Virtual experiments in megastudies: a case study of language and emotion

Victor Kuperman

McMaster University, Canada

Running Head: Virtual experiments

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

Abstract A recent dramatic increase in the number and scope of chronometric and norming lexical megastudies offers the ability to conduct *virtual experiments*, that is, to draw samples of items with properties that vary in critical linguistic dimensions. This paper introduces a bootstrapping approach which (a) enables testing research hypotheses against a range of samples selected in an uniform, principled manner and (b) evaluates how likely a theoretically motivated pattern is in a broad distribution of possible outcome patterns. We apply this approach to conflicting theoretical and empirical accounts of the relationship between psychological valence (positivity) of a word and recognition of that word by conducting multiple virtual experiments based on two lexical decision databases. Analyses of the means and distributional analyses of RTs within and across virtual experiments point to the monotonic negative relationship between valence and lexical decision RT, predicted under the gradient automatic vigilance account, as the dominant pattern of the interplay between word recognition and emotion.

Keywords: emotion, word recognition, distributional analysis, megastudies, virtual experiments

1 Introduction

The last decade has witnessed a dramatic increase in the number and scope of large-scale chronometric and norming lexical megastudies, granting access to behavioral responses to thousands of words obtained from hundreds and thousands of participants. Among other advantages, megastudies offer the ability to conduct *virtual experiments*, that is, to draw samples of items with properties that vary in critical linguistic dimensions. Behavioral responses to the items in such a virtual experiment can then be treated as if they were the outcomes of an experiment designed with the critical manipulation in mind (Keuleers, Diependaele, & Brysbaert, 2010; Sibley, Kello, & Seidenberg, 2009). The benefits of virtual experiments are many. Once the megastudies are made accessible, an infinite number of virtual experiments can be conducted without any extra data collection. While many practical factors such as time and personnel limitations constrain the duration of small-scale experiments, virtual experiments face no such restrictions on the number of items they can include. In addition, smaller-scale experimental lists are typically created to implement specific manipulations resulting in an overrepresentation of items with extreme values of lexical properties. Because megastudies are not created to implement such manipulations, they arguably provide more naturalistic ranges (see Keuleers et al., 2010 for extensive discussion). However, the utility of virtual experiments as replications of small-scale studies hinges on the assumption that task differences between these formats of data collection do not affect Type I and Type II error rates when testing statistical hypotheses on the same stimuli list. This assumption has been debated. Sibley et al. (2009) analyzed naming latencies reported in the English Lexicon Project megastudy (Balota, Yap, Cortese, Hutchinson, Kessler et al., 2007) as well as in Kessler, Treiman, & Mullennix (2002) and Seidenberg and Waters (1989), and failed to replicate the frequency by regularity interaction that had previously been robustly presented in 5 experimental studies, thereby attesting to the inflated Type II error rates when using megastudies. Possible explanations for this failure to replicate include task demands (participants usually respond to a much larger number of words over time in a megastudy than in a small-scale experiment leading to different fatigue and practice effects, and priming and list effects), and item-wise correlations between megastudies. Conversely, Keuleers et al. (2010) reported a perfect convergence between the results of over 10 small-scale lexical decision experiments, covering a range of word processing phenomena, and the data patterns obtained by analyzing lexical decision RTs to the same sets of stimuli, as reported in their Dutch Lexicon Project megastudy. To sum up, the relationship between *actual* and *virtual* experiments remains contested, and the sole tool used so far to (in)validate this relationship appears to be the cross-study comparison of responses to the same stimuli list.

The present study introduces a different utilization of virtual experiments as a robust tool for validating novel or well-established experimental effects. The proposed procedure is to (a) identify selection criteria for stimuli that put to the test a research hypothesis and control for undesirable variance, (b) draw multiple random samples from a set of items for which requisite behavioral and norming data are available, (c) retain for analyses those samples that satisfy criteria predefined in (a), and finally (d) evaluate the probability of theoretically motivated response patterns, both in the individual samples and in the distribution of responses aggregated across all samples. We illustrate this approach by conducting a series of virtual experiments organized around a well-established experimental finding in the area of language and emotion that words with relatively extreme values of psychological valence (rated as very pleasant or very unpleasant) elicit shorter lexical decision RTs than neutral words (Kousta, Vinson, & Vigliocco, 2009).

The effect of valence on word processing

An influential theoretical account of the interplay between emotionality of lexical meaning and word recognition advocates the viewpoint that affective states are grounded in two basic motivational systems, defensive (avoidance) and appetitive (approach), cf. Lang, Bradley, and Cuthbert (1990, 1997). A positive word meaning is likely to engage the approach system, while a negative meaning is likely to trigger an avoidance response. Importantly, stimuli that engage either motivational system, approach or avoidance, benefit from a preferential allocation of attention, as such stimuli are crucial for avoiding danger or gathering resources necessary for survival. Stimuli associated with extreme affective states are predicted to elicit faster processing than those that do not attract motivated attention: thus both positive and negative words are expected to elicit faster responses in lexical decision due to their motivational relevance. Moreover, as the motivational systems are argued to influence behavioral responses equally strongly, no difference between the speed of responding to positive and negative word is expected.

The predicted inverse U-shape of the functional relationship between a word's valence and the lexical decision RT to that word was originally observed in Kousta, Vinson, and Vigliocco's (2009) small-scale experiment with 120 critical words (40 triplets of words) representing the positive, neutral and negative range of valence and are matched triplet-wise on a comprehensive range of lexical and sublexical properties. Later studies using the same stimuli list with different participant cohorts replicated this inverse U-shaped effect of valence on RTs in two standard lexical decision experiments (Vigliocco et al., 2013; Experiment 1 in Yap & Seow, 2013) and one go/no-go lexical decision experiment (Experiment 2 in Yap & Seow, 2013). Moreover, Kousta et al. (2009) found the same qualitative pattern of shorter lexical decision RTs to positive and negative words in a sample of 1446 words from the English Lexicon Project (ELP; Balota et al., 2007), and Vinson et al. (2013) in a sample of 1374 words from the British Lexicon Project megastudy (Keuleers, Lacey, Rastle, & Brysbaert, 2012): for discussion see Kuperman et al. (in press).

