveal that human ratings of word valence in Warriner et al. (2013) have a much stronger

[1]A more up-to-date set of computational estimates by Hollis, Westbury, & Lefsrud (2016) was not released at the time of this paper's submission and is not accounted for here.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

correlation with the word's semantic distance from the emotion label PLEASANT than from the emotion label UNPLEASANT. The reason for this imbalance, as suggested by Westbury et al., is that the scale for valence is not properly anchored—extremes of the scales are associated with mild, colloquial terms like "pleasant" and "unpleasant". Yet word lists that raters encounter include words like "homicide", "rapist", and "castration", i.e. "things that are unquestionably very much worse than just unpleasant." (Westbury et al., 2015, p. 1615). As a result, Westbury et al. (2015, p. 1615) hypothesize the following rating strategy:

> Faced with the impossible task they have been given, we suspect that what subjects do (quite sensibly) is to recalibrate the scale they were given so that "Unpleasant" is taken to mean (what it does not mean) "Absolutely terrible" and "pleasant" is taken to mean (what it does not mean) "Absolutely wonderful". After that necessary recalibration, the alleged anchors "Pleasant" and "Unpleasant" (and similar terms that are used to anchor the valence ratings scales) will be middling words clustered close to the centre of a scale that actually goes from "Really terrible" to "Really wonderful".

Given that instructions in Warriner et al. (2013) replicate with very minor modifications instructions to the ANEW dataset (Bradley & Lang, 1999), the 'recalibration' issue might plague an entire line of research that generated human affective norms. A closer look at instructions to the two studies examined in Westbury et al. (2015) reveal that both Adelman et al. (2013) and Warriner et al. (2013) specifically describe the extremes of the scale as choices to be made when one feels "extremely negative/positive" (Adelman et al., 2013) or "completely unhappy/happy" (Warriner et al., 2013), see Supplementary materials S1. Also, every rating task in Warriner et al. (2013) began with a practice set of 10 anchor words that introduced the rater to very positive (e.g., *free, joke*), very negative (*jail, invader*) and relatively neutral words, before launching the experimental task. Yet it is possible that a description of low valence as a state of being "sad, scared" and of high valence as "happy, contended" in the instructions leads raters to start an experiment by anchoring their relatively mild emotional responses to extremes of the scale. Below, we explore whether recalibration is indeed a valid concern for the ratings scale using ANEW instructions, as well as for other tasks that implement altered instructions and proper anchoring of the scale.

In sum, the established behavioral method of collecting affective norms to words via rating-scales is open to criticism based on at least three of its characteristics: the ordinal nature of the measurement it produces and the concomitant statistical constraints, the coarse-grained resolution afforded by a discrete scale with a total of 7 to 9 points, and a potentially inadequate anchoring of extreme points of the scale resulting from unclear instructions. The present study aimed to either eliminate or alleviate these issues by presenting a new experimental paradigm. In this new *slider* task, participants moved a manikin—a stick figure of a person—towards or away from a word presented at the top or the bottom of the screen. The distance between the word and the manikin at the time when the response was submitted served as our metric of valence. We predicted that participants would place the schematic figure closer to positive items, and further from negative items.

The slider scale was continuous with the available response range covering 600 pixels, and had a spatial resolution of 1 pixel: thus it met the desiderata of a fine-grained interval scale.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We also altered instructions in several ways to achieve proper anchoring of the new scale extremes: see below. Our goals were to (i) validate the new tool as a reliable index of words' valence in both homogeneous and heterogeneous populations, under more and less controlled experimental conditions, and (ii) explore whether it is sensitive to group (i.e., sex) and individual (i.e., personality) differences in valence evaluation. We also discuss the utility of the tool for research questions that are unattainable through ratings scale data collection.

We conducted five experiments using the new paradigm. In Experiment 1, we introduced the paradigm and established its validity as a metric of valence. Prior studies of affective norms demonstrate a strong U-shaped functional relationship between valence and arousal (words with more extreme valence are more arousing; Bradley & Lang, 1999; Kuppens, Tuerlinckx, Russell, & Barrett, 2013; Warriner et al., 2013). To examine a possible influence of arousal, we collected responses to a balanced set of words representing the entire range of both dimensions. If the new slider tool generates a valid metric of valence, we predict a strong correlation between valence ratings obtained via rating scales and distances between the manikin and the word. We further predicted that arousal would most likely interact with valence, with more extreme distances being made in response to congruent affect (positive, calm and negative, arousing respectively) than to incongruent affect, which may cause a conflict-induced amelioration of distance.

In Experiment 2, we examined personality variables (e.g., shyness and sociability) to test whether individual differences equally affected valence ratings collected using the conventional ANEW-like procedure and valence-driven distances from the word obtained in the new task. We were specifically interested in behavioral approach or behavioral inhibition styles (Carver & White, 1994). Previous research (Puca, Rinkenauer, & Breidenstein, 2006) has highlighted the potentially mediating factor that temperament can have on affect–behaviour relations: participants with a high avoidance temperament did not show an effect for positive over negative words in a joystick task, whereas those with low avoidance temperament did. In another study, participants with high social anxiety were faster to push a joystick in response to both smiling and angry faces, whereas non-anxious controls showed no difference, despite the fact that valence ratings of those same faces did not differ between groups (Heuer, Rinck, & Becker, 2007). In a third study, stronger self-reported spider-related fear was correlated with slower approach responses with a joystick (Reinecke, Becker & Rinck, 2010). In view of the affinity between manipulating distance to the object with a joystick and distance to the word as a behavioral outcome of our slider task, we expected individual differences in relevant personality variables (e.g., shyness and sociability; Schmidt & Buss, 2010) to influence behavioral outcomes in our task also, and thus included personality traits as covariates in our models.

We predicted that shy people would adopt a hesitant stance and thus stay further away from all stimuli; in contrast, sociable people would adopt an exploratory stance and get closer to stimuli.

Experiments 1 and 2 used an undergraduate participant pool, which restricted the age range and several other social characteristics to those typical of an undergraduate population. Also, due to the sex imbalance in the pool, Experiment 2 only used women as responders. To

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

remedy these limitations, Experiment 3 validated the results from Experiments 1 and 2 against a more diverse population recruited online. Furthermore, Experiment 4 addressed the issue of poor anchoring in the task instructions which have been claimed to introduce unnecessary task demand and recalibration (Westbury et al., 2015), and aimed at validating the results of Experiments 1–3.

Finally, in Experiment 5, we tested whether the well-documented sex differences in affective ratings to words (Bradley & Lang, 1999; Warriner et al., 2013) translate into sex differences in distances obtained with this new paradigm. We predicted that participants would choose to move the manikin closer to those words that were rated more positively by their sex.

Experiment 1

Here we tested whether the task would reveal a systematic relation between distance of the manikin from the presented word and valence or arousal. The words were chosen to represent the full range of valence and arousal and to ensure that these two factors were uncorrelated (cf. Warriner et al., 2013).

Method

Participants—Forty-six students at McMaster University participated in this experiment in exchange for partial course credit. The data from three participants were removed for not making a response on more than 25% of trials. The data from two participants were removed for not being native English speakers. The remaining 41 (34 women, 7 men) participants ranged in age from 17 to 25 years (M = 19.10 years, SD = 2.00).

Affective Stimuli—We selected 13,763 words with valence and arousal ratings from Warriner et al. (2013) that also had frequency information available from a 51 million-token corpus SUBTLEX based on subtitles to the US films and media (Brysbaert & New, 2009). We further restricted this word set to monosyllabic words, to constrain their variability in length and phonological complexity. Remaining words were divided into 25 bins, by crossing quintiles of valence and arousal. We selected 10 words from each bin for a total of 250 words. In this subset, valence and arousal were uncorrelated (r_s = −0.019, p> .05). The mean word length was 4.4 characters (range [3, 6]). Mean natural-log SUBTLEX frequency was 6.3 (range [3.1, 10.7]). Frequency was not reliably correlated with valence or arousal (ps > 0.5).

Procedure—Participants were tested in groups of up to ten at a time in a computer lab. Each participant was seated in front of a monitor with a resolution of 1024×768 pixels. Responses were made with a mouse. The experiment was programmed using the Experiment Builder software (SR Research, Kanata, ON, Canada).

Participants answered demographic questions about age, sex (with available options “Male”, “Female”, “Other”), handedness, and education level. Then they were instructed as follows:

> [… On a] screen, you will see a word at the top of the screen with a vertical line below it. There will be a person in the centre of that line. The person represents you. Your job is to assess how close you would like to be to the word and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

> communicate that by clicking a point on the line to position the person (you). For example, if the word was DISASTER, you'd probably want to be far away and would click somewhere on the line far away from the word. But if the word was TRIUMPH you might want to be close and would somewhere on the line really close to the word. […]

Importantly for the potential issue with recalibration (Westbury et al., 2015), current instructions eliminated all adjectives (e.g., “scared, contended”) used in the ANEW instructions to describe the emotional state that would lead to an extreme response. One word was chosen for each anchor: an extremely negative word (DISASTER, mean valence = 1.71) and an extremely positive one (TRIUMPH, mean valence = 7.34), as evaluated by raters in Warriner et al. (2013).

The manikin was initially centered at the 400th pixel of the slider scale (see Figure 1 for a sample screenshot). Participants were able to move the manikin either by clicking on its new position or by dragging and dropping the manikin at that new position. After five practice words, participants were asked if they had any questions before proceeding with the remaining stimuli. Each participant saw all 250 words. Order of word presentation was randomized for every participant: all words appeared on top of the scale. The experiment took approximately 30 minutes and was counterbalanced in its order of presentation with an unrelated 30-minute experiment (when presentation order was entered into the models, it was not significant).

Variables—The dependent variable of interest consisted of the distance (in pixels) from the center of the anthropomorphic manikin in its final position to a line just below the word. The distance occupied a range of 600 pixels, from 100 pixels (closest to the word) to 700 pixels (farthest from the word). Participants could move the manikin as many times as they chose by clicking on different points on the line, or by dragging and dropping (sliding) the manikin. The manikin’s final position when the participant clicked the ‘Continue’ button was the only variable of interest. Manikin positions after each mouse button release, the number of such releases, and response time for each release were all recorded, but did not shed any additional light on the emotional effects on participants’ responses and are not reported here.

Independent variables were valence and arousal ratings for each stimulus word (Warriner et al., 2013). As the ease of word recognition is a plausible modulator of any behavioral response to a word, we also included word frequency from the 51 million-token SUBTLEX corpus and word length as statistical controls. Word length did not affect performance in the task and is not reported further.

Statistical Analyses—We used linear mixed-effects multiple regression models with participants and words as crossed random effects (cf., Baayen, Davidson, & Bates, 2008; Pinheiro & Bates, 2000), as implemented in package lme4 version 0.999999–2 (Bates Maechler, Bolker & Walker, 2013) for R version 3.0.1 (R Core Development Team, 2015). This method enables a simultaneous exploration of multiple factors and covariates, while accounting for between-participants and between-items variance. Each model was initially

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

fitted with a maximal random-effects structure (Barr, Levy, Scheepers, & Tily, 2013) and trimmed down to only contain the random effects that significantly improve the model's performance, as indicated by a series of likelihood ratio tests that compared a model with a given random effect and a model without this random effect. Using the same test in the backwards elimination procedure, we removed from the models all fixed effects that did not improve the model's performance. No model reached a harmful level of collinearity (condition index < 13). To reduce the influence of outliers, the frequency estimates were (natural) log-transformed, as indicated by the Box-Cox power transformation test.

