The graded effect of valence on word recognition in Spanish

Javier Rodríguez-Ferreiro

Universitat de Barcelona

Rob Davies

Lancaster University

Author note: Javier Rodríguez-Ferreiro, Departament de Cognició, Desenvolupament i Psicologia de l'Educació, Institut de Neurociències, Universitat de Barcelona; Rob Davies, Department of Psychology, Lancaster University. We thank González-Nosti et al. (2014) for the lexical decision data. This work was supported by grant PSI2016-80061-R (AEI/FEDER, UE). Correspondence concerning this article should be sent to Javier Rodríguez-Ferreiro, Departament de Cognició, Desenvolupament i Psicologia de l'Educació, Universitat de Barcelona, Psg de la Vall d'Hebron 171, 08035, Barcelona, Spain, eMail: rodriguezferreiro@ub.edu

Abstract

The influence of emotional content on language processing remains unclear. Previous research conducted in English has obtained contradictory results regarding the effects of valence on word recognition. Whereas some studies indicate that valence predicts recognition latencies such that negative words are processed more slowly than positive words, other studies indicate facilitation of responses to emotional (both positive and negative) compared to neutral words. We examined the influence of valence and arousal on word recognition reaction time using large-scale word naming and lexical decision data-sets in Spanish. We found that linear mixed-effects model estimates revealed a valence but not an arousal effect on reading latencies. The influence of valence was better captured by a graded (RTs to positive words < neutral < negative) than by a categorical (positive < negative) valence effect. A categorical emotional vs. neutral effect was not reliably observed. In an advance on previous research, our analyses showed that the valence effect is substantially more prominent in lexical decision than in pronunciation. These results mirror some of those reported previously in English, adding evidence to support their validity, and demonstrating important parallels in word recognition processes in orthographically shallow as well as deep languages.

Keywords: valence; word recognition; Spanish; reading; lexical decision

The interplay between emotion and cognition is central to the study of human psychology (Russell, 2003). Emotion is usually characterized in a bidimensional space framed by the theoretically orthogonal dimensions of valence and arousal (Bradley & Lang, 1999; Osgood, Suci, & Tannenbaum, 1957; Russell, 2003), and is argued to modulate our focus of attention, directly influencing word processing (Lang, Bradley, & Cuthbert, 1997). Valence refers to the pleasurable nature of a stimulus, ranging from negative or unpleasant to neutral to positive or pleasant, whereas arousal refers to the degree of activation elicited by a stimulus, ranging from calming to exciting. Despite considerable effort expended in previous studies, based either on direct experimental manipulation of the emotional qualities of words (Kanske & Kotz, 2007; Kousta, Vinson, & Vigliocco, 2009) or on analyses of data gathered from large-scale studies (Algom, Chajut, & Lev, 2004; Estes & Adelman, 2008a; Kousta et al., 2009; Kuperman, Estes, Brysbaert, & Warriner, 2014; Vinson, Ponari, & Vigliocco, 2014), the effects of valence and arousal during word recognition remain unclear. We report findings from an analysis of observations on word recognition in Spanish, in the word naming and lexical decision tasks. Our analyses help to clarify the influence of emotion on word recognition, in a shallow orthography, and under varying task demands.

According to the Automatic Vigilance model of emotion (Pratto & John, 1991), undesirable aversive events are more likely to retain attention than neutral or pleasant ones. This feature of negative stimuli complicates the disengagement of attention, delaying a possible response (Fox, Russo, Bowles, & Dutton, 2001). The effect of such a bias could then be taken to explain evidence of increased reaction times for negative words in a variety of tasks including the Stroop test (Algom et al., 2004; Pratto & John, 1991) as well as lexical decision or word naming tasks (Algom et al., 2004; Estes & Adelman, 2008a; Kuperman et al., 2014; Yao et al., 2016).

The model of Motivated Attention and Affective States (Lang et al., 1997), on the other hand, proposes that motivationally relevant events, including both positive and negative stimuli, are more likely to attract attention compared to affectively neutral events, thus speeding responses to emotional words. This hypothesis is supported by emotional facilitation effects obtained in lexical decision experiments in which negative and positive stimuli elicited faster reaction times than neutral words (Citron, Weekes, & Ferstl, 2013; Kanske & Kotz, 2007; Kousta et al., 2009; Palazova, Mantwill, Sommer, & Schacht, 2011; Vinson et al., 2014).

A further inconsistency among the results of previous studies relates to the role of arousal during word recognition and its possible interaction with valence effects. Thus, whereas Kousta et al. (2009) and Vinson et al. (2014) reported effects of emotional valence with no significant influence of arousal, Estes and Adelman (2008a) and Kuperman et al. (2014) observed independent effects of both arousal and valence, with arousing words being recognized more slowly than calming words.

Finally, there is a debate concerning whether valence effects are graded or categorical. Kousta et al. (2009), who observed facilitatory effects of both negative and positive stimuli, and Kuperman et al. (2014), who observed increased reaction times for negative words, attributed their otherwise contradictory results to a graded measure of valence. In contrast, Vinson et al. (2014) observed significant facilitation for positive and negative words, compared to neutral words, as a categorical emotion effect. Furthermore, Estes and Adelman (2008b) showed that the interaction between arousal and valence observed by Larsen et al. (2008) appeared only when valence was entered in the analyses as a continuous factor but not when it was coded as a categorical, positive vs. negative, variable.

In the current article, we present a set of analyses in which we tested the influence of affective content on word naming and lexical decision reaction times obtained from previously gathered data in Spanish (Davies, Barbón, & Cuetos, 2013; González-Nosti, Barbón, Rodriguez-Ferreiro, & Cuetos, 2014). Given the inconsistencies observed among the results of previous research, we aimed to clarify the form of the valence effect on word recognition. Does emotional valence have an effect, and, if it does, what is the best measure for capturing the effect?

Spanish is a language in which the spelling-to-sound mappings are regular so that its orthography is characterized as shallow or transparent. Research in English has limited the observation of the influence of semantic content on reading performance to low frequency irregular words that are harder to encode phonologically (Plaut, McClelland, Seidenberg, & Patterson, 1996; Strain, Patterson, & Seidenberg, 1995; Woollams, Lambon-Ralph, Plaut, & Patterson, 2007; but see Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Cortese & Khanna, 2007; Monaghan & Ellis, 2002). However, Davies et al. (2013) and González-Nosti et al. (2014) reported that a semantic component is apparent among the factors that influence oral reading in Spanish, suggesting that the influence of semantic content may be found more broadly than has previously been found (cf. Ricketts, Davies, Masterson, Stuart, & Duff, 2016). Davies et al. (2013) and González-Nosti et al. (2014) did not investigate the effects of valence or arousal. Finding such effects would therefore add evidence suggesting that reading processes are richly influenced by semantic information, not just imageability or, arguably, Age-of-Acquisition (Balota et al., 2004; Brysbaert & Ghyselinck, 2006; Cortese & Khanna, 2007; Davies et al., 2013), but also by valence or arousal. Extending observations on the shape of the emotion effects to a transparent orthography is thus a critical contribution of the present study.

Another important contribution lies in the fact that by comparing emotion effects on reading in different tasks we were able to examine the locus of the effects. Psycholinguistic effects -- most prominently, effects associated with lexical or semantic knowledge -- are typically reported to be larger in lexical decision than in reading aloud (e.g. in English, Balota et al., 2004; Cortese & Khanna, 2007; Davies, Arnell, Birchenough, Grimmond, & Houlson, 2017). Critically, a comparison between word naming and lexical decision in Spanish provides valuable information on the extent to which the effects of emotion variables were moderated by the effects of task demands and therefore (Cortese & Khanna, 2007) the extent to which the impact of emotion variables could be linked to reliance on semantic processing in word recognition (Chumbley & Balota, 1984) in transparent orthographic systems.