This inverse-U effect of valence characterizes the statistical behavior of the mean RT across the range of valence. Yap and Seow (2013) substantially complemented this finding with a distributional analysis of RTs to the same set of words. The core of this analytical technique is the assumption that an RT distribution can be closely approximated by an ex-Gaussian distribution, i.e., a confluence of a Gaussian distribution with the mean μ and standard deviation σ as parameters, and an exponential distribution with the mean (and standard deviation) τ (Balota & Spieler, 1999; see also a review of applications in Balota & Yap, 2011). Ex-Gaussian parameters are estimated for each condition by fitting an ex-Gaussian function to the observed RTs. Differences in estimated ex-Gaussian parameters across experimental conditions are theoretically revealing, as a difference in μ reflects an overall shift of the distribution and points to an effect that influences fast and long responses equally strongly: in other words, a difference in μ indicates an early effect of the manipulated variable. Observing a distributional shift in μ as a component of the effect of valence would imply that this is either "an early preconscious effect that facilitates the perceptual identification of stimuli" (Kousta et al., 2009) or an early lexico-semantic effect (Kuperman et al., in press). Conversely, a difference in τ indicates a late effect that preferentially influences long responses in the heavy right tail of the RT distribution, and thus indicates a slow underlying process. Such an effect could indicate an influence of valence on attentionally demanding stages of response execution in lexical decision (Yap & Seow, 2013). If conditions differ both in μ and τ, a compound, multi-stage effect is to be assumed.

Interpretation of distributional patterns is aided by the nonparametric visualization technique of vincentile plots (Ratcliff, 1979; Vincent, 1912). To produce a vincentile plot, all RTs to a condition are arranged in the increasing order, the mean RTs are calculated for each decile/vincentile, and the mean RTs in each decile are plotted against the decile's ordinal number. When vincentile plots are created separately for each experimental condition, a difference in the mean of the Gaussian distribution μ shows as a constant contrast between mean RTs in every decile: one line is simply shifted upwards against the other line. An increase in the mean of the exponential distribution τ

is reflected in the increasing contrast between the mean RTs that comes with an increase in the decile's ordinal number: conditions diverge more in longer responses. Effects can also be found to influence both μ and τ, causing both an overall contrast between by-condition vincentile plots and a change in the contrast's magnitude across deciles.

Yap and Seow's comparison of RTs to positive, neutral and negative words in the 120-word sample adopted from Kousta et al.'s (2009) study revealed significant differences in both μ and τ between valenced (positive and negative) words and neutral words: neutral words were processed slower (higher μ values) and had a heavier right tail (higher τ values). No significant difference on any ex-Gaussian parameter was found when comparing positive and negative words, in line with the prediction of their equal ability to engage an approach or an avoidance system. The findings of both the rightward distributional shift and a stronger effect on long responses in neutral words led Yap and Seow to conclude that the effect of psychological valence has an early component, potentially reflecting the preconscious facilitation of word perception hypothesized by Kousta et al. (2009). Importantly, however, valence also exerts influence on late stages, typically associated with decision-making and response execution in meta-linguistic tasks like lexical decision.

The "motivated attention" account, along with its body of supporting evidence, is not uncontested (see Larsen, Mercer, & Balota, 2006 and references below). Estes and Adelman (2008a) and Kuperman et al. (in press) propose that the interaction of language and emotion is regulated by the automatic vigilance mechanism (Erdelyi, 1974; Pratto & John, 1991). The central assumption of this account is that the perceptive system is attuned to potentially dangerous, negative stimuli which both capture one's attention faster and for a longer time than positive stimuli do (Fox, Russo, Bowles, & Dutton, 2001; Öhman & Mineka, 2001). As negative words are "released" for processing later than positive ones, they elicit longer processing times. If one further assumes that automatic vigilance is gradient (Kuperman et al., in press), the account predicts a monotonic decrease in RTs as a word's valence increases. This pattern was indeed observed in a regression study by Kuperman et al. (in press) of lexical decision and naming RTs to over 13,000 words from ELP (Balota et al., 2007): more positive words elicited faster responses, with the strength of this speed-up being modulated by word frequency. Given the larger sample of words in their regression analysis (12363 vs 1446 words in Kousta et al., 2009) with its naturalistic range of valence, Kuperman et al. (in press) concluded that the monotonic near-linear negative relationship between valence and lexical decision RTs finds stronger support in the data than the inverse U-shaped relationship observed in Kousta et al.'s (2009) analysis of stimuli from the same ELP megastudy.

An adjudication between conflicting theoretical accounts and accompanying bodies of empirical data is currently incomplete. Only one large-scale regression study exists that supports the automatic vigilance account and the gradient negative relationship between valence and behavioral RTs (Kuperman et al., in press), while the "motivated attention" model and the predicted inverse U-shaped relationship are replicated in three tightly controlled experiments (Kousta et al., 2009; Vigliocco, Clarke, Ponari, Vinson, & Fucci, 2013; Yap & Seow, 2013). To address this debate and demonstrate the merit of virtual experiments, the present paper harnesses the power of available megastudies: ELP (Balota et al., 2007), BLP (Keuleers et al., 2010), frequency lists from the US and UK films and media (Brysbaert & New, 2009; Van Heuven, Mandera, Keuleers, & Brysbaert, 2013), and recent datasets of affective ratings, concreteness ratings and age-of-acquisition ratings (Brysbaert, Kuperman, & Warriner, in press; Kuperman, Stadthagen-Gonzales, & Brysbaert, 2013; Warriner, Kuperman, & Brysbaert, 2013). Our validation procedure takes the selection criteria of Kousta et al. (2009) as a point of departure, draws samples from ELP and BLP megastudies which satisfy these selection criteria, and estimates the probability of observing, within and across samples, the patterns predicted by different theoretical accounts. We then follow Yap and Seow (2013) in complementing analyses of mean RTs via distributional analyses, which go beyond the central tendency in the data and enable a temporal interpretation of the effects.