In all analyses below, we examined the effect of valence and arousal on the manikin's distance from the word using a range of parametric tests, including the mixed-effects multiple regression. Even though this practice is suboptimal for these ordinal variables (see discussion above), no alternative non-parametric regression techniques exist that would meet the statistical assumptions associated with the ordinal scale. Thus, somewhat ironically, we have to retreat to the practice we criticize and acknowledge a potential for inaccuracy in the present analyses. The new slider scale is designed precisely to eliminate this methodological inconsistency in the future.

Results and Discussion

Although we instructed participants to click on (or drag and drop) the slider on every trial, even if they wanted to leave the anthropomorphic manikin in the center, some did not and simply clicked 'Continue' (we had not programmed a failsafe that would prevent this). We ran our analyses with the full set of data and a set that was trimmed in two ways: 1) we removed participants who failed to register their response more than 25% of the time (see 'Participants'), and 2) we removed about 1% of the trials with RTs that are so short as to preclude a visual inspection of the target word (less than 80 ms) and improbably long (over 5 s). Then trials with RTs that were more than 2.5 SD from the participant's mean RT were removed too (2.1% of trials)[2]: for the trimming procedure see Baayen and Milin (2010). The results patterned the same for both the untrimmed and the trimmed datasets, and thus we report analyses of the trimmed data (9,859 trials) below. Tables S1 and S2 in the Supplementary materials summarize fixed and random effects of the linear mixed-effects multiple regression model.

When individual responses were considered, distance of the manikin from the word was negatively and linearly related to the valence of the word (Pearson's $r = -0.62$, df = 9,857, 95% CI [−0.63, −0.61], $p < .001$): distance to positive words was shorter. When distance was averaged across participants for each word, results revealed an even stronger relation ($r = -.86$, df = 248, 95% CI [−0.82, −0.89], $p < 0.001$) between the word's valence and slider position. This tendency was confirmed in the linear mixed-effects multiple regression model which estimated the effect of valence on the distance over and above other predictors and over individual variability in the average distance from the manikin and the strength of valence effect [$b = -92.3$, $SE = 4.8$, $t = -19.0$, $p < .001$]. The preference for approaching

[2]Although RT was not used in any of the subsequent analyses, it was representative of how on-task participants were—whether they became distracted or were failing to actually read the words. For any trials where no movement of the manikin occurred, we used the RT of pressing the 'Continue' button in place of the RT of the first click on the manikin.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

positive words and withdrawing from negative ones was strong: Each point on the 1–9 valence scale corresponded to about 90 pixels, or 15% of the available distance range. This experiment provided data to examine the issue of recalibration in the slider task: we discuss this issue using these and other data in Experiment 2.

Arousal did not influence the manikin's distance from the word, nor did it interact with valence to influence that distance (all $|t|$-values < 1.5 in the regression models, model with the interaction not shown). Frequently occurring words were approached more readily than uncommon words (Pearson's $r = -0.33$, df = 248, 95% CI [−0.34: −0.31], $p < .001$), even when valence was controlled for in the regression model [$b = -15.1$, $SE = 2.6$, $t = -5.7$, $p < .001$]. For each unit of log frequency, the manikin moved closer to the word by 15 pixels, or 4% of the available range: we elaborate on frequency effects in the General Discussion.

Negative correlations between by-participant adjustments to intercepts and slopes in the random effects structure (Column 3 in Table S2 in Supplementary materials) pointed to individual variability in behavioral responses. Participants who tended to keep a larger distance from the word overall were also more sensitive to the effects of both valence and frequency of the word. For these participants, an increase in one unit of valence (or frequency) translated into a bigger reduction of the distance from the word. The same increase in valence had a weaker effect on participants who maintain a shorter distance from words overall. This observation is consistent with a well-established base-rate effect, whereby a larger magnitude of a response in one condition (i.e., a longer latency, larger amplitude, longer duration) tends to come with a stronger effect (e.g., a larger amount of change) associated with a critical predictor (see Butler & Hains, 1979; Faust, Balota, Spieler, & Ferraro, 1999).

In conclusion, we observed a strong relation between the word's valence rating and the manikin's distance from the word, suggesting that the slider task is a valid well-calibrated tool for measuring valence. This relation was particularly noteworthy given that our participants responded behaviorally to stimuli that other individuals (those tested in Warriner et al., 2013) evaluated as happy or unhappy. Thus, the observed pattern both validates the slider methodology and supports the generalizability of Warriner et al.'s (2013) ratings to a different population and a different task.

Experiment 2

We observed in Experiment 1 that the position of the manikin relative to a word reflects overall emotionality of that word. Our next step was to test the sensitivity of the scale to subtler individual differences in personality traits, which may influence the participants' biases towards (un)pleasant or (non)arousing phenomena. We also added variability to the procedure of Experiment 1 by presenting words at both the top and the bottom of the slider task. Also, we furthered validation of the slider tool by comparing within-participant responses to affective stimuli produced using the slider task and using the ANEW-like rating scale. In view of the sex imbalance in the available participant pool, we restricted participation solely to females in this experiment (but see Experiment 4).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Method

Participants—Thirty-nine female McMaster University students participated in this experiment in exchange for partial course credit. None of them took part in any other experiment. The data from 4 participants were removed for not making a deliberate response on more than 25% of trials. The data from an additional 4 participants were removed for not being native English speakers. Thirty-one participants remained (age range: 18 to 21, M = 19.03, SD = 1.02).

Affective Stimuli—Stimuli were the same as in Experiment 1 (i.e., a balanced representation of the entire affective space without regard for sex differences in valence or arousal ratings).

Procedure and Measures—The procedure for this study including instructions was the same as for Experiment 1 except that all participants in Experiment 2, in addition to the regular slider task, rated all presented words for both valence and arousal via a web-based form; task order was counterbalanced across participants. Task order did not affect the ratings or the slider responses (model not shown) and is not discussed further.

Additionally, all participants completed four personality questionnaires: 1) the Behavioral Approach/Behavioral Inhibition Scale (BAS/BIS; Carver & White, 1994), which identifies the degree to which people focus on avoiding punishment, fear, sadness, etc. versus focusing on acquiring rewards and achieving goals and is measured on a 4 point Likert scale from “very true for me” to “very false for me”; 2) the Alexithymia Scale (TAS-20; Bagby, Parker, & Taylor, 1994), which measures how strongly people weight sensorimotor information over emotional information (high score) and vice versa (low score) and is measured on a 5 point Likert scale from “strongly disagree” to “strongly agree”; 3) the Affective Style Questionnaire (ASQ; Hofmann & Kashdan, 2010), which measures people’s tendencies to use three different strategies to handle emotional reactions; and 4) the Cheek and Buss Shyness and Sociability Scale (SSS; Cheek, 1983; Cheek & Buss, 1981), which included the five highest loading shyness items (Bruch, Gorsky, Collins, & Berger, 1989) from the original Cheek and Buss (1981) shyness measure and the 5 item sociability scale from the Cheek and Buss (1981) measure. The ASQ and SSS were measured on a 5 point Likert scale from “not true of me at all” to “extremely true of me” and indexed individuals’ personality styles towards social approach and social avoidance tendencies. We note that the use of Likert scales in these questionnaires for measuring personality traits may be subject to shortcomings similar to the ones examined in this paper. We relegate this question to future research and use these tools as suggested by the literature. Altogether the experiment took approximately 1 hour.

The order of the words was randomized for each participant. The words were counterbalanced to appear at the top or the bottom of the slider: all participants saw the word in the same position. We acknowledge in retrospect that this design decision was suboptimal and point to examination of the word position effect as a topic for further research. Yet, given that word position is not systematically correlated with word valence or arousal, or

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with participant sex or personality, we do not expect this to introduce a systematic bias to our data.

Twenty-nine of the 31 remaining participants completed all four scales—two either missed or chose not to complete one or more of the scales. All subsequent analyses use data from only these 29 participants. Mean scores and ranges for these questionnaires are reported in Table 1. All participants completed the slider portion of the experiment on a monitor with a 1600 × 900 pixel resolution.

Data Analyses—We recorded the same information as in the previous experiment. We removed 4 participants who did not make a deliberate response on more than 25% of trials (see 'Participants'). We removed any trials that were more than 2.5 SD from the mean as calculated by participant. Doing so removed 2.4% of the data. Then we trimmed the data set as a whole by removing 1% of trials from both ends of the first click RT distribution. We compared this trimmed set with the original full dataset and found no differences in the pattern of results. The remaining data pool contained 7,072 trials.

Results and Discussion

We calculated by-word averages of distance from the word and participants' own valence ratings. These values demonstrated a near perfect negative correlation ($r = -.973$, 95% CI [−0.978, −0.966], $p < .001$), lending strong support to the notion that the slider tool is an accurate gradient analogue of valence ratings. The two sets of mean affective ratings (those in the Warriner et al.'s (2013) norming study, and those collected during the experiment) showed a strong correlation (Experiment 2: Valence $r = .817$, 95% CI [0.77,0.85], $p < .001$; Arousal $r = .514$, 95% CI [0.42,0.59], $p < .001$) and produced a nearly identical pattern of effects on slider distance. The correlations were comparable in magnitude to the inter-group correlations (i.e., old vs. young, male vs. female, high vs. low education) reported in Warriner at al. (2013) in which Pearson's correlation coefficients for valence ranged from .79 to .83 and for arousal from .47 to .52. For comparability with prior studies and across present experiments we use Pearson correlations. Also, for replicability in future studies, we only report regression analyses made with the mean ratings from Warriner et al. as independent variables.

The linear mixed-effects regression model fitted to the manikin's distance from the word replicated effects observed in Experiment 1 (see Tables S3 and S4 in Supplementary materials). Participants moved the manikin closer to relatively positive and more frequent words and increased the distance to more negative or rare words [valence: $b = -97.0$, $SE = 4.9$, $t = -2.4$, $p = .025$, frequency: $b = -18.7$, $SE = 2.4$, $t = -1.7$, $p = .083$]: the effect of frequency was only marginally significant. Similarly, the base-rate effect was replicated in the random effect structure of the model (see the negative correlation between individual intercepts and slopes in Table S4 in Supplementary materials). Participants who tended to maintain a larger rather than a shorter distance from the target word showed a larger amplitude in avoidant or approaching behavior as a function of valence or frequency. As well, there was a weak and insignificant effect of the position of the word; participants were more prone towards approaching words that were positioned at the top rather than the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

bottom of the slider scale (see Table S3 in Supplementary materials). There was no significant main effect of arousal nor any interactions between arousal and other variables.