One potentially important source of the discrepancies between the results of previous studies is related to the differing extent to which possible psycholinguistic confounds were taken into account in different analyses. In a review of 32 studies, Larsen, Mercer and Balota (2006) showed that lexical variables like length, frequency or orthographic lexical density were confounded with valence differences and therefore the effects of these variables were potentially responsible for what had been interpreted as valence effects in the Stroop task. Although recent studies using lexical decision and word naming tasks have invested considerable effort in the control of potentially confounding lexical variables, different research groups have focused on different sets of variables. This could explain, at least in part, the inconsistency among results. For example, Kuperman et al. (2014) included in their models different measures of word length, lexical density and lexical frequency, as well as Age-of-Acquisition (AoA), Context Distinctiveness (CD) and initial phoneme information, whereas Vinson et al. (2014) included only one predictor variable each to capture effects of length, frequency and density, but introduced positional bigram frequencies and extended the control of lexical-semantic factors with the inclusion in models of, not only AoA, but also concreteness, imageability and familiarity measures. In our analyses, we included a large set of variables as control factors to facilitate comparability with previous research and to strengthen our results by isolating the impact of emotional factors, over and above the effects of better known psycholinguistic variables.

We investigated the composition of emotion effects in reading, examining the impact of valence and arousal on word recognition in Spanish. We investigated the shape of the valence effect, examining whether the valence effect should best be described as a graded (positive-neutral-negative) or a categorical positive-negative valence effect, or as a categorical emotional-neutral effect. Close examination of recent reports (e.g. Kuperman et al., 2014; Vinson et al., 2014) makes it apparent that a number of alternate routes can be taken, and have been taken, through the process of analysing word recognition data to uncover the effects of emotional variables. Gelman and Loken (2014) characterize such variation as resembling a "garden of forking paths", and Silberzahn and Uhlman (2015; see also Simmons, Nelson, & Simonsohn, 2011) have demonstrated the consequences of variation in analytic approach in relation to differences in the size and direction of the effects that can be estimated. Our approach to analyzing the psycholinguistic effects on word recognition therefore, firstly, assimilated critical alternative steps employed in previous studies. Secondly, we examined the impact on our findings of variation in analytic choices by comparing results across critical alternative permutations in analysis steps. We share our data and analysis code to enable readers to review our choices or to examine alternative approaches.

Method

Data

We gathered reaction time data from previous large-scale studies that had examined word recognition in Spanish using the word naming (Davies et al., 2013) and lexical decision (González-Nosti et al., 2014) tasks. Davies et al. (2013) recorded word naming reaction times from 25 monolingual speakers of Spanish using randomized lists of 2,765 words. Their stimuli set included all nouns, verbs and adjectives between 3-10 letters-long, excluding compounds, from the LEXESP (Sebastián-Gallés, Carreiras, Cuetos, & Martí, 2000) database, which is one of the most used psycholinguistic databases in Spanish. González-Nosti et al. (2014) obtained lexical decision reaction times for the same words from a group of 36 participants. These data were combined with valence and arousal values gathered through Emofinder (Guasch, Padrón, Haro, Ferré, & Fraga, 2017), a web-based search engine for Spanish word properties from different normative databases (Ferré, Guasch, Moldovan, & Sánchez-Casas, 2012; Guasch, Ferré, & Fraga, 2016; Hinojosa et al., 2016; Redondo, Fraga, Comesaña, & Perea, 2005; Redondo, Fraga, Padrón, & Comesaña, 2007; Stadthagen-Gonzalez, Imbault, Pérez Sánchez, & Brysbaert, 2017), resulting in lexical decision and word naming data for a set of 2,555 words. Affective norms were gathered using nine-point scales for valence and arousal dimensions by means of the self-assessment manikin standard method (Bradley & Lang, 1994), a non-verbal pictorial assessment technique that allows direct measurement of these dimensions using simple non-verbal icons to depict various points along each of them. When data for a given word were available in various databases we used averaged values.

In addition, we compiled lexical characteristics known to influence word recognition including: word length measured as number of letters, phonemes and syllables; initial phoneme; written subtitle-based lexical frequency, as CD or as word form occurrence count; mean positional bigram frequency; orthographic and phonological neighbourhood size (N, Coltheart, Davelaar, Jonasson, & Besner, 1977) as well as averaged Orthographic Levenshtein Distance (OLD, see Yarkoni, Balota, & Yap, 2008) measures of lexical similarity neighbourhoods; and subjective ratings of familiarity, imageability, concreteness and AoA. All the values were gathered from the EsPAL database (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013) except for the AoA data which were obtained by Davies et al. (Davies et al., 2013). Lexical frequency values represent per-million occurrences from a 462,611,693 token data-set constructed from movie subtitles in Spanish. CD refers to the percentage of movies in which a given word appeared in the corpus, out of a total of 98,339 distinct movies. Bigram frequency and neighbourhood values were taken from the same database. Familiarity and imageability ratings correspond to averaged scores obtained with seven-point scales from at least 30 participants, except for the AoA data which correspond to averaged scores from 25 informants. A summary of the normative values of the psycholinguistic variables is given in Table 1. Note that the distribution of valence and arousal values in our stimuli sample is substantially similar to that obtained in previous large-scale normative studies conducted in Spanish (Stadthagen-Gonzalez et al., 2017).

(Table 1 about here)

Results

We begin by reporting steps taken to clean the data-set for analysis, removing outliers and missing values. We then report the correlations between psycholinguistic variables and the steps taken to reduce the problem of multicollinearity indicated by the correlations. We firstly report an analysis of the combined cross-task data-set. We then report the results of separate task-specific analyses of the lexical decision and word naming data, in a step analogous to simple effects analyses, designed to aid the interpretation of interactions between the effects of task and of the psycholinguistic variables.

Data treatment

We analysed trial-level data corresponding to the latencies of correct responses made by 61 participants to the 2,555 stimulus words for which we had complete critical psycholinguistic variable information. After excluding errors, outlier fast responses (responses associated with RT < 200ms) or responses to words for which data were missing on one or more psycholinguistic variables, we had a data-set of 60,690 word naming latencies and 79,616 lexical decision latencies for the task-specific analyses. A total of 140,306 observations was available for the primary cross-task analysis.

Preparation of predictor variables

Correlations between predictor variables must be examined because of the potential problem of multicollinearity. This problem arises in a linear model or, by extension, in a linear mixed-effects model, when the information associated with predictors overlaps, as indicated by high pairwise correlations ($r > .8$, according to a commonly used threshold) or condition numbers (kappa > 12, according to another common threshold; compare Baayen, Davidson, & Bates, 2008; Cohen, Cohen, West, & Aiken, 2003).

Table 2 presents the pairwise correlations between critical psycholinguistic variables for the stimulus words in our data-set.