2 Virtual experiments

2.1 Methods

We obtained ELP lexical decision latencies for all correct responses to existing words. The set was further restricted to words for which all of the following lexical variables were available: affective ratings of valence and arousal (Warriner et al., 2013), frequency counts in the 51 million-token SUBTLEX-US corpus based on subtitles to US films and media (Brysbaert & New, 2009), age-of-acquisition ratings (Kuperman et al., 2013), concreteness ratings (Brysbaert et al., in press) and other sublexical form-related lexical measures available from the ELP (see the full list below). The resulting set of 12363 words was split into negative, neutral and positive tertiles with valence values of 4.65 and 5.68 as cut-off points. A similar procedure was applied to the British Lexicon Project using the 201.3 million-token SUBTLEX-UK (Van Heuven et al., in press) as a source of frequency counts. As a result, lexical decision latencies were obtained for 7125 words from the BLP database with tertile valence cut-off points of 4.74 and 5.71.

We set out to test the generalizability of the inverse-U pattern of the valence effect by conducting a series of virtual experiments, each satisfying the stringent matching criteria proposed by Kousta et al. (2009). We drew 5000 samples with replacement from the ELP and, separately, the BLP data sets, such that each sample contained 40 words from the positive, neutral and negative ranges of valence (see above for valence cut-off points), for a total of 120 words. In every sample, each pair of 40-word subsets (positive-negative, positive-neutral and negative-neutral) was tested for a significant difference of the mean in all of the following lexical dimensions: concreteness, age of acquisition, familiarity, log frequency, orthographic neighborhood, number of letters, number of syllables, number of morphemes, and mean positional bigram frequency. Positive and negative words within each 120-word sample were further tested for differences in arousal. Two-sample independent t-tests of the mean were used. Only those samples in which each individual t-test failed to reject the null hypothesis for each pair of lexical subsets at the 5% level were considered matched. The list of variables that positive, neutral and negative words were matched on was similar to the one used for stimuli creation in Kousta et al. (2009) with two exceptions. First, we used pairwise matching rather than triplet matching. Second, we omitted imageability as a matching criterion: as imageability ratings exist for a relatively small number of words (on a megastudy scale). Using them would halve the number of words available to us and reduce the statistical power of the bootstrapping procedure considerably. We note, however, that we were able to replicate all qualitative data patterns in the much smaller number of samples obtained with imageability as one of matching criteria.

2.2 Results and discussion

These criteria yielded 163 matched samples of 120 words drawn from the 12363 words in the ELP database and 46 matched samples from the 7125 words in the BLP database. No two samples, drawn either from ELP or BLP, shared more than three words, and thus they were practically independent.

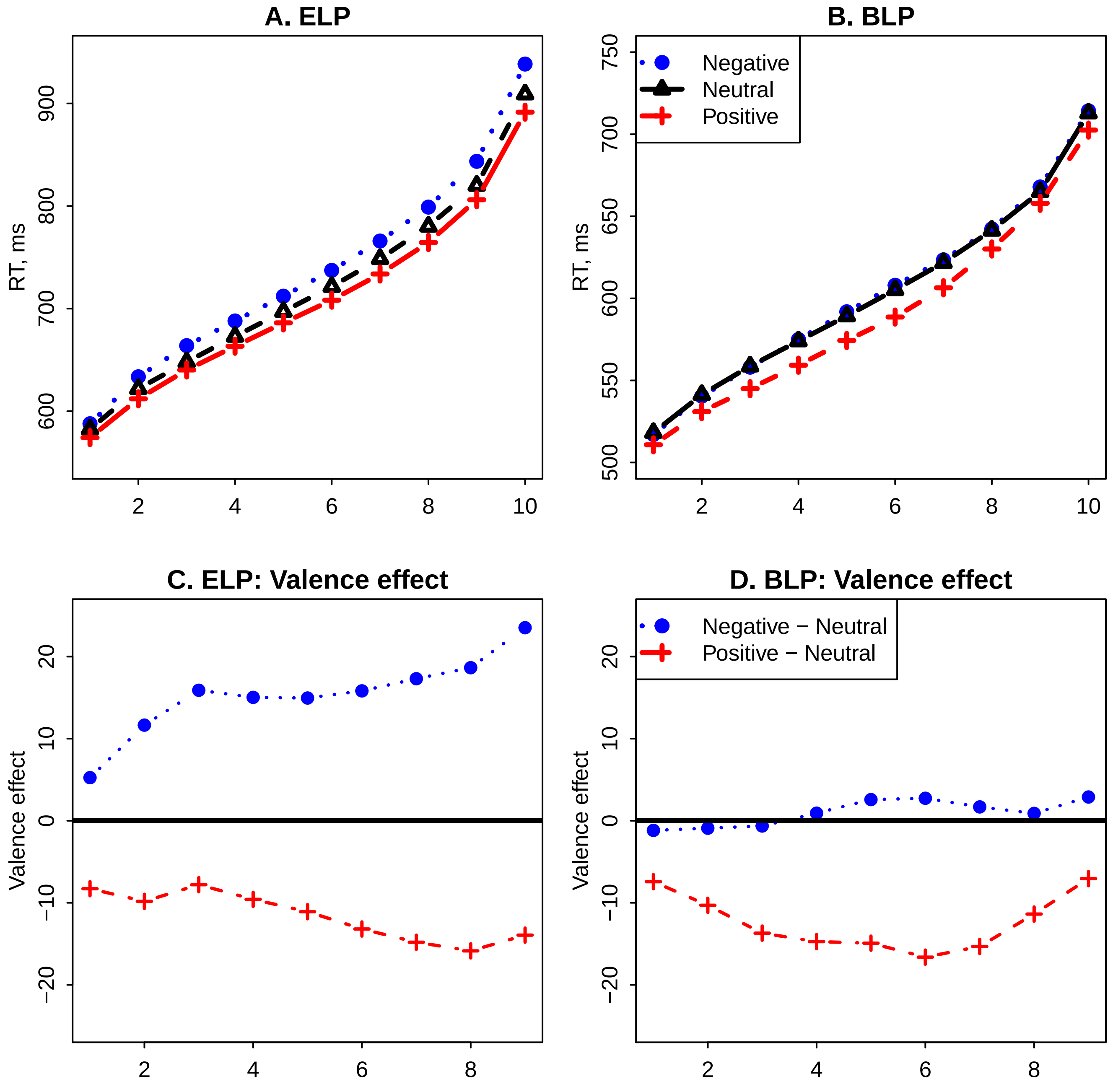
As a first step, we considered an RT distribution aggregated over samples from each source. The design of the ELP megastudy (Balota et al., 2007) was such that participants were assigned to items randomly. As there was no consistent cohort of ELP participants for every word and every experimental condition, we calculated average RTs per word across all participants who responded to that word. For each ELP sample and condition, by-item averages were sorted in increasing order and binned into deciles (vincetiles). An average RT was calculated for each decile in each sample and each condition, such that ten RT values characterized the RT distributions to positive, neutral and negative words in each sample. The final aggregation was to average RTs in decile 1 (2, 3, ...10) across 163 samples for each condition separately: this yielded an aggregate RT distribution for positive, neutral and negative words. For comparability of results, we applied this same procedure to the 46 samples from BLP: RTs to words were averaged across all participants who responded to the word correctly (rather than calculating the decile's mean RT per participant and then averaging the by-participant mean RTs). Resulting vincentile plots are reported in Figure 1A for ELP and 1B for BLP. Bottom panels (Figure 1C and 1D) report differences between the negative and neutral, and the positive and neutral conditions for ELP- and BLP-derived samples respectively.