Personality measures—The battery of personality measures that we collected revealed main effects of sociability and shyness[3] on the manikin's distance from the word. As predicted, on average, participants with higher sociability scores tended to move the manikin closer to all words [$b = -4.0$, $SE = 1.5$, $t = -2.7$, $p = .01$], while those with higher shyness scores placed the manikin at a larger distance [$b = 2.8$, SE = 1.2, $t = 2.4$, $p = .02$] (Figure 3). Thus, in an average response, a person with the highest shyness score (20) would place the anthropomorphic manikin some 56 pixels, or 10% of the available scale, farther away from the word than the person with the lowest shyness score (1). The effect range was similar for sociability. As expected, shyness and sociability showed a moderate negative correlation ($r = -0.49$, 95% CI [−0.51, −0.47], $p < .001$). To establish whether the difference in effect magnitude was significant between shyness and sociability, we employed the method advocated by Hotelling (1940)[4]. We reverse-coded sociability, such that its effect shows the same polarity as that of shyness. We further obtained inferential estimates from a regression model that contained a variable representing the sum of shyness and reverse sociability and a variable representing their difference, as well as several control variables. As expected, the effect of the sum variable was strong and significant, but importantly the effect of the difference was not (b = 0.2, SE = 1.2, p = 0.891). Thus, there was no reliable difference in the magnitude of shyness and sociability effects, and we cannot establish on the basis of this sample whether effects of shyness and sociability are separable and independent.

Neither shyness nor sociability showed interactions with valence or arousal. This finding suggests that the effect of these personality traits in the tested cohort shows in the overall withdrawal from or proximity towards words. That is, people varying in shyness and sociability experience differences in valence and arousal similarly, and their approach–avoidance behavior is affected in much the same way. Other tests of individual differences (BAS, ASQ, alexithymia) did not affect the participants' performance in a consistent way. We conclude that the slider task is sensitive to individual differences across selected personality traits, with avoidant behavior more prevalent in individuals who also self-report a preference to socially avoidant behavior.

Scale resolution—This Experiment enables a direct comparison of the resolution that is offered by the continuous slider scale and the ordinal 9-point rating scale used to collect participants' ratings of valence. As reported above, there were reliable effects of shyness and sociability on the manikin's distance from the word, amounting to a difference in about 10% of the scale between the least and most sociable person. Critically, no reliable effect of either personality traits was detected in regression models fitted to participants' own ratings (all p's > 0.05; model not shown), despite the near perfect correlation that these ratings have with slider distances. We attribute this discrepancy to the inability of the 9-point rating scale to capture subtle personality effects. Consider the following simplified example as an

[3]Effects of BAS, Alexithymia, and ASQ were primarily non-significant or small and inconsistent across experiments. We chose not to report them here in the interest of space but models can be provided by the authors upon request.
[4]We thank an anonymous reviewer for this suggestion.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

illustration. Suppose that a range of participants encounter a word that is perfectly neutral and would elicit a rating of 5.0 in a person who is in the middle of the sociability scale. A difference of 10% in the evaluation of valence between extremes of the sociability scale would translate into a difference of 5% above and below that average person, yielding (for a 9-point scale) a rating of 4.55 for the least sociable persons and 5.45 for the most sociable ones. Both extreme ratings would, however, round to a rating of 5, making the effect of sociability undetectable. The fine-grained 1-pixel resolution of the slider scale does not meet with this problem, and uncovers even subtle changes in valence.

Recalibration—This experiment offers two datasets to examine the proposed undesired effect of recalibration on valence estimation, i.e., ratings of valence on the ratings scale and responses to the slider task. Instructions to the former were virtually identical to the ANEW-style instructions (see the Introduction and S1), while instructions to the latter used better anchors (see Experiment 1). A straightforward prediction for Westbury et al.'s (2015) criticism of the ANEW-style instructions is that in the beginning of an experiment, words that are only mildly pleasant or unpleasant would be evaluated as very positive or very negative and given extreme ratings. In the course of an experiment, when a participant encounters words denoting really terrible or really wonderful phenomena, these phenomena would be given extreme ratings while words with milder emotionality would move towards the middle of the rating scale. Thus, if recalibration is present, we should expect a reduction in the range of ratings for relatively mild words in the course of an experiment, whereas words with extreme emotional connotations would elicit equally extreme responses across the entire experiment. This prediction enables us to verify whether recalibration affects valence ratings made with ANEW instructions and on the ratings scale, a corrected type of instructions and the slider scale, neither or both.

We tested whether this set of predictions were borne out by fitting a regression model to distance of the manikin from the word, and separately to participants' valence ratings, as a dependent variable, and a critical interaction of a valence estimate by the trial number in the experiment. Importantly, since all ratings in Warriner et al. and similar datasets are suspect to recalibration issues, we used computational estimates of word valence by Westbury et al. (2015), which are independent of human intuition and thus are immune to potential scale issues. The correlation between Warriner et al.'s and Westbury et al.'s estimates of valence for 226 words in our stimulus list that had ratings available in both datasets was $r = 0.67$, $p < 0.001$.

To allow for nonlinearity in the change of valence effect over the course of the experiment, we made use of generalized additive mixed effects models (GAMM; see, e.g., Hastie & Tibshirani, 1990; Wood, 2006) as implemented in the mgcv package 1.8–7 (Wood, 2006, 2011) of the R statistical computing software (R Core Team, 2015). For detailed description and worked examples of the use of generalized mixed-effects additive models in psycholinguistics see Baayen, Kuperman, and Bertram (2010), Balling and Baayen (2012), Kryuchkova et al. (2012), Matuschek, Kliegl, and Holschneider (2015); Smith and Levy (2008) and Tremblay and Baayen (2010). These types of models implement the use of tensor product smooths, enabling the multidimensional modeling of nonlinear, 'wiggly' surfaces created by interactions of numeric variables. Our regression model had distance as a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

dependent variable, a tensor product of trial number and Westbury et al.'s estimate of word valence as a critical predictor, as well as random intercepts per word and random smooths of trial number per participant (full model specification is available in Supplementary materials S3, Table S9). Figure 2 (top left) reports model-estimated effects of trial number on ANEW-type ratings for quartiles of valence, while Figure 2 (top right) reports these effects for distance from the word in the slider task. The latter interaction was virtually identical to the interaction observed in Experiment 1's data (not shown in Figure 2).

The observed pattern shows no evidence of recalibration in the direction predicted by Westbury et al. (2015) for either conventional ANEW-like ratings or for the slider task. Instead of being stable throughout the experiment, extremely valenced words occupy a narrower range of distance towards the end of the experiment relative to the beginning, probably due to participants' habituation to the task and fatigue. This reduction is akin to the tendency of participant to gravitate towards neutral ratings on a rating scale towards the end of the experiment (Warriner et al., 2013). As the order of words was randomized for each participation, no systematic effect of trial order on distance is expected to emerge in our experiment: our regression models confirm this null effect (not shown). In sum, contrary to the prediction of undesired task demand, mildly valenced words are unaffected by the progression of the experiment. We conclude, contra Westbury et al.'s (2015) hypothesis, that neither the ANEW-like rating task nor the novel slider task were affected by recalibration.

Experiment 3

Experiments 1–2 above were conducted in the laboratory with undergraduate students recruited from the convenience pool as participants. This setup imposed typical restrictions on the age, self-selection, educational level and, in our case, sex composition of the tested cohorts. To test the validity of our results in a more diverse and representative population, we implemented a web-based version of the slider task and recruited participants online.

Method

Participants—We used Amazon Mechanical Turk (www.mturk.amazon.com), a web-based service where a pool of anonymous web surfers can earn money by completing tasks supplied by researchers, for data collection. Basic demographics, statistics and best practices of use of the Amazon Mechanical Turk have been recently reviewed in Mason and Suri (2012). We restricted our recruitment to those individuals whose IP address was located in the USA. Participants were paid $2 for their participation, even if they chose to withdraw from the study. None of the participants were involved in any other experiment in this study.

Affective Stimuli—Stimuli were the same as in Experiment 1 and 2 (i.e., a balanced representation of the entire affective space).

Procedure and Measures

The slider task was implemented as a web application using jQuery and PHP. Participants executed the task on their own computers and were required to do so in Full screen mode to maximize screen real estate and eliminate the need for scrolling. We collected information

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

about the operating system and the browser they used: these did not have an effect on any variable of interest. Participants could operate either a mouse or a trackpad (touchpad) to select the position of the manikin and submit their responses.

The appearance of the scale, the manikin and the 'Continue' button were similar to the ones used in Experiment 2, where words appeared on either the top of the bottom of the slider (see Figure 1). An additional URL was shown in the bottom right corner of the screen, enabling a participant to withdraw from the experiment. The length of the vertical slider occupied a range of 648 pixels, from 1 (the closest to the word) to 649 pixels (farthest from the word). The manikin was initially centered at the 325th pixel. For comparability across studies, we linearly transformed the current scale to a range from 100 to 700 pixels, as in Experiments 1–2 and 5 (we divided the scale by 649, multiplied by 600 and added 100). The remainder of the procedure was the same as in Experiments 1–2 including instructions.

Before completing the slider task, participants were asked to provide information on a number of demographic variables: age, sex, handedness, education level, native language(s), and the US state where they currently reside. Additionally, they completed the Cheek and Buss Shyness and Sociability Scale questionnaire (see Experiment 2). We did not collect participants' own ratings of valence or arousal. The entire experiment took between 15 and 35 minutes.

Results and Discussion

Forty-three participants took part in the experiment. The data from eight participants were removed for not making a response on more than 25% of trials. The data from three participants were removed for not being native English speakers. The data was trimmed similarly to Experiment 2. All critical patterns (including the effect of valence and recalibration) were found to be identical in both the untrimmed and trimmed datasets, so only the trimmed data are reported below. The remaining pool consisted of 7897 trials and 32 (13 women, 19 men) participants who ranged in age from 19 to 55 years (M = 34.06, SD = 8.68). The mode education level was a completed bachelor's degree (10 participants). Shyness scores ranged from 0 to 19 (M = 9.40, SD = 5.20) and sociability scores from 0 to 20 (M = 9.28, SD = 5.26).

The web-based version of the slider task replicated all critical patterns observed in Experiments 1–2, even though it was administered to a more diverse cohort. Distance to the word showed a strong negative correlation with the word's valence: r = −0.58, 95% CI [−0.59, −0.56], p < 0.001. The correlation was even stronger when the distance to the word was aggregated across participants: r = −0.88, 95% CI [−0.91, −0.85], p < 0.001. The tendency to move the manikin closer to positive words and farther away from negative ones was confirmed by the linear mixed-effects regression model (Tables S5 and S6 in Supplementary materials) [b = −84.5, SE = 6.2, t = −13.1, p < .001]. As in Experiments 1–2, each point on the 1–9 valence scale corresponded to about 84 pixels, or 14% of the available distance range. Participants tended to withdraw from relatively arousing items [b = 14.3, SE = 3.5, t = 4.1, p < .001]: the effect was stronger than in previous experiments with 14 pixels or 2.3% of the available scale corresponding to one unit of change in word arousal. In line with other Experiments, participants moved the manikin closer to frequent words [b = −7.8,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

SE = 2.2, t = −3.5, p < 0.01], see Table S5 in Supplementary materials. Importantly, on average, shy people kept a greater distance from all words [$b = 2.7$, $SE = 1.1$, $t = 2.5$, $p = .018$]. The effect size was virtually identical to that in Experiment 2, with the shiest person placing the manikin some 52 pixels farther away from an average word than the least shy person. Sociability did not show a reliable main effect, and the two-way interactions for combinations between valence, arousal, shyness and sociability did not reach significance at the 5% level either. Finally, as the random effects structure suggests (see Table S6 in Supplementary materials), people who were overall further away from words showed stronger effects of valence and frequency.