(Table 2, about here)

It can be seen that there are correlations $r >= .7$ for a number of pairs of variables, including correlations between length (letters, phonemes, syllables), and orthographic similarity measures (orthographic neighbourhood size, phonological neighbourhood size, and orthographic Levenshtein distance), as well as between the different measures of frequency (word form frequency and CD), and semantic content (imageability and concreteness). These correlations or, rather, the multicollinearity they indicate, would if ignored pose the risk that analyses would not be capable of estimating the unique contributions to outcome variance of separate predictor variables, or would estimate effects that would not be stable between different samples (Cohen et al., 2003). Therefore we took the following steps to address the multicollinearity, prior to conducting our formal analyses: (1.) we combined the length measures, number of letters and number of phonemes, by averaging them together to create a new variable, “length”; (2.) we combined the orthographic and phonological neighbourhood size measures by averaging them together to create another new variable, “N-size”; (3.) we selected the CD measure of frequency for use in the analyses, given recent findings (Adelman, Brown, & Quesada, 2006; Brysbaert & New, 2009) indicating its superior performance compared to word form frequency in explaining variance in reading performance; (4.) between concreteness and imageability, we selected imageability as a measure of semantic content for inclusion as a model predictor, given its common use in previous analyses of large data samples (e.g. Cortese & Khanna, 2007); and (5.) we standardized all continuous numeric predictor variables. We include the aggregated length and N-size variables in the Table 2 correlations, for information.

We examined whether these choices influenced our results. We checked if the choice of frequency, orthographic similarity or semantic measure affected the estimates of emotion effects. We fitted models: (1.) using either the CD or the word form frequency measure, but not both, in separate analyses; (2.) using either the aggregated N-size or the OLD measures of neighbourhood similarity (but not both) in separate analyses; and (3.) using either the imageability or concreteness measures (but not both) in separate analyses. We found that alternation in the choice of frequency, orthographic similarity or semantic measures did not substantially influence the size or direction of the estimates of the valence or arousal effects. The interested reader is referred to the Supplementary Materials for detailed information.

Standardizing continuous numeric predictors removes non-essential collinearity due to scaling (Cohen et al., 2003) and it is critical for the estimation of interaction or curvilinear effects because lower- and higher-order terms are collinear if numeric predictor variables are not first rescaled to center on zero. Although Vinson et al. (2014) and Kuperman et al. (2014) chose to center their numeric predictors on mean values, we preferred to standardize predictors because transforming the variables to the same scale allowed straightforward comparison of effects.

Construction of categorical valence variables

We followed previous authors (Estes & Adelman, 2008a; Vinson et al., 2014) in constructing categorical valence predictor variables: (1.) a variable coding for whether word valence was positive or negative, termed positive-negative valence; (2.) a variable coding for whether word valence was emotional or neutral, termed emotional-neutral valence. We constructed categorical positive-negative and emotional-neutral valence terms for the cross-task (naming and decision), the word naming, and the lexical decision data-sets.

In our cross-task data-set, raw valence ratings varied from 1.2 to 8.7, with a mean (SD) of 5.3 (1.4); in the valence ratings studies (e.g. Stadthagen-Gonzalez et a., 2017), ratings had been elicited for a scale ranging from 1 (unhappy) to 9 (happy) via 5 (neutral, neither happy nor sad). To create the positive-negative factor, we divided the data by coding words with rated valence < 5 as negative, and words with rated valence >= 5 as positive. This division categorised 96,475 observation as responses to positive words and 43,831 observations as responses to negative words. To create the emotional-neutral factor, following Vinson et al. (2014), we divided the data at lower and upper bounds, respectively, of the neutral valence rating value of 5 plus or minus 1.5, categorizing 41,035 observations as responses to emotional words (valence < 3.5, or valence > 6.5) and 99,271 observations as responses to neutral words (valence >= 3.5, or valence <= 6.5). The distributions of observations in relation to valence values are illustrated in the barchart plots shown in Figure 1.

(Figure 1, about here)

In the word naming data-set, the same categorization scheme resulted in the classification of 19,326 observations as concerning responses to negative words, 41,364 concerning positive words, 17,478 concerning emotional words, and 43,212 concerning neutral words. In the lexical decision data-set, the same scheme resulted in the classification of 24,505 observations as concerning negative words, 55,111 concerning positive words, 23,557 concerning emotional words, and 56,059 concerning neutral words. It can be noted that we used linear mixed-effects models to analyse latencies, and that such models are robust to imbalances in numbers of observations.

Cross-task analysis

We examined the latencies of correct responses to words in both the lexical decision and naming tasks, in a cross-task analysis, fitting linear mixed-effects models to estimate effects using the lme4 package version 1.1-14 (Bates et al., 2017) in R version 3.4.2 (R development core team, 2017). All predictors were entered simultaneously.

We report the results of analyses of the effects of psycholinguistic variables on reading response RT but it is common practice to transform the outcome variable to log10(RT) to ameliorate skew in the distribution of latencies. We checked if the choice of outcome variable transformation made any difference to our results. We therefore repeated the final models (described later), for each valence measure, for the cross-task and for the task-specific lexical decision and word naming data-sets. To anticipate, we found that the significance and, more critically, the direction and the relative size of psycholinguistic effects were replicated in alternate log10(RT) or -1/RT compared to the RT models (see Supplementary Materials).

Following Vinson et al. (2014) and Kuperman et al. (2014), we began our analyses by specifying a baseline model. Because our primary focus was on the cross-task data, the baseline model had to incorporate effects due to task, psycholinguistic variables, and interactions between the effects of task and of the psycholinguistic variables. We report, firstly, our observations from the process of specifying an adequate baseline model. We report, then, the results from subsequent analysis steps conducted to evaluate the contribution of valence and arousal to our account of the variance of response latencies in reading in Spanish. In these steps, as we explain following, we examined whether the addition of the affective variables was warranted by improved model fit to data. We evaluated model fit using information criterion (e.g. Burnham & Anderson, 2004) and Likelihood Ratio Test (LRT, see, e.g., Baayen, 2008; Baayen et al., 2008; Pinheiro & Bates, 2000) comparisons.

The research questions investigated in our analysis were:

1. Does valence have an effect, and, if it does, what is the best measure for capturing the effect?

2. Does arousal have an effect?

3. Do the effects of valence and arousal interact?

4. Do the effects of valence and arousal interact with the effect of task?

5. Are the effects of valence or arousal modulated by the influence of word frequency in interactions between frequency and emotion effects?

We structure our results reporting correspondingly. We addressed each question in turn, examining whether the addition to our model of a term corresponding to the effect of interest, for example, of valence, improved model fit to data. We did this separately for each valence measure. We compared model fits for models with vs. without the effect of interest using the LRT. In addition, we report the results of t-tests of the coefficient estimates for each effect of interest, employing Satterthwaite approximations to denominator degrees of freedom (p-values were derived with the lmerTest package, Kuznetsova, Brockhoff, & Bojesen Christensen, 2016). At present, different methods are commonly used to examine the utility of hypothesised effects or, equivalently, the relative utility of alternative models (with or without the effects). It was reassuring to find, as we shall report, that, concerning the utility of hypothesized effects, the indications from the model fit comparisons and the hypothesis tests coincided.

We began by comparing models that varied in fixed effects, corresponding to psycholinguistic effects, but were consistent in the inclusion of random effects due to unexplained differences between sampled participants or between items in intercepts (random intercepts). We fitted terms corresponding to all effects of interest, ultimately. Model comparisons are reported as tests of the utility (for model fit to data) of the inclusion of terms corresponding to these effects, not as the basis for including the terms. We conclude this section by presenting a summary of the full model including all effects of interest.