Analyses of the central tendency: The multi-fold (163- and 46-fold) bootstrapping of the experiment first conducted in Kousta et al. (2009) through a series of virtual experiments gave rise to several important observations. First, the patterns observed across matched samples ran counter to the ones observed by Kousta et al. (2009), Vigliocco et al. (2013), and Yap and Seow (2013). Overall, positive words elicited faster responses than negative or neutral words: the neutral condition was not the slowest but rather positioned between positive and negative words. The difference between negative and neutral words varied by source: it was positive for ELP-derived samples and virtually non-existent for BLP-derived samples. Second, an increased contrast between positive-neutral and negative-neutral words was evident in the RT distribution aggregated over ELP samples, but was absent in the BLP-based distribution. The visual patterns in Figure 1 found confirmation in analyses of mean RTs observed in each condition: see Table 1 for means and standard deviations of RTs to all words observed in respective conditions.

	mean RT	sd RT	mean μ	sd μ	mean σ	sd σ	mean τ	sd τ
ELP								
negative	740.16	103.32	663.21	39.60	67.63	23.97	74.68	41.36
neutral	721.52	97.93	649.45	32.91	63.97	23.28	71.51	36.51
positive	703.68	94.81	638.94	34.35	61.25	21.92	70.04	36.40
BLP								
negative	606.51	60.37	564.97	26.70	44.13	17.25	39.59	25.00
neutral	604.91	58.81	563.97	27.11	40.67	16.78	39.69	27.52
positive	587.24	57.53	535.59	16.04	30.72	14.62	58.14	19.80

Table 1: Means and standard deviations of observed RTs, and of the ex-Gaussian distribution parameters averaged across 163 ELP samples and 46 BLP samples

Figure 1: Vincentile plots for positive, neutral and negative words obtained from 163 matched samples from ELP (panel A) and 46 BLP (panel B). Differences between the negative and neutral, and the positive and neutral conditions, as a function of valence are reported for ELP (panel C) and BLP (panel D). Error bars in panels C and D represent the standard error of the mean difference. Error bars are not plotted in panel A for the sake of legibility.



On average, positive words were responded to significantly faster than neutral or negative words in both the ELP and BLP (all ps < 0.01 in two-tail two-sample paired t-tests). Additionally, in the distribution of RTs to words from ELP samples, neutral words were responded to faster than negative words ($p < 0.05$), whereas the difference did not reach significance in the BLP samples ($p > 0.1$).

We supplement the analysis of distributional patterns observed in the entire set of samples by a consideration of patterns within individual samples. A central claim of Kousta et al. (2009), corroborated in all subsequent studies using their sample, is that the functional relationship of valence and lexical decision latencies has an inverse-U shape. An additional specification of the inverse-U shape is that negative and positive words are processed equally fast, reflecting the respective engagement of two equally strong motivational systems (approach and avoidance). To establish the number of 120-word sets in which this pattern holds, we applied to each set a one-tail two-sample paired t-test comparing mean RTs to positive (faster) and neutral words, a one-tail two-sample paired t-test comparing mean RTs to negative (faster) and neutral words, and a two-tail two-sample paired t-test comparing mean RTs to negative and positive (equally fast) words. The significance threshold was set at $0.05/3 = 0.017$, using the Bonferroni correction for multiple comparisons. The number of samples that satisfied all criteria for the inverse-U shape was 3 or 2% of samples drawn from ELP, and 1 or 2% of samples drawn from BLP. Thus, the inverse U-shaped relationship with equal processing advantages to positive and negative words does occur in individual samples, but is not supported by the cross-sample distributions as a dominant pattern.

An alternative view advocated in the regression analysis of Kuperman et al. (in press) is that valence has a gradient monotonic negative effect of word recognition latencies, such that relatively positive words are responded to faster than relatively negative words. On this account, positive words are expected to show shorter RTs than both neutral and negative words, and neutral words than negative words. To each of the 120-word sets we applied three one-tail two-sample paired t-tests of the mean: the significance threshold was set to 0.017. The number of samples from ELP in which all paired t-tests showed a significant difference between condition means was 28 or 17% of samples drawn from ELP, and 19 or 41% of samples drawn from BLP. The pattern of the gradient decrease in RTs that comes with increasing valence finds a stronger support in the variety of virtual experiments that adopted the original experimental manipulation. While clearly a dominant pattern, the negative relationship between valence and RT is observed in a lower-than-expected number of samples. We elaborate on this issue in the General Discussion.