In sum, this experiment replicated both the overall strong relationship between the slider task and affective norms, and the modulating role of personality traits that we found under laboratory conditions with a homogenous population of undergraduate students in Experiments 1–2. Recalibration was not an issue either, as shown in Figure 2 (bottom left). This suggests that the patterns are stable across the demographic diversity in the tested population, the minor differences in the appearance of the slider, as well as the method of administering the task (lab-based vs web-based).

Experiment 4

Experiments 1–3 used the set of instructions that required “to assess how close you would like to be to the word” by moving the manikin along the slider (see Method in Experiment 1). One might argue that this set of instructions may have measured an intended emotional response rather than an actual one: “I would like to treat *diet* as a positive thing, so I will move my manikin close to the word, whereas in fact I don’t like dieting”. Moreover, a low-valence high-arousal word “disaster” was used as an example of the word to withdraw from and a high-valence high-arousal word “triumph” as an example of the word to be approached. This re-opens a possibility of inadequate anchoring in the case that participants vary as to how extremely positively or negatively they feel about these words. This present experiment removes all references to affect from the instructions, including anchor words, and tests whether the influence of valence or arousal is still detectable in the manikin’s distance to the word.

Method

Participants were recruited using the Amazon Mechanical Turk platform, as in Experiment 3. Affective stimuli were the same as in Experiments 1–3. Procedure was the same as in Experiment 3 with the exception of instructions, which read:

> [,,,] On each of the following screens, you will first see a plus sign in the centre. That's to center the mouse for the next screen. Click on the plus and you will see a word either at the top or the bottom with a vertical line below or above it. There will be a person in the centre of that line. The person represents you. You can move "yourself" closer to or further away from the word. Position yourself where you prefer to be.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results and Discussion

Forty-four participants took part in the experiment. None of the participants were involved in any other experiment in this study. Data from three were lost due to not responding in over 25% of trial and from another five participants for not being native English speakers. The data was trimmed similarly to Experiment 2. Again, critical effects were found in both untrimmed and trimmed datasets, so only the trimmed data are reported below. The resulting pool consisted of 8858 trials and 36 participants (13 women, 23 men) who ranged in age from 21 to 60 ($M = 34.42$, $SD = 8.95$). The mode education level was a completed bachelor's degree (12 participants). Shyness scores ranged from 0 to 20 ($M = 10.81$, $SD = 5.61$), and sociability scores from 1 to 18 ($M = 10.25$, $SD = 4.14$).

As in all prior experiments, distance to the word was strongly negatively correlated with the word's valence: $r = -0.49$, 95% CI [−0.51, −0.48], $p < 0.001$. When the distance to the word was aggregated across participants, the observed correlation was virtually identical in magnitude to those obtained in Experiments 1, 2 and 3: $r = -0.87$, 95% CI [−0.89, −0.84], $p < 0.001$. The linear mixed-effects regression model fitted to distance to word replicated the observation that participants approached more positive words [$b = -75.5$, $SE = 6.2$, $t = -12.2$, $p < 0.001$; 75 pixels per unit of valence or 12.5% of the range] and avoided more arousing words [$b = 16.6$, $SE = 3.6$, $t = 4.6$, $p < 0.001$; 16 pixels per unit of arousal or 2.3% of the range]. Shyness showed a weak effect on distance in the expected direction [$b = 1.6$, $SE = 1.2$, $t = 1.3$, $p = 0.20$], which did not reach statistical significance at the nominal threshold. Sociability did not show a reliable main effect either, nor did the two-way interactions formed by combinations between valence, arousal, shyness and sociability. Also, participants moved the manikin closer to more frequent words [$b = -10.5$, $SE = 2.7$, $t = -3.9$, $p < 0.001$] and were more prone towards approaching words that were positioned at the top rather than the bottom of the slider scale [$b = -13.3$, $SE = 6.6$, $t = -2.0$, $p = .044$] (see Table S7 in Supplementary materials). As Table S8 (in Supplementary materials) shows, people who were overall further away from words showed stronger effects of valence and frequency. Finally, Figure 2 (bottom right) confirms that a hypothesized recalibration effect was not an issue for this experiment either, despite the removal of anchor terms and re-wording of instructions. Distance to mildly valenced words did not noticeably change over the course of the experiment, while extremely negative words came with a longer distance from the word in the beginning compared to the end of the experiment.

Magnitude of the valence effect—The effect of valence was somewhat weaker in Experiment 4 as compared to other experiments using the same stimuli [estimated regression slopes of the valence effect for Experiment 1: $b = -92.3$; Experiment 2: $b = -97.0$; Experiment 3: $b = -84.5$; Experiment 4: $b = -75.5$]. This raises the possibility that the removal of the task demand in this Experiment's instructions lessened the influence of affect. To address this concern, we fitted a linear multiple regression model to the average distance to each word calculated separately for each study. Figure 4 presents the regression lines for distance as a function of valence for Experiments 1–4, i.e. all experiments that used a balanced word set representing the entire range of valence and arousal. The interaction between valence and experiment was a critical predictor in the model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2 reports the outcome of the regression model that chose contrast coding for the factor with experiment labels as levels, and level "Experiment 4" as a reference level. The model demonstrates that the regression slope in Experiment 4 (a) did not significantly differ from that in Experiment 3 [$b = -4.5$; $SE = 4.6$; $t = -1.0$; $p = 0.330$], but (b) significantly differed from those in Experiments 1 [$b = -16.5$; $SE = 4.6$; $t = -3.6$; $p < 0.001$] and 2 [$b = -13.7$; $SE = 4.6$; $t = -3.0$; $p = 0.003$]. Changing the reference level, we also verified that laboratory-based studies (Experiments 1 and 2) did not produce significantly different slopes from one another [$p > 0.05$], but both web-studies (Experiments 3 and 4) had different slopes from those found in both lab-studies [$p < 0.01$]. This pattern reveals that the method of administering the task and the minor differences in the scale between the web and the laboratory versions led to a small discrepancy in the effect size of valence (on the order of 15 pixels per unit of valence). This observation dovetails with reported differences in rating tasks conducted in the lab or online (Barenboym, Wurm, & Cano, 2010; Wurm & Cano, 2010; Wurm, Cano, & Barenboym 2011). Crucially, however, this discrepancy was not due to the presence or absence of the task demand, as it was equally found in Experiments 3 and 4. The convergence of the effects that valence showed in Experiments 1–4 rules out the task demand as a potential cause for the observed link between affect and approach-avoidance behavior.

Experiment 5

One of the questions we posed was whether the slider task was sufficiently sensitive to group differences in valence ratings. Given the sex differences in emotional responses to words as reported by Warriner et al. (2013), we set out to test whether those differences in affective norms would be mirrored in the behavior of men versus women in the slider task.

Method

Participants—Eighty-seven students at McMaster University participated in this experiment in exchange for partial course credit. None of them participated in any other study. Using the same exclusion criteria as in Experiment 1, the data from fifteen participants were removed for not making a deliberate response on more than 25% of trials (see Experiment 1). Data from 8 participants were removed for not being native English speakers. The remaining 64 (35 female, 29 male) participants who participated in the study ranged in age from 17 to 23 years ($M = 18.97$ years, $SD = 1.39$).

Affective Stimuli—Warriner et al.'s (2013) dataset reported ratings of valence and arousal averaged by sex. For each word, we calculated the difference between average male and female ratings for both valence and arousal. We selected 50 words with the most extreme positive, and 50 words with the most extreme negative difference scores for valence, i.e. 50 words associated with higher valence ratings from men than women (*beer, gun, topless, hotshot*), and 50 vice versa (*flower, caterer, faith, parent*). We also selected 30 words associated with higher arousal ratings from men than women (*panties, hunting, scuffle, velvet*), and 30 words vice versa (*nerd, limo, skinny, toddler*). Words eliciting different valence or arousal ratings across sexes totaled 160 stimuli. To make sure that not all stimuli represent extreme sex differences in valence and arousal, we prepared filler words that had

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

little to no sex differences in ratings. For this purpose, the remaining dataset was divided into 25 bins (crossing quintiles of valence and arousal), and 5 words were randomly chosen from each bin for a total of 125, making a total of 285 words. One word was subsequently lost due to a program error. In the final stimulus set, valence and arousal were uncorrelated ($r_s = -0.101$, $p > .05$) and difference scores for both were approximately normally distributed, as indicated by the Shapiro-Wilk normalcy test. Mean affective ratings for the entire stimulus list and for each subset of words are reported in Table 3.

Procedure—The procedure for this study including instructions was the same as for Experiment 1 with the following differences. First, the initial 49 participants completed the task on a monitor with a 1024 × 768 pixel resolution while the last 38 completed it on a monitor with a 1600 × 900 pixel resolution. There was no difference in responses based on screen resolution, $t(607) = -0.18$, 95% CI [−17.37, 14.43], Cohen's $d = 0.016$. Second, the participants in this Experiment (as in Experiment 2) rated all words for both valence and arousal via a web-based form; task order counterbalanced across participants. As in Experiment 2, mean ratings from Warriner et al.'s (2013) norming study, and those collected during the experiment correlated strongly and influenced the slider distance in nearly identical ways. To enable replicability, we only report the analyses made with the mean ratings from Warriner et al. as independent variables.

Variables—Dependent variables included the average distance from each word as chosen by male participants and the average distance as chosen by female participants. The difference ranged from −249 pixels (men closer to the word than women) to 258 (women closer to the words than men).

Independent variables were sex-specific mean ratings of valence and arousal from Warriner et al. (2013). Additionally, we considered the difference in valence and arousal ratings per word as a predictor: positive when a rating given by men was higher (showing a happier, or more excited response) than that given by female raters.

Results and Discussion

We trimmed the data in a similar manner to Experiment 1. Participants who did not move the anthropomorphic manikin more than 25% of the time were removed (see 'Participants'). One additional female participant was removed for having an average first click RT greater than 2.5 SD's than the mean of rest of the participants. We removed any trials that were more than 2.5 SD from the mean as calculated by participant. Doing so removed 2.6% of the data. We then trimmed the data set as a whole by removing 1% of trials from both ends of the first click RT distribution. The resulting dataset contained 17,068 trials, with 34 male and 28 female participants. Results patterned the same for both trimmed and untrimmed datasets, and thus we report analyses of the trimmed dataset only.

First, we tested whether behavioral patterns of elicited by words with sex-different responses were similar to the patterns elicited by words representing the entire range of valence and arousal (as observed in Experiment 1). A linear mixed-effects multiple regression model was fitted to the manikin's distance from the word as a dependent variable, with average (non-sex specific) ratings (from Warriner et al, 2013), frequency (from SUBTLEX-US) and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

position as critical predictors (model not shown). Both word and participant were entered as random factors. This model replicated the findings of Experiment 1—higher valence ratings were related to the tendency to move the manikin closer to the word (approximately 77 pixels closer for each unit increase in valence) [$b = -76.8$, $SE = 4.3$, $t = -17.8$, $p < .001$]. There was no effect of arousal, but there was a marginal main effect of frequency—higher frequency words were approached more than lower frequency words [$b = -3.7$, $SE = 2.0$, $t = -1.9$, $p = .06$]. Also, as the random-effects structure indicated, participants showed stronger effects of valence on distance (i.e., changed the manikin position by more pixels in response to the same change in valence) when their distance from the word was overall larger. The distance from the word was slightly larger (by 14 pixels) if the word was positioned on the top as opposed to the bottom of the slider [$b = 14.2$, $SE = 5.9$, $t = 2.4$, $p = .02$]. When sex was added as a predictor, there was no main effect; however, there was an interaction with both valence and arousal, which will be explored subsequently.