We then examined the utility for model fit of adding random effects due to differences between participants in the slopes of the (within-subjects) psycholinguistic effects or between items in the slope of the (within-items) task effect. Matuschek, Kliegl, Vasishth, Baayen, & Bates (2017; see, also, Barr, Levy, Scheepers, & Tily, 2013) argue that an adequate balance between the relative sensitivity and conservatism of an analysis can be found by examining whether the inclusion of random effects terms improves model fit to data. We did this by fitting a model with all fixed effects of interest and all random effects permitted by the study design, then removing random effects terms until we arrived at a model with a parsimonious random effects structure (as complex as appeared defensible, given the data). We present the cross-task and task-specific models, finally, with this random effects structure. We note that the results of comparisons between models varying in fixed effects did not differ if we specified only random intercepts (as presented) or instead the more complex, but parsimonious, random effects structure identified in our later checks (code and results for both sets of models are presented in the Supplementary Materials).

Baseline task x psycholinguistic effects models

We began by fitting a baseline model. We examined four candidate baseline models. In all models, we included terms corresponding to the following key variables: a word initial phoneme coding variable; a variable coding for reading task (word naming vs. lexical decision); the CD frequency measure; AoA; familiarity; imageability; the aggregate word length measure (the average of length in letters and in phonemes); word length in syllables; and the aggregate neighbourhood size measure (the average of orthographic and phonological neighbourhood size). The models incorporated fixed effects terms corresponding to the effects of the psycholinguistic variables and, at this stage, random effects terms corresponding to variance due to unexplained differences between sampled participants or words in intercepts (random intercepts).

In model 1, the fixed effects included the effects of task, phoneme, and the critical psycholinguistic variables except valence or arousal. No interactions were included. All numeric predictor variables were specified as terms corresponding to linear effects.

Previous observations have indicated that the effects of some psycholinguistic variables on response latencies, like the effects of word frequency or length, may be curvilinear in English (e.g. Baayen, Feldman, & Schreuder, 2006; New, Ferrand, Pallier, & Brysbaert, 2006) and in Spanish (Davies et al., 2013). Studies of the impact of emotion on word recognition have included reports suggesting non-linear effects of valence (Estes & Adelman, 2008b; Kousta et al., 2009; but see Kuperman et al., 2014). We therefore examined, in model 2, if curvilinearity should be allowed for the effects of any of the psycholinguistic variables (excluding task, initial phoneme, and length in syllables). A comparison of model 1 and model 2, where, in the latter, all numeric predictors were fitted to latencies using restricted cubic splines (with up to k = 3 knots), indicated that the model allowing for curvilinearity fit the data better (LRT comparison, $\chi^2(7) = 357.7$, $p < .001$). (See the Supplementary Materials for summaries of all models.)

We examined curvilinearity in the effects of psycholinguistic variables using restricted cubic splines (e.g. Baayen, 2008; Davies et al., 2013) but checked if the influence of emotion effects was the same in analyses using polynomial (up to quadratic) terms. This is because, while Vinson et al. (2014) preferred to fit polynomial terms to estimate potentially non-monotonic emotion effects, Kuperman et al. (2014) preferred to fit restricted cubic splines (in Generalized Additive Models, GAMs; see also Kousta et al., 2009). The results of the check analyses indicated that the size and direction of critical effects estimates were not substantially different if polynomials or splines were used to capture curvilinearity in effects (see Supplementary Materials).

The effects of CD, AoA, familiarity, length and N-size were associated with significant curvilinear components (model 2, t-tests on corresponding coefficients, $p < .05$). Task, imageability and length in syllables were associated with significant linear effects only (model 1, all t-tests on corresponding coefficients, $p < .01$). Bigram frequency was associated with a marginal linear effect (t-test, $p = .074$) and a non-significant curvilinear effect (t-test, $p > .10$). To fit the most parsimonious defensible baseline model, given our data, we specified the CD, AoA, familiarity, length and N-size effects as curvilinear, and specified all other effects as linear, in all further models. The simplified model (model 3) fit the data as well as (or not detectably different than) model 2 ($\chi^2(2) = .6$, $p = 0.758$).

In the final *baseline interactions* model, we specified the described linear or curvilinear psycholinguistic effects, plus the effect of task, and the effects of all possible two-way interactions between the effect of task and the effect of each of the psycholinguistic variables. An LRT comparison indicated that including interactions between task and psycholinguistic effects improved model fit ($\chi^2(36) = 1953.2$, $p < .001$). Task differences significantly modulated the effects of CD, AoA, familiarity and length (t-tests of task by psycholinguistic interaction effects, all ps < .05). A summary of the model is given in Table 3. Response latencies decreased with increasing CD frequency, familiarity and neighbourhood size though the impact of each effect diminished for higher variable values. Latencies increased with unit increase in AoA, word length measured in syllables or with increasing bigram frequency. The effect of the aggregate length variable was curvilinear such that latencies decreased slightly for increasing length, at first, and then increased with increasing length for longer words. The effect of AoA was curvilinear such that the impact of AoA was greater for later-acquired words. Each psycholinguistic effect was more pronounced in lexical decision than in naming.

(Table 3, about here)

In the following sequence of analyses, to address each of the research questions, over a series of models, we successively added terms corresponding to the effects of interest. The addition of terms was cumulative so that later models included all terms specified in earlier models. At each step, we first added the term as a “main effect”, that is, ignoring any potential interaction with task. We then added the term as both the lower-order component and as the task x effect interaction. Stepping up the increments in model complexity in this way allowed us to evaluate whether the addition of the effect was warranted with or without allowing for the modulation of the effect by task differences. In the following, we report the results of LRT comparisons of the model with versus without the additional term, added as a “main effect”, then of the model with the additional “main effect” versus the model with the additional term added as a “main effect” and as a task by effect interaction.

Test of the effect of valence

Our first research question was: Does valence have an effect, and, if it does, what is the best measure for capturing the effect? To answer the first part of that question, we added the main effect of valence to the *baseline interactions* model, in separate models, one model for each valence measure: graded valence, categorical positive-negative valence, or categorical emotional-neutral valence. By comparing the fit of the *baseline interactions* model to the fit of the model including a valence measure, we evaluated if the addition of valence was useful in accounting for observed variance in Spanish reading. Likelihood ratio test comparisons showed that the addition of valence was justified by significantly improved model fit with the addition of graded valence ($\chi^2(1) = 25.3$, $p < .001$) or positive-negative valence ($\chi^2(1) = 27.7$, $p < .001$) but not of emotional-neutral valence ($\chi^2(1) = .8$, $p = .363$). (Allowing the effect of graded valence to be curvilinear did not improve model fit, $\chi^2(2) = 3.5$, $p = .178$.)

In the second step, we examined whether the valence effect was moderated by the effect of task differences. We compared the fit of a model including the *baseline interactions* terms plus valence to the fit of a model with the same *baseline interactions* terms plus valence and the task by valence interaction. LRT comparisons showed that the addition of a task by valence interaction was justified by significantly improved model fit for the models including graded valence ($\chi^2(1) = 16.3$, $p < .001$), positive-negative valence ($\chi^2(1) = 7.2$, $p = .007$) but not emotional-neutral valence ($\chi^2(1) = .3$, $p = 599$). We term these models the *baseline-plus-valence* models.

In evaluating competing models using information criteria, we are concerned with the relative, not the absolute, AIC values. Criteria with lower values (smaller values if positive, closer to negative infinity if negative) indicate that models have higher likelihood (log likelihood, scaled by multiplication by -2), that they incorporate effects estimates that allow better prediction of observed latencies, minimising Kullback-Leibler information loss (Burnham & Anderson, 2004; McElreath, 2016). The graded rated valence or categorical positive-negative valence models better approximated Spanish word recognition performance data than did the baseline or categorical emotional-neutral models. A summary of the *baseline-plus-valence* models is shown in Table 3. The influence of rated valence on word recognition RTs, and its greater prominence in lexical decision, is clearly illustrated in Figure 2.