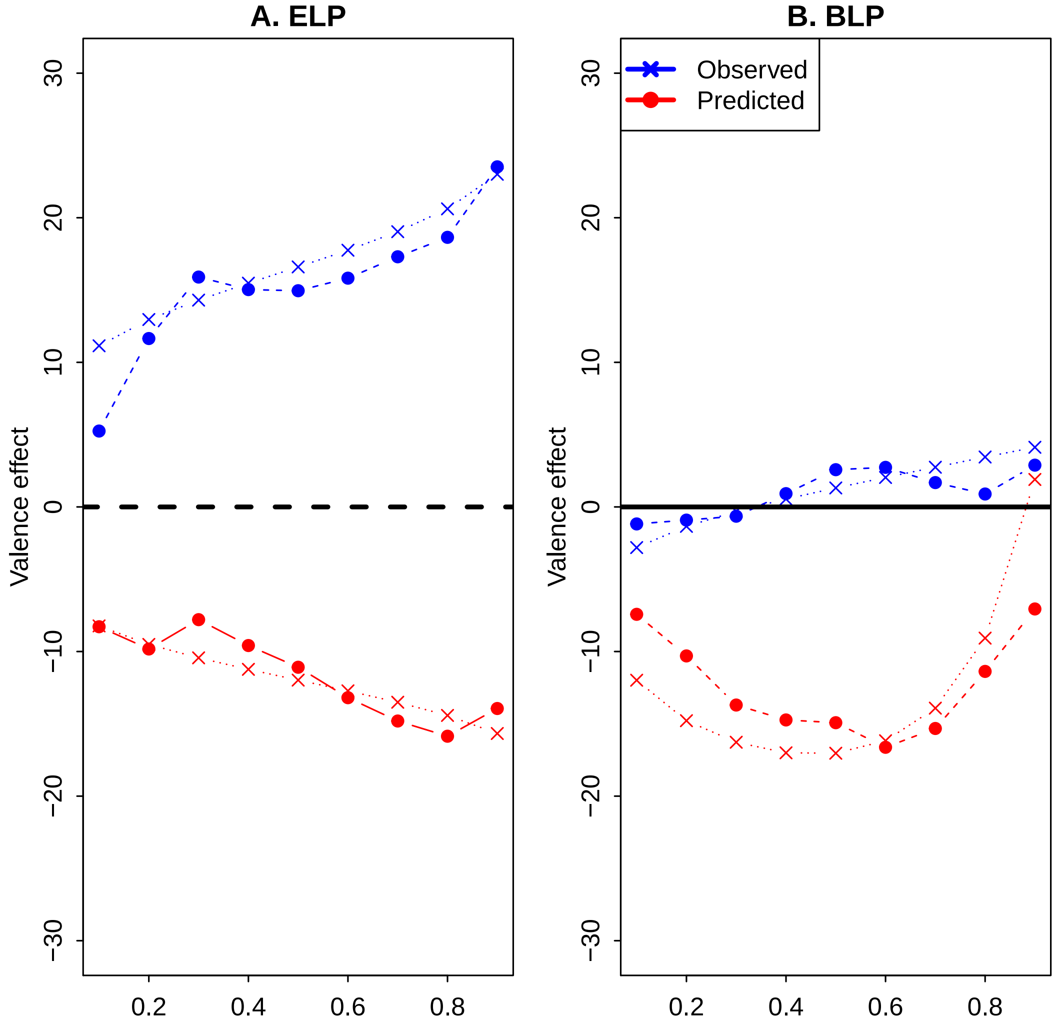
Distributional RT analyses: We further fitted ex-Gaussian distributions to word-averaged RTs for positive, neutral and negative words in each of the ELP- and BLP-derived samples. The quantile maximum likelihood estimation procedure in the QMPE software package (Cousineau, Brown, & Heathcote, 2004; Heathcote, Brown, & Cousineau, 2004) was used to estimate ex-Gaussian parameters μ, σ, and τ. No more than 60 iterations was required for all fits to converge successfully. Table 1 reports parameters of ex-Gaussian distributions averaged across the matched samples from ELP and BLP. Figure 2 demonstrates the high accuracy of predicted RTs that are based on estimated parameters of the ex-Gaussian distribution. Both for the RT distribution averaged across ELP samples (panel A) and BLP samples (panel B), the predicted differences between positive and neutral, and neutral and negative words were within a 5 ms range from the observed ones. As Table 1 demonstrates, in the ELP samples, μ values were significantly smaller for neutral than negative words, and for positive than neutral or negative words (all ps <= 0.005 in paired t-tests, i.e. significant after the Bonferroni correction for multiple comparison). Critically, neither σ nor τ values differed significantly in any of the pairwise comparisons (all ps > 0.1). Taken together, these results (a) disconfirm the inverse-U shape of the emotion effect and rather point to the gradient speed-up associated with positivity, and (b) suggest that the effect of valence is early and does not preferentially influence exceedingly long RTs.

Patterns observed across BLP patterns reveal that RTs to negative and neutral words form virtually identical ex-Gaussian distributions. Positive words, however, come with significantly smaller μ and σ values and a significantly larger τ value than other conditions (all ps < 0.01 in paired t-tests). This indicates a compound two-stage process. Positive words appear benefit from an early advantage, reflected in a speed-up of the entire distribution of responses (as in the ELP data). This advantage for positive words but not for negative ones is unexpected under the “motivational relevance” account but does not contradict the gradient automatic vigilance account. However, this advantage is reversed in positive words that elicit very long responses, caused by a slow-down to the positive words in the right tail of the RT distribution.

3 General Discussion

This study explores the currently underutilized potential of behavioral and norming mega-studies to bootstrap factorial manipulations via a series of virtual experiments. The typical notion of a virtual experiment is a one-time replication of a completed experimental study through the cross-

Figure 2: Vincentile plots of the differences between observed and predicted RTs to positive, neutral and negative words in samples drawn from ELP (left) and BLP (right).



check of behavioral responses to its stimuli list study in a response database. A comparison of the responses obtained in the hypothesis-driven small-scale experiment and the hypothesis-blind large scale megastudy is typically interpreted as (in)validation of the results of the small-scale experiment or, more commonly, of that megastudy. We propose a more robust validation procedure that takes an existing or novel experimental manipulation as a point of onset, draws multiple random samples from the database of behavioral responses and retains only those that follow the stringent matching criteria of that chosen manipulation. A series of virtual experiments conducted in such a way allows for an exhaustive characterization of behavioral patterns associated with the manipulation, including an estimation of the probability of observing a specific pattern in any given experiment. The *sine* *qua non* of this approach is availability – through megastudies – of both extensive collections of behavioral data (represented for the lexical decision in English by the English and British Lexicon Projects, Balota et al., 2007, Keuleers et al., 2010), and equally broad collections of objective or subjective norms for a variety of lexical variables. In what follows we first summarize our series of virtual experiments originating from the study of language and emotion by Kousta et al. (2009), and then discuss what multiple virtual experiments can and cannot achieve as a methodological paradigm.