The second analysis, which addressed the central point of this Experiment, tested the relation between sex differences in valence and sex differences in manikin distance. For this analysis, we calculated the average distance from each word separately for men and women. These distances were pitted against the rating differences of valence and arousal (Warriner et al., 2013). In other words, sex-specific distances were correlated with valence ratings given by the same sex. Figure 5 and Table 4 summarize the outcome of the linear multiple regression model (a mixed-effects model was not used as only one value of the dependent variable was associated with each word).

Figure 5 points to a tendency for participants of one sex to preferentially approach words rated as more pleasant or arousing by raters of the same sex [valence: $b = -18.1$, $SE = 1.8$, $t = -9.8$, $p < .001$; arousal (not plotted): $b = -6.0$, $SE = 2.0$, $t = -3.0$, $p = .003$]. On average, women moved the manikin about 18 pixels (or 3% of the 600 pixel range) closer to a word whose valence ratings given by female raters was 1 point higher than that given by male raters (e.g. *adoring, drink, manuscript*). Between the extremes of the sex difference in valence ratings (−3.40 *mommy* to 4.48 *threesome*, where positive numbers indicate higher male ratings), the sex difference in locations reached a substantial magnitude of 139 pixels or 23% of the available position range.

There was also a similar, though weaker, tendency to move the manikin closer to the words judged as more arousing by the same sex. A sex difference in 1 point of arousal ratings to a word came with a difference of about 6 pixels (or 1% of the distance range) in the distance of the manikin from the word. Between the extremes of the sex difference in arousal (−3.30 *seafood* to 4.25 *musket*), the magnitude of the distance difference was 44.5 pixels or 7% of the available position range.

The symmetrical nature of the relation in Figure 5 is further confirmed by the value of the intercept: for completely neutral words, i.e. those with no sex difference in either valence or arousal, the sex difference in the distance to the word is minimal: only 13 pixels or 2% of the available range. Thus, as suggested by a weak main effect of sex in the linear mixed effects model above, there is no overall difference between sexes in approach–avoidance behavior in response to emotional stimuli. Both sexes reduced distance to pleasant and arousing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

words, yet—crucially—they did so more when the words were judged as particularly pleasant or arousing by that sex. We conclude that the slider task is sensitive to group, specifically sex, differences in emotional responses to stimuli.

General Discussion

In all five experiments, we showed that when the stimulus was positive, people approached it, and when the stimulus was negative, people withdrew from it. Distance was inversely proportional to the magnitude of valence, and correlations between by-word average distances and valence ratings ranged from strong to near perfect (all $r > 0.8$). Using this method, we were also able to identify both group and individual level variability.

At the group level, in Experiment 5, we showed that participants' distance choices paralleled the affective ratings given by their respective sex in Warriner et al. (2013). Women moved closer to words that female participants in a different study had rated more positively and further from words that women had rated less positively. Men moved closer to words that male participants from a different study had rated more positively and further from words that men had rated less positively. There was no average difference between sexes, meaning that women did not consistently approach words more than men or vice versa. The difference was specific to those words that were rated differently and in magnitude only, not direction.

Additionally, individual differences were observed in all studies. Thus, in all Experiments we found that participants who chose to move farther away from a word on average were more responsive to valence. We further demonstrated that the functional relation between valence and manikin distance was co-determined by personality differences. Participants who scored high on shyness (in Experiments 2 and 3) tended to stay further away from all stimuli while those who scored high in sociability (in Experiment 2 only) tended to move closer to all stimuli. Effects of shyness and sociability did not reach statistical significance in Experiment 4 (see below).

In sum, we argue that the proposed slider scale is a valid tool for an evaluation of psychological valence on a scale that is (i) continuous, (ii) has a high resolution, and is (iii) well-anchored. Considerations (i)–(iii) meet concerns raised against the commonly used rating scale of affective ratings (Bradley & Lang, 1999 and derived scales). First, the interval nature of the slider scale enables the use of parametric tools of descriptive and inferential statistics. Second, the scale's fine-grained resolution uncovers subtle effects of personality traits that would go undetected if a rating scale was used. Our discussion in Experiment 2 and converging results (for shyness) in Experiment 3 exemplify the utility of the higher resolution in that the slider scale reveals effects of shyness and sociability when participants' own ratings on a rating scale fail to. Third, two new sets of instructions for the slider task—presented in full in Experiments 1 and 4—do not give rise to the issue of the scale anchoring advocated by Westbury et al. (2015), and neither do the commonly used ANEW-like instructions. Figure 2 demonstrates that—contrary to Westbury et al.'s concerns—there is no tendency for participants to treat mildly negative or positive words as very negative or very positive, either in the beginning, the middle or at the end of the experiment, regardless of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

instructions, or quality of anchors. We conclude that our new tool resolves some shortcomings of previous tasks requiring affective evaluation (i.e. the ordinal scale and scale resolution), while maintaining a very strong relationship to the magnitude of emotional responses that those tasks elicited. In consideration of space, we present our examination of the slider's reliability elsewhere (Imbault, Shore, & Kuperman, in preparation), but we have observed that the task has a high reliability both within participants, between participants, and over short and long period of time.

In what follows, we consider effects on manikin distance other than valence; discuss the stability of results across experiments; examine theoretical implications of using the slider scale; and outline experimental tasks that can be implemented with the slider tool but not with a rating scale.

Other effects on distance

Our ability to consider multiple affective and lexical factors simultaneously clarified several effects that are understudied in the literature, e.g. those of arousal and word frequency. While our predictions regarding the role of valence were fully confirmed, the role of arousal in approaching or avoiding stimuli was less clear-cut. We did not observe an interaction between valence and arousal in determining distance choices. Perhaps such an interaction will only be evident in speed rather than degree of response. The studies that have shown such an interaction between valence and arousal have been reaction time paradigms showing that congruency of affect facilitated quicker responses (Robinson et al., 2004; Eder & Rothermund, 2010). Comparably, it may be that deciding to pull away from a congruently negative and highly arousing word like "danger" might be quicker than deciding to pull away from something that is incongruently negative and calm like "sadness" while the withdrawal distance might be equivalent. However, our methodology in this paper did not allow us to capture accurate reaction times and this interaction was not evidenced in the measurement of distance. Where observed (Experiments 3 and 4), higher arousal was associated with a tendency to withdraw from the stimulus. This behavior may be related to a long-observed facilitation of approach to moderately intense stimuli and a facilitation of avoidance to highly intense stimuli, reactions evident even in lower animals and infants (Schneirla, 1959 and Izard, 1993, as cited in Robinson et al., 2004).

We also observed a clear effect of word frequency on performance in the slider task, even after valence was controlled for. In all five experiments, participants moved closer to higher frequency words than to lower frequency ones: effects were marginally significant in Experiments 2 and 5 ($p = 0.08$ and 0.06, respectively). More frequent or familiar words are rated more positively, an observation in line with a well-established positive correlation between subjective and objective indices of word familiarity and word valence (see Warriner & Kuperman, 2015; Westbury, 2014, for a detailed discussion).

Stability of results

As already mentioned, the effect of arousal on distance was not stable across studies. It was weak and unreliable in Experiments 1 and 2, but was sizable in Experiments 3 and 4. It is possible that distance is not an appropriate measure with which to capture arousal effects. In

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

addition, estimation of arousal is known to be less stable across studies and samples (Warriner et al., 2013). Similarly, the effect of shyness was reliably present in Experiments 2 and 3, but not in Experiment 4 (sociability was only reliable in Experiment 2). It is possible that discrepancies stemmed from the greater age and educational diversity of the cohorts recruited online, as well as a different sex balance. A second possibility relates to the different environment in which the task was undertaken—a computer laboratory with multiple participants and an experimenter present vs. at home by oneself. Shyness and sociability would presumably be variables more likely to be affected by the social context of a study. We also note that the stimuli used were not chosen for their social relevance which makes the finding of an effect of shyness and sociability perhaps even more interesting. This effect might be stronger and more reliable if socially relevant stimuli were selected. Further explorations of relevant participant-related dimensions are clearly necessary to establish the magnitude and reliability of the weaker effects across the population at large.

Future directions

We couched the interpretation of the slider task as an alternative way of evaluating psychological valence, which offers methodological and statistical benefits. However, the choice to use distance as a corollary to valence stems from a long-standing proposal that valence is unconsciously linked to motivational systems which drive appetitive and aversive responses (Carver & White, 1994; Lang, 1995). The classic finding is of a congruency effect in which people are faster to approach positive stimuli than negative and faster to avoid negative stimuli than positive (pushing or pulling a lever or joystick—Chen & Bargh, 1999, Fishbach & Shah, 2006; Rinck & Becker, 2007; taking steps forward or backwards—Stins, Roelofs, Villan, Kooijman, Hagenaars, & Beek, 2011; or even making facial expressions—Neumann, Hess, Schulz, & Alpers, 2005). The advantage of this method is its ability to capture unconscious, automatic associations; the disadvantage is the way in which it dichotomizes valence and does not consider arousal. The slider task could potentially bridge the gap, providing a way to measure approach and avoidance in a gradient way. In its current form, the slider is not a measure of automatic association, however, small changes such as making the task speeded may extend its utility in this regard.

It is an open question whether the results that we report with the current slider paradigm can be attained if a more traditional interval rating scale were used. It is clear however that our slider paradigm opens possibilities for experimentation unattainable or more cumbersome with the ratings scale paradigms. One possibility is eliciting behavioral responses to emotion in conflict situations, say, when two words are presented at the extremes of the scale and are manipulated to represent a small, medium or large difference in valence, arousal, or both affective dimensions. A closer distance to one of the two words is then expected to indicate how much more positive that word is than its counterpart: too small a difference would elicit a response right in the middle of the slider scale, and too large a difference would cause the participant to move the manikin all the way to the more positive word. A task like this would be able to identify thresholds of emotional sensitivity, i.e., the minimum and the maximum differences in valences that produce differential responses in the manikin distance. It can also be used to assess the role of arousal and its interaction with valence evaluation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Another possibility is to manipulate the manikin image (a male vs female figure; a figure of an old vs young person; or a figure of a neutral vs sad person) and instruct participants to evaluate their emotional response to words as the person depicted as the manikin. This "figure change" paradigm enables researchers to tap into the ability of participants to mimic emotional states experienced by individuals of a different sex, age, mood etc. The amount of change is easily estimated as an adjustment in the intercept and the slope of the regression lines characterizing one's responses in one capacity (e.g. male) vs another capacity (female), see Kuperman, Imbault, Shore (2015). Although one can similarly collect the ANEW-style ratings instructing the participant to rate words as a male or a female, the presence of a sex-specific symbol on the slider, as well as the interval nature of the slider scale, make both the data collection and analysis more reliable.