(Figure 2, about here)

Test of the effect of arousal

Our second research question was: Does arousal have an effect, either as a main effect or in a task by arousal interaction? We answered this question by comparing the fit of the *baseline-plus-valence* model with the fit of models including the same terms as the *baseline-plus-valence* model plus, successively, the main effect of arousal, and the effects of arousal and the task by arousal interaction.

For models representing valence as a graded measure, likelihood ratio tests indicated that, compared to the *baseline-plus-valence* model, the addition of arousal did not significantly improve model fit if added as a main effect ($\chi^2(1) = 2.0$, $p = .158$). Compared to the *baseline-plus-valence* and arousal model, the addition of arousal as main and task by arousal interaction effects did not improve fit ($\chi^2(1) = 1.3$, $p = .258$). For models incorporating valence as a positive-negative factor, the same pattern of results was found. LRTs indicated that, compared to the *baseline-plus-valence* model, the addition of arousal did not significantly improve model fit if added as a main effect ($\chi^2(1) = 2.2$, $p = .140$). Compared to the baseline-plus-valence and arousal model, the addition of arousal as main and interaction effects did not improve fit ($\chi^2(1) = .2$, $p = .622$). For models incorporating valence as an emotional-neutral factor, likewise, the addition of arousal did not significantly improve model fit if added as a main effect ($\chi^2(1) = .3$, $p = .603$) or as main and interaction effects ($\chi^2(1) = .7$, $p = .403$).

Adding arousal, as a main effect, or as main and task by arousal interaction effects, did not improve the fit to data, compared to models incorporating baseline and valence terms. The limited impact of arousal in either task is clearly illustrated in Figure 3. We termed the models including the arousal and task by arousal interaction effects the *baseline-plus-affect model*.

(Figure 3, about here)

Test of the interaction between the effects of valence and arousal

Our third research question was: Do the effects of valence and arousal interact? To answer this question, we compared the fit of the *baseline-plus-affect* model with the fit of models including the same terms plus, successively, the valence by arousal interaction effect, and the valence by arousal as well as the task by valence by arousal interaction effects.

For models representing valence as a graded measure, LRTs indicated that, compared to the *baseline-plus-affect* model, the addition of the valence by arousal interaction did not significantly improve model fit ($\chi^2(1) = .2$, $p = .695$). However, compared to a baseline-plus-affect model that also included a valence by arousal interaction, further adding the task by valence by arousal interaction effect did improve model fit ($\chi^2(1) = 5.4$, $p = .020$).

For models representing valence as a positive-negative factor, a different pattern of results was found. LRTs indicated that, compared to the *baseline-plus-affect* model, the addition of the valence by arousal interaction did not significantly improve model fit ($\chi^2(1) = 2.4$, $p = .118$). Nor, if added as valence by arousal and task by valence by arousal interaction effects, did that addition improve model fit to data ($\chi^2(1) = .1$, $p = .796$).

For models representing valence as an emotional-neutral factor, likewise, LRTs indicated that, compared to the *baseline-plus-affect* model, the addition of the valence by arousal interaction did not significantly improve model fit ($\chi^2(1) = .0$, $p = .859$). Nor, if added as valence by arousal and task by valence by arousal interaction effects, did that addition improve model fit ($\chi^2(1) = .4$, $p = .550$).

In sum, a potential interaction between the effects of valence and arousal was apparent but it was expressed in different ways depending on the measure of valence incorporated in the model. For models in which valence was represented using a positive-negative or an emotional-neutral categorical factor, the inclusion of the interaction between valence and arousal did not improve model fit. For models in which valence was represented as a graded valence measure, the impact of the valence by arousal interaction appeared to be constrained by task differences. We termed the models including the valence, arousal, valence by arousal and task interaction effects the *baseline-plus-affect-interaction model*.

Evaluating the modulation of valence and arousal effects by task differences

Our fourth research question was: Do the effects of valence and arousal interact with the effect of task? We addressed this question by estimating potential interactions between the effect of task and the effects associated with critical psycholinguistic variables. Our observations indicated, as seen, that psycholinguistic effects are modulated by task differences, with variation in the size and shape of the effects of frequency, AoA, familiarity and valence in lexical decision compared to word naming. These differences were explored in the task-specific analyses reported in a following section.

Evaluating the modulation of valence and arousal effects by frequency

Our final research question was: Are the effects of valence or arousal modulated by the influence of word frequency in interactions between the frequency and emotion effects? To answer this question, we compared the fit of the *baseline-plus-affect-interaction* model with the fit of models including the same terms plus, successively, both CD frequency by valence and CD frequency by arousal interaction effects, and models including these interactions as well as terms corresponding to the modulation of the interactions by task differences.

We found that the addition of interactions between the effects of frequency and the effects of valence or arousal did not improve model fit to data, irrespective of the valence measure, whether comparing the fit of *baseline-plus-affect-interaction* models to models with the same terms plus just the frequency by valence and frequency by arousal interaction effects (graded valence, $\chi^2(4) = 3.6$, $p = .470$; positive-negative valence, $\chi^2(4) = 2.4$, $p = .662$; emotional-neutral valence, $\chi^2(4) = 3.7$, $p = .445$) or comparing the fit of the latter models to models with the same terms plus the task by CD frequency by valence or task by CD frequency by arousal interactions (for graded valence, $\chi^2(4) = 5.1$, $p = .278$; for positive-negative valence, $\chi^2(4) = 2.8$, $p = .597$). For models including emotional-neutral valence, the addition of terms corresponding to interactions between task, frequency and valence or arousal together did improve model fit to data (emotional-neutral valence, $\chi^2(4) = 14.5$, $p = .006$).

We termed the models including the frequency by valence, frequency by arousal, and corresponding task interaction effects, the *baseline-plus-affect-frequency-interaction model*. Our conclusion is that frequency did not significantly modulate the effects of valence or arousal except where, for models including valence coded as an emotional-neutral factor, the main effect of categorical valence was not, itself, reliably detected as a main effect.

Comparison of model fit across different measures of valence

We found that a comparison of information criteria statistics indicated that models representing the valence effect with a graded valence or a categorical positive-negative measure fit the data better than either a baseline model not including a valence measure, or a model including the categorical emotional-neutral measure (see Table 3). A comparison of information criteria statistics showed that the ranking of the relative utility of models incorporating different valence measures remained the same after models had been expanded to include effects associated with arousal. We evaluated, for each valence measure, the models including the baseline effects plus the effects of valence, arousal, and the valence by arousal interaction, as well as the interactions between these effects and the effects of frequency and task. We found that information criteria values indicated better fit to data for the model representing valence as a graded measure (AIC = 1666762) compared to the model representing valence as a positive-negative measure (AIC = 1666776), while both graded and categorical positive-negative valence models were better fits than a model including the categorical emotional-neutral variable as the valence measure (AIC = 1666800).

Comparison of model fit when only valence and arousal are entered as predictors

Readers may ask if the observed utility of the valence or arousal effects would appear to be different if only valence or arousal were entered as predictors or if the order in which valence or arousal were entered was varied. (We thank an anonymous reviewer for this suggestion.) We should note that the psycholinguistic variables were entered simultaneously in each of the reported models. However, entering valence and arousal as the only fixed effects (alongside random effects due to between-subjects or between-items differences in intercepts) allowed us to estimate a further measure of relative fit, to bring converging evidence to bear on the question of how valence or arousal influenced word recognition latencies in Spanish.