Validation of the valence effect on lexical decision latencies

Recent work reporting an effect of valence on lexical decision latencies provided the focus of this study. A series of standard and go/no-go lexical decision experiments (Kousta et al., 2009; Vigliocco et al., 2013, Yap & Seouw, 2013) used the same set of stimuli with psychological valence (positivity) as a critical manipulation and tight matching control over a number of potential lexical confounds. All studies reported an inverse U-shaped relationship between valence and RTs, such that positive and negative words elicited equally fast responses, and both conditions were responded to significantly faster than neutral words. Distributional analyses of Yap and Seow additionally showed that valence does not exclusively affect early stages of word processing (as reflected in the shift of μ between the positive, neutral and negative words) but also impacts long responses, implicating an influence on slow, attentional processes in word recognition. This body of evidence supports the "motivational relevance" account, with very positive and negative words benefitting from allocation of additional attentional resources associated with the approach or avoidance motivational systems (Lang, Bradley, & Cuthbert, 1990, 1997).

The present series of virtual experiments adopted all essential aspects of the stimulus selection procedure proposed in Kousta et al. (2009) to create 163 samples from the ELP database and 46 samples from the BLP database, with each sample of 120 words equally representing the three subranges of valence and matched on a variety of lexical and sublexical properties. The key results of the analyses of the means, and distributional analyses of the samples are as follows. First, the dominant pattern in the response time distribution within samples and across samples is one in which positive words elicit faster responses than negative and neutral words, and negative words were either slower than neutral ones (ELP) or equally slow as the neutral ones (BLP). This pattern characterizes both the RT distribution aggregated over ELP and BLP samples, and the estimates of the ex-Gaussian parameter μ, see Table 1. Moreover, the gradient advantage of positivity is attested in a larger percentage of individual samples, as compared to the inverse-U alternative. The inverse-U shape of the valence effect on lexical decision latencies, with both positive and negative words eliciting equally fast responses and both being significantly faster than responses to neutral words, is found in less than 3% of the samples from either behavioral database.

Second, distributional analyses revealed differences between behavioral databases in the temporal locus of the valence effect. The RT distribution based on the ELP samples showed an impact of valence on the mean μ of the Gaussian component only: the values of μ increased significantly from positive to neutral to negative words. This finding suggests that the effect of valence is as early as this analytic paradigm is capable of detecting, and affects the entire distribution of RTs. It also supports the lexico-semantic locus of the valence effect and dovetails well with the interpretation of the interaction of valence by frequency in Kuperman et al. (in press), Scott, O'Donnell, & Sereno, (2012), and Sheikh and Titone (2013). No evidence is found in this set of samples in favor of the specific influence of valence on the response execution phase of the lexical decision task, contra Yap and Seow (2013).

Conversely, the RT distribution based on the BLP samples showed that positive words come with a lower value of μ and also a higher value of τ than neutral or negative words: no difference was found in the parameter estimates for neutral vs negative words. This indicates an early advantage to positive vs other words, which is compatible with the lexico-semantic locus of the effect discussed above. The pattern is however coupled with a delayed disadvantage that inflates very long responses to positive words. We do not have an explanation for the behavior of this component. We do note that estimates of ex-Gaussian parameters for the BLP data should be treated with caution for two reasons. First, there is an overall 100-ms advantage in RTs that participants in the BLP study demonstrate as compared to those in the ELP study. It is possible that patterns observed in BLP samples bear more similarity to the patterns of very fast participants in the ELP than the entire cohort of ELP participants: this testable hypothesis requires further investigation. Second, ratings of valence that we use were collected from US responders rather than British ones, yet they were applied for analyses of both the ELP and BLP data. This fact may account for the apparent lack of difference between negative and neutral words in BLP data. We leave the methodological scrutiny of the discrepancy between the ELP and BLP data to future research and confine ourselves to an observation that neither the ELP- nor the BLP-derived patterns replicate the findings of Yap and Seow (2013) based on lexical decision latencies to the 120-word sample of Kousta et al. (2009).

Why is there such a drastic discrepancy between the present results and results of the original study (Kousta et al., 2009) and subsequent replications (Vigliocco et al., 2013; Yap & Seow, 2013)?