Finally, the slider task is easy to administer either in the laboratory or online, across populations, ages and levels of ability. Anecdotally, our participants, including school-age children, report that they perceive this task as less boring and monotonous than producing responses on a rating scale. This slider method may then offer a way of using emotionality of words to identify the presence of particular pathologies (e.g., extreme shyness), or conduct research at an earlier age that would be feasible with other tasks.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Margo Wilson and Martin Daly Ontario Graduate Scholarship (OGS) to the first author. The last author's contribution was supported in part by the Social Sciences and Humanities Research Council of Canada (SSHRC) Insight Development grant 430-2012-0488, the Natural Sciences and Engineering Research Council or Canada (NSERC) Discovery grant 402395-2012, the NIH R01 HD 073288 (PI Julie A. Van Dyke), the Early Research Award from the Ontario Ministry of Research and Innovation, and the Canada Research Chair (Tier 2) award. LAS was supported by SSHRC, NSERC, and the Canadian Institutes of Health Research operating grants. DIS was supported by an NSERC Discovery Grant. CLI was supported by the OGS fellowship.

References

Adelman JS, Estes Z. Emotion and memory: A recognition advantage for positive and negative words independent of arousal. Cognition. 2013; 129(3):530–535. [PubMed: 24041838]

Adelman JS, Marquis SJ, Sabatos-DeVito MG, Estes Z. The unexplained nature of reading. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2013; 39(4):1037.

Allen IE, Seaman CA. Likert scales and data analyses. Quality Progress. 2007; 40(7):64.

Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. 2008; 59(4):390–412.

Baayen RH, Kuperman V, Bertram R, Baayen R. Frequency effects in compound processing. Compounding, Amsterdam/Philadelphia: Benjamins. 2010:257–270.

Baayen RH, Milin P. Analyzing reaction times. International Journal of Psychological Research. 2010; 3(2):12–28.

Bagby RM, Parker JD, Taylor GJ. The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. Journal of Psychosomatic Research. 1994; 38(1):23–32. [PubMed: 8126686]

Balling LW, Baayen RH. Probability and surprisal in auditory comprehension of morphologically complex words. Cognition. 2012; 125(1):80–106. [PubMed: 22841290]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language. 2013; 68(3):255–278.

Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-4. 2013 [Accessed online: December 2013]

Bestgen Y, Vincze N. Checking and bootstrapping lexical norms by means of word similarity indexes. Behavior Research Methods. 2012; 44(4):998–1006. [PubMed: 22396137]

Bradley, MM., Lang, PJ. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida; 1999. p. 1-45.

Bruch MA, Gorsky JM, Collins TM, Berger PA. Shyness and sociability reexamined: A multicomponent analysis. Journal of Personality and Social Psychology. 1989; 57(5):904.

Brysbaert M, New B. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods. 2009; 41(4):977–990. [PubMed: 19897807]

Butler B, Hains S. Individual differences in word recognition latency. Memory & Cognition. 1979; 7(2):68–76.

Carver CS, White TL. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. Journal of Personality and Social Psychology. 1994; 67(2):319–333.

Cheek JM, Buss AH. Shyness and sociability. Journal of Personality and Social Psychology. 1981; 41(2):330–339.

Cheek, JM. The revised Cheek and Buss shyness scale. Wellesley, MA: Wellesley College; 1983. Unpublished manuscript

Chen M, Bargh JA. Consequences of automatic evaluation: Immediate behavioural predispositions to approach or avoid the stimulus. Personality and Social Psychology Bulletin. 1999; 25(2):214–224.

Eder AB, Rothermund K. Automatic influence of arousal information on evaluative processing: Valence–arousal interactions in an affective Simon task. Cognition and Emotion. 2010; 24(6): 1053–1061.

Faust M, Balota D, Spieler D, Ferraro F. Individual differences in information-processing rate and amount: Implications for group differences in response latency. Psychological Bulletin. 1999; 125(6):777–799. [PubMed: 10589302]

Fishbach A, Shah JY. Self-control in action: Implicit dispositions toward goals and away from temptations. Journal of Personality and Social Psychology. 2006; 90:820–832. [PubMed: 16737375]

Hastie, TJ., Tibshirani, RJ. Generalized additive models. Vol. 43. CRC Press; 1990.

Heuer K, Rinck M, Becker ES. Avoidance of emotional facial expressions in social anxiety: The approach–avoidance task. Behaviour Research and Therapy. 2007; 45:2990–3001. [PubMed: 17889827]

Hofmann SG, Kashdan TB. The affective style questionnaire: development and psychometric properties. Journal of Psychopathology and Behavioral Assessment. 2010; 32(2):255–263. [PubMed: 20495674]

Hollis G, Westbury C, Lefsrud L. Extrapolating Human Judgments from Skip-gram Vector Representations of Word Meaning. The Quarterly Journal of Experimental Psychology. 2016:1–45.

Hotelling H. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. The Annals of Mathematical Statistics. 1940; 11(3):271–283.

Izard CE. Four systems for emotion activation: cognitive and noncognitive processes. Psychological Review. 1993; 100(1):68–90. [PubMed: 8426882]

Jamieson S. Likert scales: how to (ab) use them. Medical Education. 2004; 38(12):1217–1218. [PubMed: 15566531]

Kloumann IM, Danforth CM, Harris KD, Bliss CA, Dodds PS. Positivity of the English language. PLoS One. 2012; 7:e29484. [PubMed: 22247779]

Knapp TR. Treating ordinal scales as interval scales: an attempt to resolve the controversy. Nursing Research. 1990; 39(2):121–123. [PubMed: 2315066]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Kousta ST, Vinson DP, Vigliocco G. Emotion words, regardless of polarity, have a processing advantage over neutral words. Cognition. 2009; 112(3):473–481. [PubMed: 19591976]

Kryuchkova T, Tucker BV, Wurm LH, Baayen RH. Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography. Brain and Language. 2012; 122(2):81–91. [PubMed: 22726720]

Kuperman V, Estes Z, Brysbaert M, Warriner AB. Emotion and Language: Valence and Arousal Affect Word Recognition. Journal of Experimental Psychology: General. 2014; 143(3):1065–1081. [PubMed: 24490848]

Kuperman, V., Imbault, C., Shore, D. Gender Differences in Emotional Empathy; Presentation at the Annual Meeting of the Psychonomics Society; Chicago, IL. 2015 Nov.

Kuppens P, Tuerlinckx F, Russell JA, Barrett LF. The relation between valence and arousal in subjective experience. Psychological Bulletin. 2013; 139(4):917–940. [PubMed: 23231533]

Lang PJ. The emotion probe: Studies of motivation and attention. American Psychologist. 1995; 50(5): 372–385. [PubMed: 7762889]

Liu B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies. 2012; 5(1):1–167.

Mandera P, Keuleers E, Brysbaert M. How useful are corpus-based methods for extrapolating psycholinguistic variables? The Quarterly Journal of Experimental Psychology. 2015; 68(8):1623–1642. [PubMed: 25695623]

Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods. 2012; 44(1):1–23. [PubMed: 21717266]

Matuschek H, Kliegl R, Holschneider M. Smoothing spline ANOVA decomposition of arbitrary splines: An application to eye movements in reading. PloS one. 2015; 10(3):e0119165. [PubMed: 25816246]

Mohammad, SM., Turney, PD. Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. Association for Computational Linguistics; 2010 Jun. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon; p. 26-34.

Neumann R, Hess M, Schulz SM, Alpers GM. Automatic behavioural responses to valence: Evidence that facial action is facilitated by evaluative processing. Cognition and Emotion. 2005; 19(4):499–513.

Norman G. Likert scales, levels of measurement and the "laws" of statistics. Advances in Health Sciences Education. 2010; 15(5):625–632. [PubMed: 20146096]

Osgood, CE., Suci, GJ., Tannenbaum, PH. The measurement of meaning. University of Illinois Press; 1957.

Pennebaker, JW., Francis, ME., Booth, RJ. Linguistic inquiry and word count: LIWC 2001. Vol. 71. Mahway: Lawrence Erlbaum Associates; 2001. 2001

Pinheiro, JC., Bates, DM. Mixed-effects models in S and S-PLUS. New York: Springer-Verlag; 2000.

Puca RM, Rinkenauer G, Breidenstein C. Individual differences in approach and avoidance movements: How the avoidance motive influences response force. Journal of Personality. 2006; 74(4):979–1014. [PubMed: 16787426]

R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. URL https://www.R-project.org/

Recchia G, Louwerse MM. Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. The Quarterly Journal of Experimental Psychology. 2015; 68(8):1584–1598. [PubMed: 24998307]

Reinecke A, Becker ES, Rinck M. Test–retest reliability and validity of three indirect tasks assessing implicit threat associations and behavioural response tendencies. Zeitschrift für psychologie. 2010; 218:4–11.

Rinck M, Becker ES. Approach and avoidance in fear of spiders. Journal of Behavior Therapy and Experimental Psychiatry. 2007; 38:105–120. [PubMed: 17126289]

Robinson MD, Storbeck J, Meier BP, Kirkeby BS. Watch out! That could be dangerous: Valence-arousal interactions in evaluative processing. Personality and Social Psychology Bulletin. 2004; 30(11):1472–1484. [PubMed: 15448310]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Schmidt, LA., Buss, AH. Understanding shyness: Four questions and four decades of research. In: Rubin, KR., Coplan, RJ., editors. The Development of Shyness and Social Withdrawal. New York: Guildford Publications; 2010. p. 23-41.

Schneirla TC. An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. Nebraska symposium on motivation. 19591959:1–42.

Shaoul C, Westbury C. Exploring lexical co-occurrence space using HiDEx. Behavior Research Methods. 2010; 42(2):393–413. [PubMed: 20479171]

Smith NJ, Levy R. Optimal processing times in reading: a formal model and empirical investigation. Proceedings of the 30th annual conference of the cognitive science society. 2008:595–600.

Stins JF, Roelofs K, Villan J, Kooijman K, Hagenaars MA, Beek PJ. Walk to me when I smile, step back when I'm angry: Emotional faces modulate whole-body approach–avoidance behaviours. Experimental Brain Research. 2011; 212:603–611. [PubMed: 21698468]

Tremblay, A., Baayen, RH. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In: Wood, D., editor. Perspectives on formulaic language: Acquisition and communication. London: The Continuum International Publishing Group; 2010. p. 151-173.

Wang ST, Yu ML, Wang CJ, Huang CC. Bridging the gap between the pros and cons in treating ordinal scales as interval scales from an analysis point of view. Nursing Research. 1999; (4):226–229. [PubMed: 10414686]

Warriner AB, Kuperman V. Affective biases in English are bi-dimensional. Cognition and Emotion. 2015; 29:1147–1167. [PubMed: 25313685]

Warriner AB, Kuperman V, Brysbaert M. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods. 2013; 45(4):1191–1207. [PubMed: 23404613]

Westbury C. You Can't Drink a Word: Lexical and Individual Emotionality Affect Subjective Familiarity Judgments. Journal of Psycholinguistic Research. 2014; 43(5):631–649. [PubMed: 24061785]

Westbury C, Keith J, Briesemeister BB, Hofmann MJ, Jacobs AM. Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. The Quarterly Journal of Experimental Psychology. 2015; 68(8): 1599–1622. [PubMed: 26147614]

Wood, S. Generalized additive models: an introduction with R. CRC press; 2006.

Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011; 73(1):3–36.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

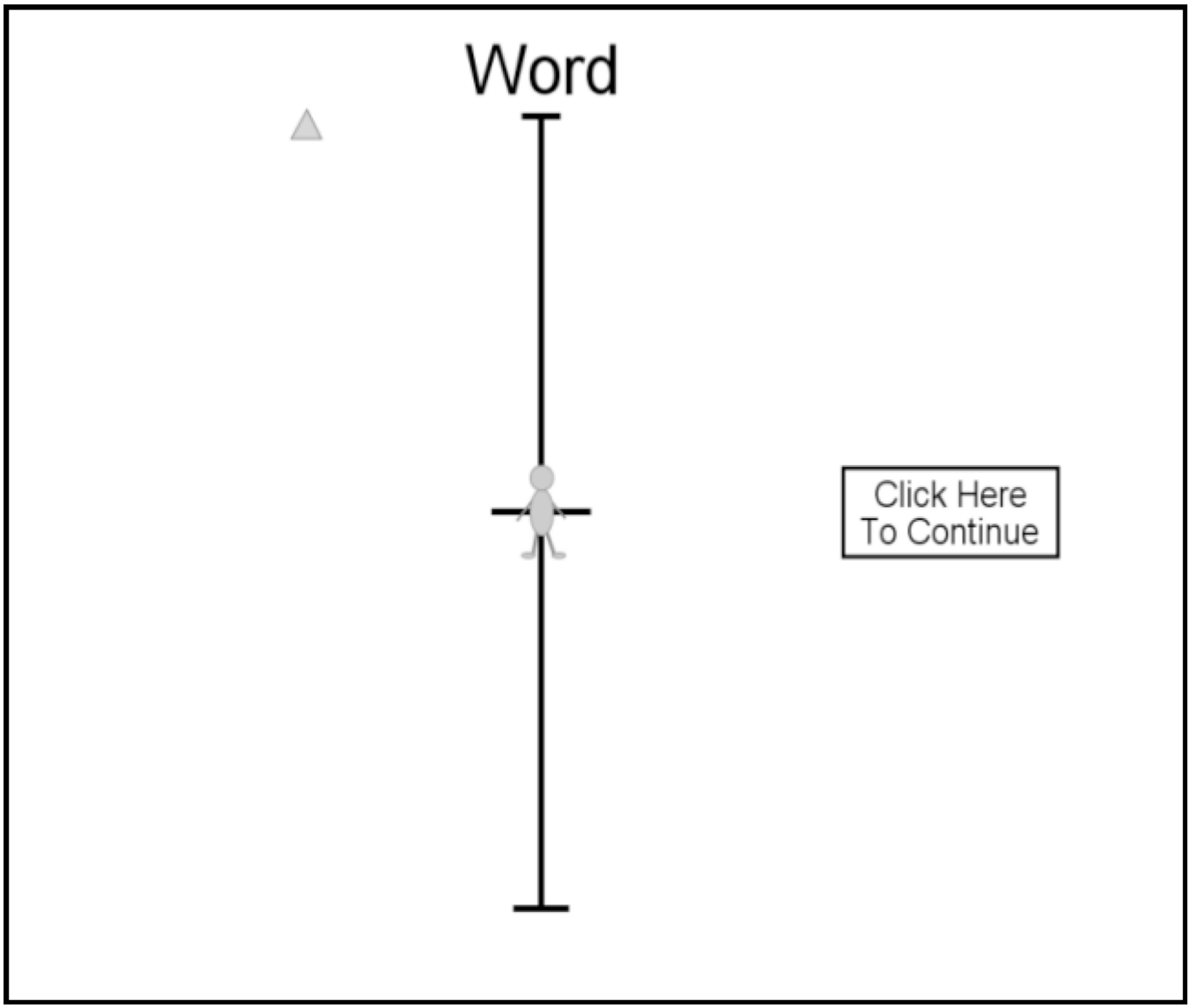

Figure 1.
The slider scale, manikin figure and word (in top position) at the beginning of each trial.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

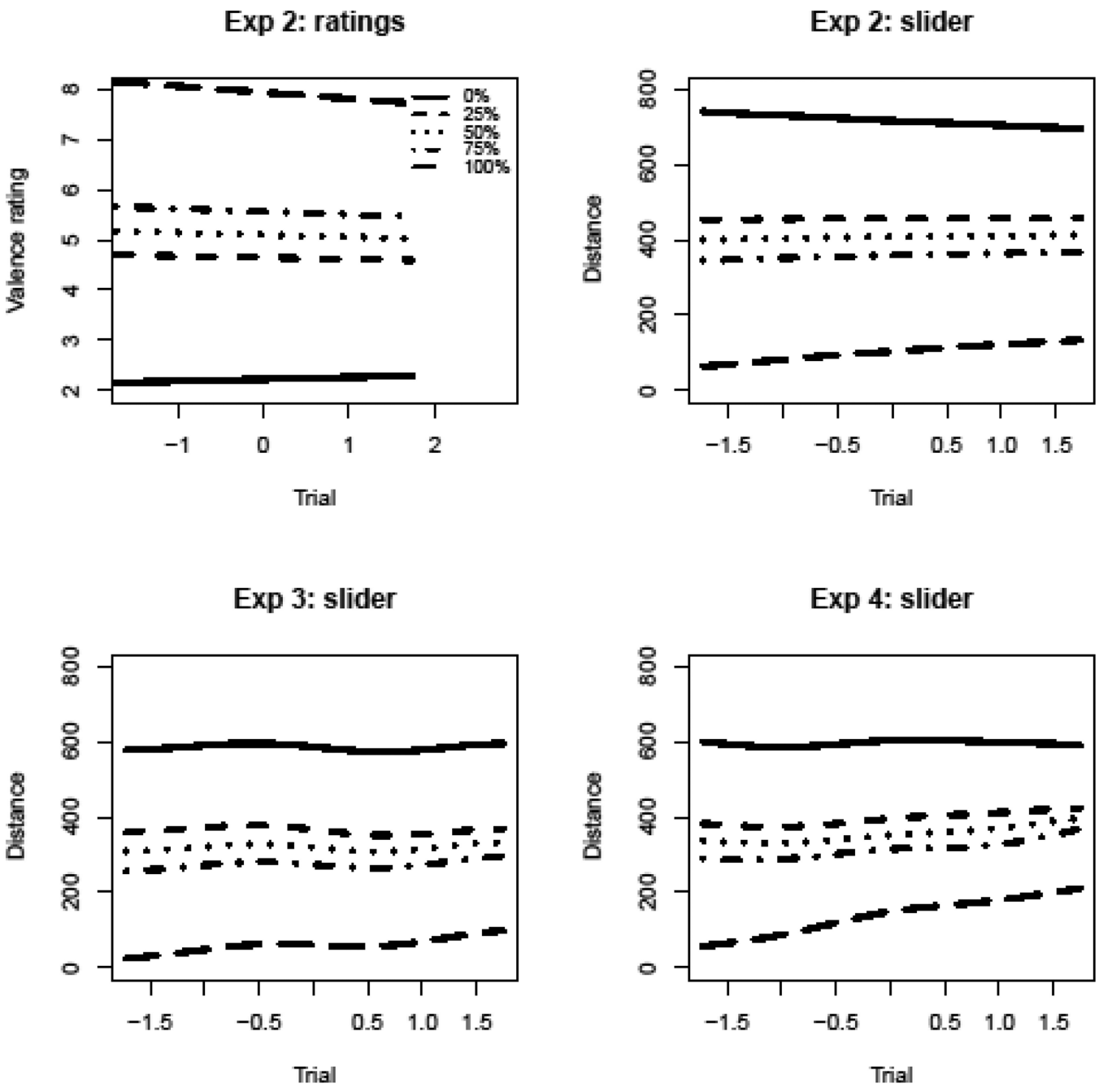

Figure 2.

The partial effect of (scaled) trial number on valence ratings by participants in Experiment 2 (top left) and distance between the manikin and the word (Experiments 2–4), by quartile of Westbury et al.'s (2015) valence estimates (from the most negative words shown as a solid line to the most positive shown as a long-dash line). Distances and ratings change towards the end of the experiment for extremely valenced words; mildly valenced words are unaffected by the progression of the experiment.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

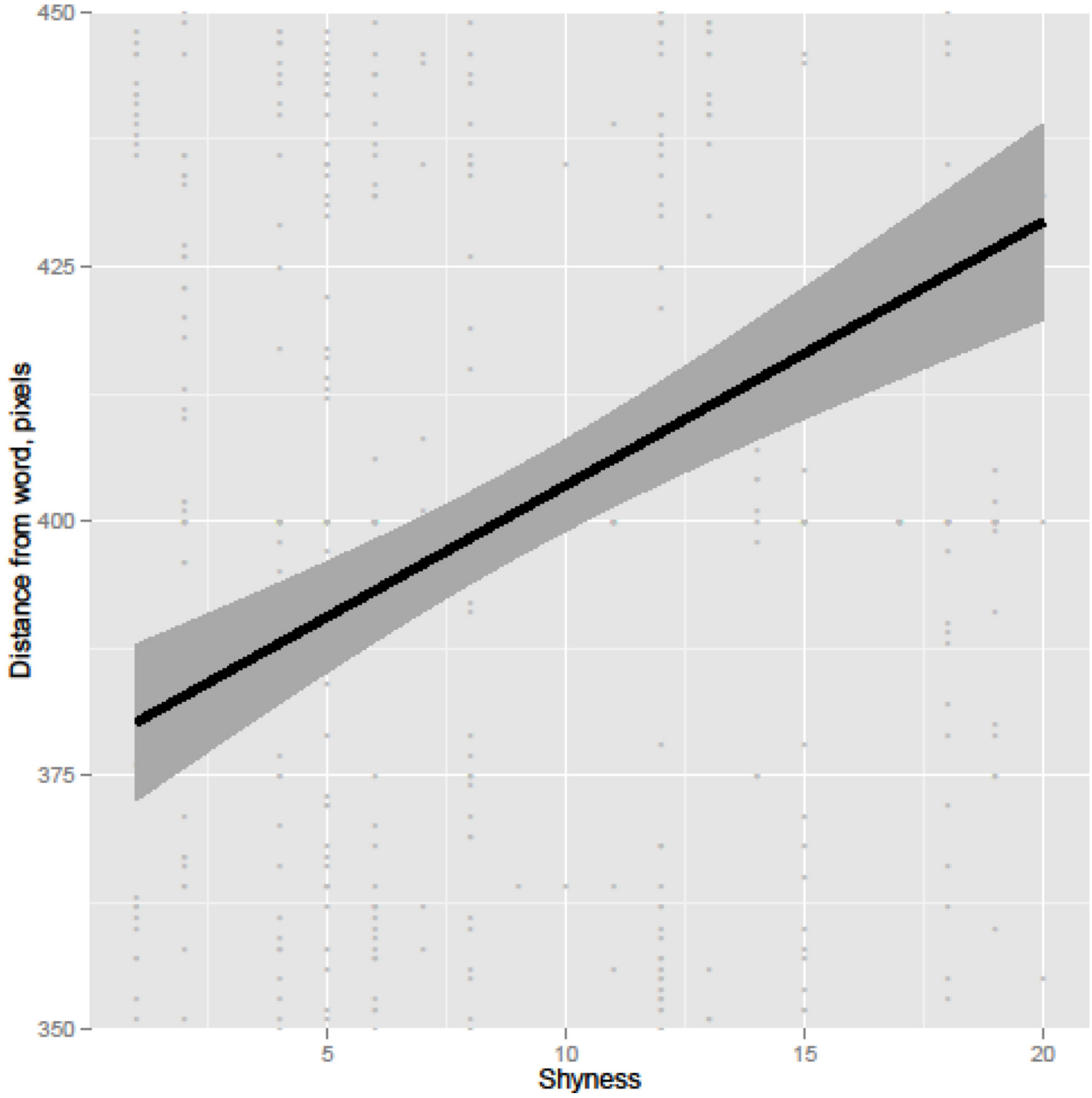

Figure 3.
Scatterplot of the manikin's distance from the word as a function of shyness scores. The individual data points are shown in white and the trend line in black, with the 95% confidence interval presented as a gray area.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