We fitted models of the cross-task data-set response latencies, separately for each valence measure, in which we specified as fixed effects: valence alone; arousal alone; valence and arousal as additive main effects; valence, arousal, and the interaction between valence and arousal. For each model, we calculated the marginal R2_m, the variance explained by the fixed effect(-s) as a proportion of the sum of all the variance components, including the fixed effects as well as the random effects and the residuals (with R2_m calculated using the MuMIn package, version 1.15.6, Barton, 2016; Johnson, 2014; Nakagawa & Schielzeth, 2013). We found that .3% of variance was explained by the graded valence effect, compared to .2% explained by the categorical positive-negative valence effect, and .1% by the emotional-neutral effect. The valence effect is small but, consistent with the results reported in the foregoing, we found that it was best captured by the graded valence measure. We estimated that .02% of variance was found to be explained by the effect of arousal, entered as a fixed effect on its own. We calculated that a valence by arousal interaction explained, at best, .05% of variance. Arousal, or the valence by arousal interaction, thus added little to our account.

Random effects

The models reported to this point have incorporated fixed effects due to the psycholinguistic variables, and random effects due to the differences between participants or between stimulus words in intercepts. We did not, up to this point, include variance terms corresponding to random differences between participants in the slopes of the within-subjects psycholinguistic effects, or between words in the slope of the within-items task effect (random slopes). This was a potentially important omission. Not including random slopes has been argued to increase the Type I error rate (Barr, Levy, Scheepers, & Tily, 2013). However, Matuschek et al. (2017) have demonstrated that some caution is required because a loss of sensitivity can be associated with including random effects not warranted by the data.

We fit a model with the same fixed effects as the final *baseline-plus-affect-frequency-interaction* models, with both random intercepts and random slopes. We excluded terms corresponding to covariances between random intercepts and random slopes, to random differences between subjects in the coefficients of the curvilinear components of the psycholinguistic effects, and to random differences between subjects in the coefficients of the word initial effect, because models including those terms did not converge. We fit a model (1) including the critical fixed effects plus random effects corresponding to random differences between subjects or items in intercepts, random differences between subjects in the slopes of the linear (main and interaction) psycholinguistic effects, and random differences between words in the slope of the task effect. This model fit the data approximately as well as a model (2) excluding terms corresponding to random differences in the slopes of interactions ($\chi^2(3) = .7$, $p = .863$), the latter fit the data better than a model (3) excluding a term corresponding to random differences between items in the task effect ($\chi^2(1) = 637.1$, $p < .001$), while the last fit the data better than a model with just random intercepts ($\chi^2(10) = 1047.5$, $p < .001$).

The model comparisons indicate that model (1) represents the best account of the Spanish reading data, including fixed effects terms that test theoretically critical questions, as well as a random effects structure that is as complex as necessary to fit the data, capturing random differences between subjects or items in intercepts and slopes. We present a summary of the final model in Table 4. We show effects estimated with a model including the graded valence measure only because that measure was found to be most useful in capturing the influence of affect.

(Table 4, about here)

Task-specific analyses

The results of the cross-task analysis show that the effects of critical psycholinguistic variables are moderated by the influence of differences between reading tasks.The psycholinguistic effects were consistent in direction but smaller in size in the word naming compared to the lexical decision task. This pattern matches previous observations in English and other languages (e.g. Burani, Arduino, & Barca, 2007 in Italian; Balota et al., 2004; Cortese & Khanna, 2007 in English). However, we observed, for the first time, interactions between curvilinear psycholinguistic effects and task differences within the same analysis. To clarify how task differences moderated the psycholinguistic effects, we fitted the same model to the lexical decision and word naming data-sets. For each task-specific analysis, we estimated the effects of the same linear and curvilinear psycholinguistic effects, including the effects of valence, arousal, and the valence by arousal interaction. For each analysis, we included the same random effects structure as we identified for the cross-task final model, minus the random effect of items on the slope of the task effect. We fit models using each different valence measure though we report in detail only the results for the models representing valence as a graded measure.

In the task-specific model of lexical decision latencies, we found significant curvilinear effects of frequency, AoA, familiarity, length and neighbourhood size, along with linear effects of bigram frequency and valence (represented as a graded measure). In the model of word naming latencies, we found significant curvilinear effects of frequency, AoA, length, and neighbourhood size, along with linear effects of familiarity, word length in syllables, and the valence x arousal interaction. In Table 5 we present summaries of mixed-effects models of the task-specific data.

(Table 5, about here)

It can be seen that the frequency effect in lexical decision was, on average, negative going (task-specific estimate of the linear component of the frequency effect, coefficient = -60.3, SE = 3.7), with more frequent words associated with faster latencies. However, for the most frequent words, the frequency effect diminished considerably (estimate of the non-linear component of the frequency effect, coefficient = 150.8, SE = 10.1). The impact of task differences was to reduce this curvilinearity so that the slope of the negative linear component (estimate of the word naming frequency effect, coefficient = -19.5, SE = 2.6), and the slope of the positive curvilinear component (estimate of the non-linear component of the word naming frequency effect, coefficient = 58.2, SE = 8.2) were both less pronounced in naming.

The AoA effect in lexical decision was, on average, positive going (estimate of the linear component of the AoA effect, coefficient = 4.7, SE = 1.8) with later acquired words eliciting longer latencies, but for words that were even later acquired the AoA effect was greater (estimate of the non-linear component of the AoA effect, coefficient = 5.4, SE = 1.7). The impact of task differences was to reduce the slope of the positive linear component strongly (task-specific estimate of the linear component of the word naming AoA effect, coefficient = -.1, SE = 1.4), and to reduce the slope of the positive curvilinear component very weakly (task-specific estimate of the linear component of the word naming AoA effect, coefficient = 4.3, SE = 1.2), so that the AoA effect remained large among responses to later-acquired words in naming.

The familiarity effect in lexical decision was, on average, negative going (estimate of the linear component of the familiarity effect, coefficient = -14.6, SE = 1.4), with more familiar words associated with faster latencies, but for the most familiar words the familiarity effect was smaller (estimate of the non-linear component of the familiarity effect, coefficient = 7.2, SE = 1.5). The impact of task differences was to reduce the the slope of the negative linear component (estimate of the linear component of the word naming familiarity effect, coefficient = -4.1, SE = 1.1), about as much as the slope of the positive curvilinear component (estimate of the word naming non-linear component of the familiarity effect, coefficient = .9, SE = 1.0).

The length effect in lexical decision was, on average, weakly negative going (estimate of the linear component of the length effect, coefficient = -8.1, SE = 3.0), with longer words associated with slightly faster latencies, on average, but for the longest words the direction of the length effects reverses so that increasing length was associated with increasing latencies (estimate of the non-linear component of the length effect, coefficient = 20.3, SE = 2.9). The impact of task was to comparatively strongly reduce the slope of the negative linear component (estimate of the linear component of the word naming length effect, coefficient = -.8, SE = 2.2) and weakly reduce the slope of the positive curvilinear component (estimate of the linear component of the familiarity effect, coefficient = 8.4, SE = 2.0). In consequence, the length effect was relatively weak or null for shorter words, but stronger for longer words, in naming compared to lexical decision.

The bigram frequency effect in lexical decision was, on average, positive going (estimate of the bigram frequency effect, coefficient = 2.2, SE = .7), with words composed of more frequent bigrams eliciting slower responses. The impact of task differences was to almost eliminate the bigram frequency effect in naming compared to decision (estimate of the word naming bigram effect, coefficient = .7, SE = .5).

The valence effect in lexical decision was, on average, negative going, with words that were associated with higher (more positive) valence ratings associated with faster responses (estimate of the valence effect, coefficient = -3.4, SE = 1.5). The impact of task differences was to strongly reduce the valence effect in naming compared to lexical decision (estimate of the word naming valence effect, coefficient = .9, SE = 1.0).