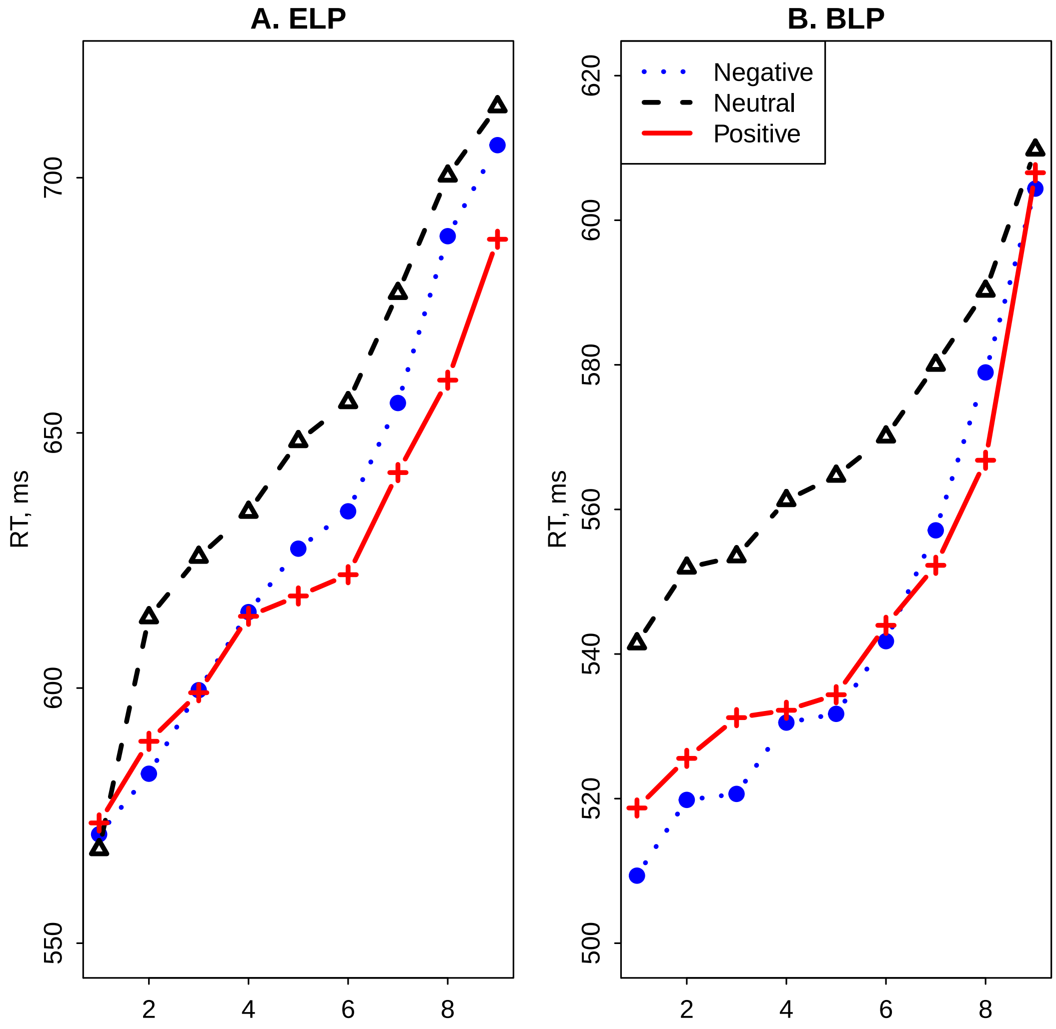
One possibility is the cross-study differences in the list and context effects: pseudowords varied between the ELP, BLP and Kousta et al.'s study, as did the participants, as did the length of the stimulus list presented for lexical decision, as did the probability of encountering a positive, neutral or negative word in the stimulus list. It is possible then that discrepancies in the task or the population altered the responses such that the effect of valence elicited by the original word list is no longer present. We ruled out this possibility by considering distributions of RTs to words from Kousta et al.'s list in both the ELP and BLP databases. The overlapping sets included 112 words from the ELP database, and 93 words from the BLP database. Figure 3 summarizes these results: in both databases, RTs to the original word list replicated well the critical findings of Kousta et al. RTs to negative and positive words were not significantly different from each other in the ELP (a two-tailed t-test: t = -0.77, df = 71.8, p = 0.44) and the BLP sample (t = 0.18, df = 57.0, p = 0.86). Positive words elicited faster responses than neutral ones (one-tailed two-sample t-test, ELP: t = -1.97, df = 70.5, p = 0.03; BLP: t = -2.00, df = 50.0, p = 0.03). RTs to negative words were significantly shorter than to neutral words in the BLP sample (one-tailed two-sample t-test: t = -2.35, df = 56.0, p = 0.01) and there was a numerical tendency in the expected direction in the ELP sample (one-tailed two-sample t-test: t = -1.19, df = 73.3, p = 0.12)[1]. Importantly, the inverse-U shape of the valence effect observed in studies using one and the same word list (Kousta et al., 2009; Vigliocco et al., 2013; Yap & Seow, 2013) was also found in the ELP and BLP data. That is, this specific set of stimuli is confirmed to elicit the same qualitative pattern of speedier responses to positive and negative words than to neutral ones in two more data sets, the ELP and the BLP megastudies. Thus, whatever the discrepancies were in the administration, partipant cohorts, or stimuli of the small-scale and large-scale experiments we discuss, they were not responsible for the change in a magnitude or direction of the valence effect.

What appears to underlie the body of evidence in favor of the inverse U-shaped effect of valence on lexical decision RTs – and thus in favor of the equally strong activation of motivational approach and avoidance systems – is a single, carefully constructed set of words that happens to represent a characteristic behavioral pattern only observed in a minority (3% or less) of samples derived from the same selection criteria. The predominant pattern is that of a monotonic gradient decrease in response times to relatively positive words, which can be best explained by the theoretical account

[1] Stable QMPE estimates of ex-Gaussian parameters require 40 or more observations in the conditions under comparison, and the number of datapoints in the samples overlapping between Kousta et al.'s word list in ELP and BLP is substantially smaller (93 vs 120 datapoints in BLP). For this reason, we did not check whether these parameters replicate the distributional effects that Yap and Seow observed.

of the gradient automatic vigilance advocated in Kuperman et al. (in press).

Figure 3: Vincentile plots for positive, neutral and negative words from Kousta et al.'s (2009) sample in the ELP (panel A) and BLP (panel B).



What virtual experiments can and cannot do

As demonstrated above, virtual experiments enable an estimation of how probable a certain pattern is when pitted against similarly selected samples of items. The present study scrutinizes issues of item selection in factorial designs. The technique we propose can also be readily used for bootstrapping participants of a megastudy, thus ensuring that the resulting patterns are not specific to a participant cohort. Furthermore, the criteria for sample formation can be changed so that the critical contrasts as well as control variables are represented in a continuous rather than categorical manner, thus enabling regression analyses rather than factorial ones.

There are several methodological issues that multifold sampling from megastudies is not designed to resolve. For instance, this technique may not be helpful if the research question crucially hinges on the context in which items occur, such as list or block effects, short- or long-distance priming, or any other manipulation which requires a certain order of and distance between experimental items. Samples from stimuli lists of the megastudy will not generally comply to specifications like these.