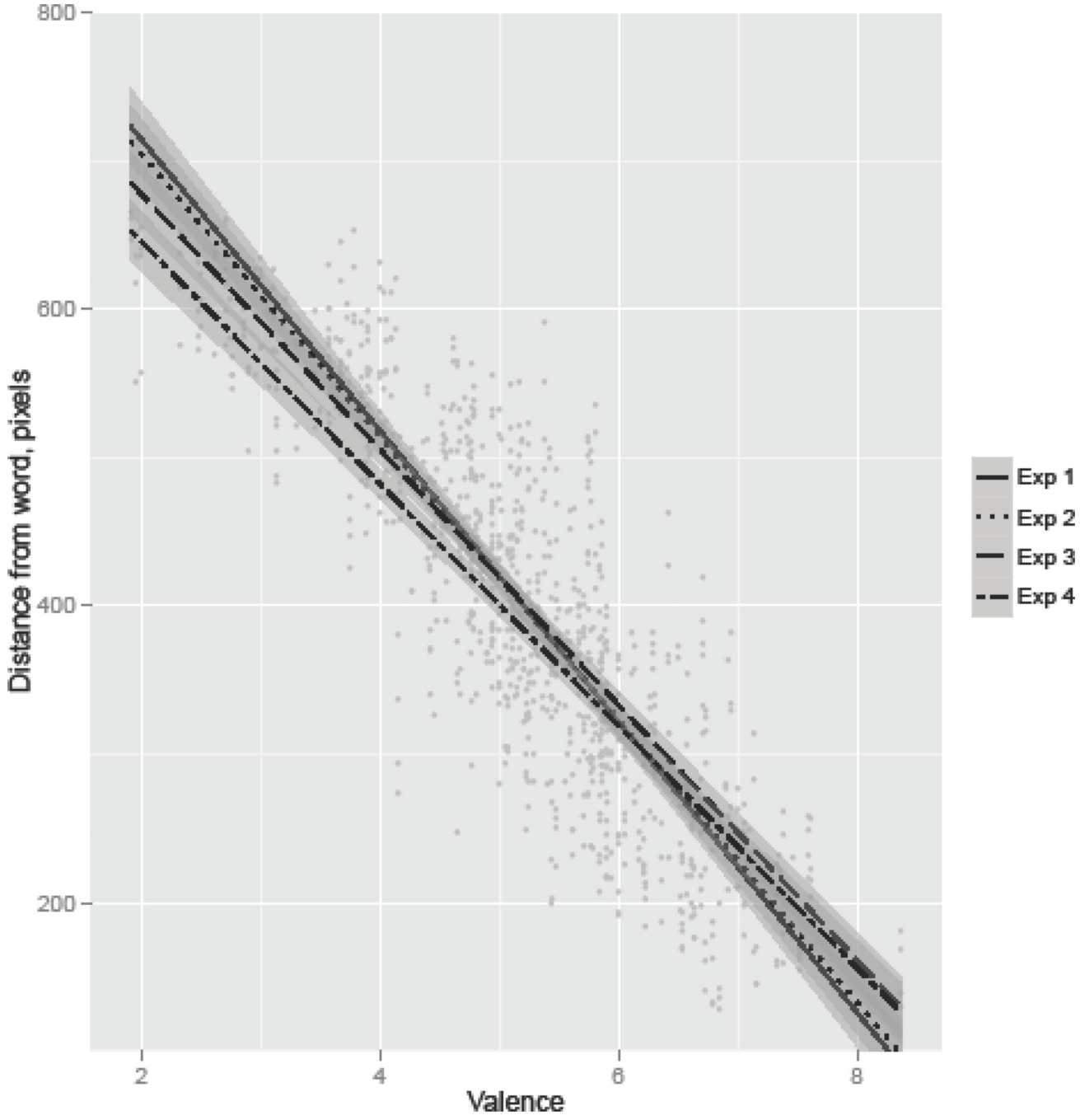

Figure 4.
Scatterplot of the manikin's distance from the word as a function of the word's valence, shown for Experiments 1–4. 95% confidence intervals are displayed in dark grey.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

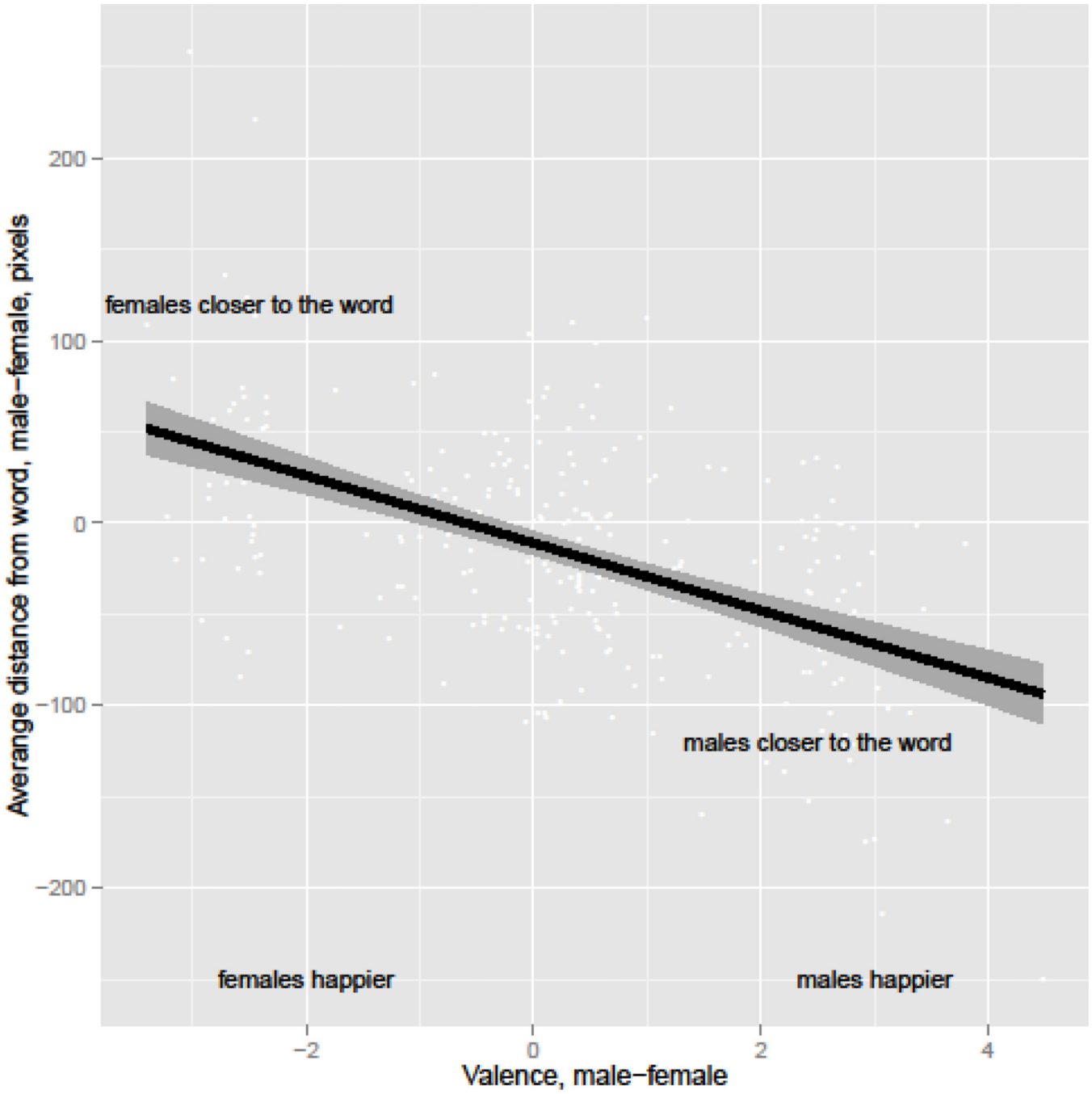

Figure 5.

Difference between distance choices made by men and women as a function of the difference between valence ratings given by men and women. The trend line is in black with the 95% confidence interval presented as a gray area.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Mean, maximum, and minimum scores for each of the scales and their subscales.

	Mean	Min	Max
BAS Drive	10.45	7	15
BAS Fun	10.93	6	15
BAS Reward	17.90	15	20
BAS Approach (D+F+R)	39.28	32	47
BAS Inhibit	23.93	15	28
Alexithymia: Feeling	17.76	8	28
Alexithymia: Describing	15.24	10	20
Alexithymia: External	18.00	13	29
Alexithymia: TOTAL	51.00	36	71
ASQ: Concealing	24.72	13	38
ASQ: Adjusting	22.10	10	34
ASQ: Tolerating	16.76	9	22
ASQ: TOTAL	63.59	43	93
Shyness	8.79	1	20
Sociability	12.93	4	20

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Multiple regression model fitted to the distance of the manikin from the word compared in Experiment 1–4, with Experiment 4 as the reference. R^2 of the model is 0.76.

Predictor	Estimate	SE	t-value	p-value
Intercept	808.243	17.202	46.985	<.001
Valence	−81.502	3.238	−25.171	<.001
Experiment 1	102.080	24.327	4.196	<.001
Experiment 2	86.623	24.327	3.561	<0.001
Experiment 3	41.171	24.327	1.692	0.091
Valence: Experiment 1	−16.514	4.579	−3.606	<0.001
Valence: Experiment 2	−13.704	4.579	−2.993	0.003
Valence: Experiment 3	−4.494	4.579	−0.981	0.33

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Average ratings and values of lexical variables across stimuli subsets showing sex differences for both valence and arousal. M = male; F = female; V = valence; A = arousal. Average natural log frequency is also reported for each subset.

	N	Male V	Fem V	Male A	Fem A	V Diff	A Diff	Log Freq	Length
M happier	50	6.05	3.28	5.43	4.74	2.77	0.69	4.76	7.30
F happier	50	4.85	7.47	4.06	3.93	−2.62	0.13	6.50	6.26
M more aroused	29	5.63	5.00	6.59	3.23	0.62	3.36	4.68	7.90
F more aroused	30	5.47	5.48	2.64	5.30	−0.01	−2.66	5.06	6.43
Remaining	125	5.17	4.92	4.40	4.13	0.24	0.26	4.88	7.72
All words	284	5.34	5.15	4.56	4.24	0.20	0.32	5.14	7.27

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Summary of the regression model fitted to the sex difference in the manikin's distance to the word with sex differences in valence and arousal ratings as predictors. $R^2 = 0.29$.

	Estimate	SE	t-value	p-value
Intercept	−13.362	3.282	−4.071	<.001
Valence difference	−18.118	1.850	−9.794	<.001
Arousal difference	−5.970	1.999	−2.987	.003