The valence x arousal interaction effect in lexical decision was, on average, small and positive (estimate of the interaction effect, coefficient = .4, SE = .5), suggesting that the valence effect was slightly smaller for higher arousal words. In word naming, a contrasting pattern was apparent. The valence x arousal interaction effect in naming was small and negative (estimate of the word naming valence x arousal interaction effect, coefficient = -1.0, SE = .4), suggesting that the valence effect was slightly larger for higher arousal words.

In summary, the graded effect of valence was significant for lexical decision but not for word naming, though there was a trend suggesting an effect of valence in naming. Consistent with the full cross-task analysis, the task-specific results indicated larger effects in lexical decision than naming for variables typically associated with lexical or semantic processes, frequency, AoA, familiarity and, critically for our study, valence. While we do not report summaries of full models including categorical positive-negative or emotional-neutral valence measures, we note that positive-negative valence was associated with a significant effect in lexical decision but not naming, while emotional-neutral valence was not associated with a significant effect in either task.

Discussion

We aimed to assess the impact of affective content on word recognition in Spanish. We explored the influence of valence on performance in the lexical decision and naming tasks. In addition, we examined the effects of arousal, and of the interaction between valence and arousal. Our analyses revealed a significant effect of valence on word recognition, as emotional negativity delayed the participants' responses in lexical decision and to some extent in naming. These results support theoretical accounts of emotional stimulus evaluation in which negative affective values delay reaction times. They contradict accounts in which emotional (negative or positive valence) words elicit faster responses than neutral words. Our results add to current understanding by showing that the valence effect is larger in lexical decision than in word naming, consistent with an interpretation of the effect as located in semantic processing. They demonstrate the importance of the valence effect in reading in Spanish, a language with a transparent orthography, significantly extending the apparent scope of the influence of emotion on reading. We discuss the theoretical implications of our observations in the following.

Pratto and John's Automatic Vigilance model of emotion (1991) proposes that undesirable stimuli grab more attention than desirable ones. According to this model, the effect occurs during automatic monitoring of the environment (i.e. monitoring without the perceiver's intent), functioning as a signal of potential danger. Based on their observations, in which undesirable stimuli retained more attention than positive ones, regardless of their relative valence, Pratto and John (1991) proposed that the valence effect was categorical in nature (see also Estes & Adelman, 2008a, 2008b). However, more recently, Kuperman et al. (2014) reported graded linear valence effects, leading those authors to argue that the automatic vigilance process is graded. The retention of attention is proportional to the negative affective value of the stimulus. In our study, word recognition reaction times were better explained by models including a graded (positive-negative) version of the valence measure, adding to the empirical support for a graded view of automatic vigilance.

We observed that the fit of models incorporating graded versus categorical positive-negative valence measures were not greatly different. It would be appropriate, then, to acknowledge that the impact of valence on word recognition can be captured by graded or by categorical measures of positive-negative valence differences. In our analysis, the graded measure of valence was found to be a more sensitive means of estimating the influence of valence on word recognition latencies. This is consistent with the greater information associated with a graded compared to a categorical measure of psychological variation (Cohen, 1983). However, as seen, the effect of valence is relatively small, the variance explained by the fixed effect of the graded valence effect was about .3% (the marginal R^2_m; Nakagawa & Schielzeth, 2013) while for the positive-negative valence it was about .2%. This means that the graded valence effect may be readily detected in the long run. For comparison, we note that Adelman and Estes (2008b) reported a valence effect of .8% (lexical decision) or .6% (word naming) but remind their readers that the theoretical importance of the effect is nevertheless large. The difference between the size of the effect of valence in Spanish compared to English is interesting but should be the topic of future research.

The important point is that the difference between speed of response to positive compared to negative valence words was reliably detected in our analyses of Spanish reading behaviour. The balance of evidence is that the difference between response latencies for positive and negative words is graded. Equally, our results are clearly in conflict with findings from previous studies that indicated emotional facilitation during word recognition. Both Kousta et al. (2009) and Vinson et al. (2014) observed inverted-

U effects of valence with faster reaction times for negatively and positively valenced stimuli compared to neutral words. Their observations favored the Motivated Attention and Affective States model (Lang et al., 1997), according to which both positive and negative affective stimuli are more likely to draw attention than neutral stimuli because emotional stimuli are motivationally relevant. In our analyses, an emotional-neutral coding of valence failed to capture the impact of valence effect that was otherwise evident (using positive-negative measures) across an extensive set of analyses.

Our sample of Spanish reading behaviour, while substantial, did not indicate an effect of arousal, either. We did not observe an effect of arousal, overall, or in an interaction, moderated by task. The lack of an effect of arousal in our analyses is congruent with the results obtained by Kousta et al. (2009) and Vinson et al. (2014), who also observed specific effects of valence but no influence of arousal on word recognition. Our results, however, contrast with the effect of arousal identified in the large-scale study conducted by Kuperman et al. (2014). Although the inclusion of arousal in our analysis did not improve the fit of our model to word recognition data, the fact that our word sample was smaller than that analysed by Kuperman et al. (2014) does not allow us to rule out the existence of a small arousal effect.

We analyzed if the effects of valence and arousal interacted. A valence by arousal interaction was not reliably detected in previous studies (Kousta et al., 2009; Vinson et al., 2014; Estes and Adelman, 2008a; Kuperman et al. 2014). In contrast, in our study, the effect of valence on word naming latencies was modulated by that of arousal, with stronger valence effects for higher arousal words. A similar interaction was not observed in lexical decision. Our observations thus suggest that an influence due to arousal may be found, to the extent that the valence effect is slightly different for words varying in arousal, in Spanish, but not to the extent that the impact of arousal is, on its own, detectable for our data. The variation in the valence by arousal interaction, depending on which valence measure is included in the model, suggests that the influence of arousal merits further investigation but will be difficult to characterize with confidence.

Critically, we tested whether the effect of valence was different in response to different task demands. The inclusion of the interaction between task (lexical decision or word naming) and graded valence in the analysis significantly improved model fit. Valence affected lexical decision responses more strongly than word naming responses. This finding extends previous observations in English (Estes & Adelman, 2008a; Kuperman et al., 2014) in which the valence effect was compared between tasks qualitatively but not formally. Importantly, our cross-task analysis allowed a direct estimate of the moderation of the valence effect by task differences as the effect of the task by valence interaction. (See Nieuwenhuis, Forstmann, & Wagenmakers, 2011, for a discussion of the inferential problems inherent in comparing effects in different data sub-sets when interactions are hypothesized but not formally tested).

The comparison between the results of lexical decision and word naming tasks is of interest because it could help to clarify the nature of the effect. Previous research (Balota et al., 2004; Cortese & Khanna, 2007) has indicated that semantic effects tend to be larger or easier to detect in lexical decision than in word naming because lexical decision response preparation is more reliant or draws more readily on such information (although see Plaut, 1997; Seidenberg & McClelland, 1989, for alternative interpretations). We do not think that the greater size of psycholinguistic effects, like the effect of valence, in lexical decision, is due to the fact that responses were slower than in word naming (as is usually observed). The average speed of response varies at random between subjects within and between tasks, as well as between items within tasks. Our use of linear mixed-effects models allowed us to include variance terms to account for such differences between sampled participants or words in response speed (while controlling for all other predictors). We thus observed the interaction between task and valence effects while taking into account differences in average speed of response. The fact that we observed stronger effects on lexical decision would, in our view, therefore support a semantic interpretation of the valence effect. According to this account, valence would join the group of variables argued to be related to semantic knowledge, like AoA, imageability or familiarity (Balota et al., 2004; Cortese & Khanna, 2007; Davies et al., 2013; Davies, Wilson, Cuetos, & Burani, 2014).