Moreover, drawing multiple samples does not decrease the probability of a Type II error in any individual sample, i.e. it does not increase statistical power. In this paper, we adopted the factorial design in which the critical manipulation is represented by 40 words in each of the three conditions. The power analysis of a one-way ANOVA with the nominal 0.05 alpha level suggests that this design has an excellent probability of detecting a large-size effect (f = 0.4, p = 0.98), but a much lower probability of detecting a medium-size effect (f = 0.25, p = 0.68) and an even lower probability of detecting a small-size effect (f = 0.1, p = 0.15): see Cohen (1992) for classification of effect sizes. In our samples, effect sizes varied from small to medium. We believe that the reduced power of factorial designs, and a relatively small sample size adopted in this specific design, can explain why only 17% of samples drawn from ELP, and 41% of samples drawn from BLP showed the dominant pattern of a gradient negative effect of valence on lexical decision RTs. That said, while increasing the size of even one sample that is manually matched on dozens of lexical variables may require major time investment for a small-scale experiment, a trivial adjustment of the program code achieves the desideratum if virtual experiments are used. Alternatively, virtual experiments can be used to create samples with natural variability of critical and control variables, to be analyzed with the help of regression techniques: such techniques are shown in Baayen (2010), among others, to increase statistical power in analyses of continuous data as compared to techniques that bin continuous variables into discrete categories.

Furthermore, drawing multiple samples that manipulate variable A and are matched on variable B does not fully protect from collinearity between A and B. For instance, relatively positive words occur in language more commonly than negative ones and thus the effect of word frequency may confound the emotion effect. Every sample that we drew from the ELP and BLP data sets was matched pairwise such that log frequencies of positive, neutral and negative words were not significantly different from each other (all ps > 0.05). However, when considered across all samples (163 from ELP, and 46 for BLP) the pool of positive words (40 in each sample) is on average significantly more frequent than that of neutral words, and both are more frequent than the pool of negative words. That is, while the differences in log frequency are above the significance threshold in individual t-tests, the differences accumulate across multiple samples and lead to a data set that replicates the correlation of frequency and valence that is found in the data population the samples are drawn from. For this reason, it is advisable to complement the analyses of the RT distribution aggregated over samples by a consideration of patterns in each individual sample, as done in this study. Another, more efficient solution would be to eschew the factorial design and apply regression techniques which are able to estimate the contribution of every variable while taking under statistical control the contributions of other, potentially collinear, variables[2].

To conclude, one of the currently underutilized benefits that megastudies afford is the ability to test one's hypothesis against a range of samples selected in an uniform, principled manner. Virtual experiments can be fruitfully used not only for replicating the results of a small-scale laboratory experiment, but also for shedding light on how likely the results of that experiment are in a much broader distribution of possible outcome patterns.

References

Baayen, R. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, 5(1):149–157.

Balota, D., Yap, M., Cortese, M., Hutchison, K.A.and Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39:445–459.

Balota, D. A. and Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, 128(1):32.

Balota, D. A. and Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry the power of response time distributional analyses. *Current Directions in Psychological Science*, 20(3):160–166.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera 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*, 41(4):977.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (in press). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*.

[2] We are indebted to James Adelman for raising the point discussed in this paragraph

Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1):155–159.

Cousineau, D., Brown, S., and Heathcote, A. (2004). Fitting distributions using maximum likelihood: Methods and packages. *Behavior Research Methods, Instruments, & Computers*, 36(4):742–756.

Erdelyi, M. H. (1974). A new look at the new look: perceptual defense and vigilance. *Psychological review*, 81(1):1.

Estes, Z. and Adelman, J. (2008). Automatic vigilance for negative words is categorical and general. *Emotion*, 8(4).

Fox, E., Russo, R., Bowles, R., and Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General*, 130(4):681.

Heathcote, A., Brown, S., and Cousineau, D. (2004). Qmpe: Estimating lognormal, wald, and weibull rt distributions with a parameter-dependent lower bound. *Behavior Research Methods, Instruments, & Computers*, 36(2):277–290.

Kessler, B., Treiman, R., and Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language*, 47:145–171.

Keuleers, E., Diependaele, K., and Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*, 1:1–174.

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods*, 44(1):287–304.

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

Kuperman, V., Estes, Z., Brysbaert, M., and Warriner, A. B. (in press). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*.

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

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological review*, 97(3):377.

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1997). Motivated attention: Affect, activation, and action. In Lang, P. J., Simons, R., and Balaban, M. T., editors, *Attention and orienting: Sensory and motivational processes*, pages 97–135. Hillsdale, NJ.

Larsen, R., Mercer, K., Balota, D., and Strube, M. (2008). Not all negative words slow down lexical decision and naming speed: Importance of word arousal. *Emotion*, 8(4):4454–52.

Öhman, A. and Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological review*, 108(3):483.

Pratto, F. and John, O. (1991). Automatic vigilance: the attention-grabbing power of negative social information. *Journal of personality and social psychology*, 61(3):380.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological bulletin*, 86(3):446.

Scott, G., O'Donnell, P., and Sereno, S. (2012). Emotion words affect eye fixations during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(3):783.

Seidenberg, M. and Waters, G. (1989). Reading words aloud-a mega study. In *Bulletin of the Psychonomic Society*, volume 27, pages 489–489.

Sheikh, N. A. and Titone, D. A. (2013). Sensorimotor and linguistic information attenuate emotional word processing benefits: An eye-movement study. *Emotion*, 13(6):1107–1121.

Sibley, D. E., Kello, C. T., and Seidenberg, M. S. (2009). Error, error everywhere: A look at megastudies of word reading. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 1036–1041.

van Heuven, W. J., Mandera, P., Keuleers, E., and Brysbaert, M. (in press). Subtlex-uk: A new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology.*

Vigliocco, G., Clarke, R., Ponari, M., Vinson, D., and Fucci, E. (2013). Feeling visible and invisible words: Emotional processing is modulated by awareness in first and second language. Presented at the 54th annual meeting of the Psychonomic Society, Toronto, Canada.

Vincent, S. B. (1912). *The functions of the vibrissae in the behavior of the white rat*, volume 1. University of Chicago.

Vinson, D., Ponari, M., and Vigliocco, G. (in press). How does emotional content affect lexical processing? *Cognition & emotion.*

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

Yap, M. J. and Seow, C. S. (in press). The influence of emotion on lexical processing: Insights from rt distributional analysis. *Psychonomic bulletin & review.*