The observation of the valence and task by valence interaction effects in Spanish has significant implications for a language-general account of reading. Our findings demonstrate that emotional content affects reading in a transparent orthography. Granted that valence can be understood as a semantic effect, this contrasts with the account proposed by some researchers, that semantic information tends to influence word recognition more prominently where words are difficult to encode, as appears to be the case, in English, for low frequency irregular words (Plaut et al., 1996; Strain et al., 1995). It may well be that semantic information influences word recognition in English more widely across the vocabulary (as reported by Balota et al., 2004; Cortese & Khanna, 2007). Our results demonstrate with certainty that word recognition is richly influenced by semantic content when the words being read have regular pronunciations.

In sum, we did not observe a significant effect of arousal in word recognition in Spanish. Further studies should be conducted to ascertain whether the lack of a reliable arousal effect in our data is due to specific characteristics of our stimuli or it rather indicates differences between the influence of this variable on word recognition in deep and transparent orthographies.

In contrast, we identified a substantial effect of emotional valence on word recognition, with positive valence words eliciting faster reaction times than negative valence words. This finding provides empirical support to the Automatic Vigilance model of emotion (Pratto & John, 1991), according to which emotionally negative information slows down cognitive activity. In contrast, our data does not support the model of Motivated Attention and Affective States (Lang et al., 1997), which predicts faster reaction times for both positive and negative stimuli. Our data suggest that the observed effect is graded, such that the latency reduction associated with positive compared to negative valence is proportional to the positivity of the stimuli. This finding is inconsistent with the categorical effect for negative stimuli predicted by the original version of the Automatic Vigilance model. Critically, our observation that the influence of valence was stronger in lexical decision than naming indicates a semantic location for the effect. This has implications for theoretical accounts of the cognitive architecture of the reading system, and of the role of semantic information in reading performance in different languages. Our observation of a valence effect in Spanish, a language with a transparent orthography, shows that emotion influencess the recognition of words with regular pronunciations. These results mirror some of those reported previously in English, thus demonstrating important parallels in word recognition processes between orthographically shallow and deep languages.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision time. *Psychological Science*, *17*, 814–823.

Algom, D., Chajut, E., & Lev, S. (2004). A rational look at the emotional Stroop phenomenon: A generic slowdown, not a Stroop effect. *Journal of Experimental Psychology: General*, *133*, 323–338. https://doi.org/10.1037/0096-3445.133.3.323

Baayen, R. H. (2008). *Analyzing linguistic data: A Practical Introduction to Statistics using R*. Cambridge, U.K: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *53*, 496–512.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing : Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barton, K. (2016). MuMIn: Multi-model inference. Retrieved from https://cran.r-project.org/web/packages/MuMIn

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv:1506.04967*.

Bates, D., Maechler, M., Bolker, B., Walker, S., Bojesen Christensen, R. H., Singmann, H., … Green, P. (2017). lme4: Linear mixed-effects models using “Eigen” and S4. Retrieved from https://cran.r-project.org/web/packages/lme4

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*, 49–59.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Gainesville: University of Florida, Center for Research in Psychophysiology.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency-related, partly frequency-independent. *Visual Cognition*, *13*, 992–1011.

Brysbaert, M., & New, B. (2009). 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*, *41*, 977–990.

Burani, C., Arduino, L. S., & Barca, L. (2007). Frequency, not age of acquisition, affects Italian word naming. *European Journal of Cognitive Psychology*, *19*(6), 828–866.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference. Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304.

Chumbley, J. I., & Balota, D. A. (1984). A word’s meaning affects the decision in lexical decision. *Memory and Cognition*, *12*(6), 590–606. Journal Article.

Citron, F. M. M., Weekes, B. S., & Ferstl, E. C. (2013). Effects of valence and arousal on written word recognition: Time course and ERP correlates. *Neuroscience Letters*, *533*, 90–95.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Earlbaum Associates.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S.Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). Hillsdale,NJ: Lawrence Erlbaum Associates Inc.

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *The Quarterly Journal of Experimental Psychology*, *60*, 1072–1082.

Davies, R. A. I., Arnell, R., Birchenough, J. M. H., Grimmond, D., & Houlson, S. (2017). Reading Through the Life Span: Individual Differences in Psycholinguistic Effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 298–1338. https://doi.org/10.1037/xlm0000366

Davies, R. A. I., Barbón, A., & Cuetos, F. (2013). Lexical and semantic age-of-acquisition effects on word naming in Spanish. *Memory and Cognition*, *41*, 297–311. https://doi.org/10.3758/s13421-012-0263-8

Davies, R. A. I., Wilson, M., Cuetos, F., & Burani, C. (2014). Reading in Spanish and Italian: effects of age of acquisition in transparent orthographies? *The Quarterly Journal of Experimental Psychology*, *67*(9), 1808–1825.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal:

One-stop Shopping for Spanish Word Properties. *Behavior Research Methods*., *45*, 1246–1258. https://doi.org/10.3758/s13428-013-0326-1

Estes, Z., & Adelman, J. S. (2008a). Automatic vigilance for negative words in lexical decision and naming: Comment on Larsen, Mercer, and Balota (2006). *Emotion*, *8*(4), 441–444. https://doi.org/10.1037/1528-3542.8.4.441

Estes, Z., & Adelman, J. S. (2008b). Automatic vigilance for negative words is categorical and general. *Emotion*, *8*, 453–457. https://doi.org/doi:10.1037/a0012887

Ferré, P., Guasch, M., Moldovan, C., & Sánchez-Casas, R. (2012). Affective norms for 380 Spanish words belonging to three different semantic categories. *Behavior Research Methods*, *44*, 395–403.

Fox, E., Russo, B., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General*, *130*, 681–700. https://doi.org/10.1037/0096- 3445.130.4.681

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

González-Nosti, M., Barbón, A., Rodriguez-Ferreiro, J., & Cuetos, F. (2014). Effects of the psycholinguistic variables on the lexical decision task in Spanish: A study with 2,765 words. *Behavior Research Methods*, *46*, 517–525.

Guasch, M., Ferré, P., & Fraga, I. (2016). Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, *48*, 1358–1369.

Guasch, M., Padrón, I., Haro, J., Ferré, P., & Fraga, I. (2017). EmoFinder: A search engine for normative ratings of Spanish words. In *XIII International Symposium of Psycholinguistics*. Braga.

Hinojosa, J. A., Martínez-García, N., Villalba-García, C., Fernández-Folgueiras, U., Sánchez-Carmona, A., Pozo, M. A., & Montoro, P. R. (2016). Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. *Behavior Research Methods*, *48*, 272–284.

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R(2)GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*, 944–946.

Kanske, P., & Kotz, S. A. (2007). Concreteness in emotional words: ERP evidence from a hemifield study. *Brain Research*, *1 1 4 8*, 138–148.

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*, 473–481. https://doi.org/10.1016/j.cognition.2009.06.007

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065–1081.

Kuznetsova, A., Brockhoff, P. B., & Bojesen Christensen, R. H. (2016). lmerTest: Tests in linear mixed effects models. Retrieved from https://cran.r-project.org/package=lmerTest

Lang, P. J., Bradley, M. M., & Cuthbert, M. M. (1997). Motivated attention: Affect, activation and action. In P. J. Lang, R. F. Simons, & M. T. Balaban (Eds.), *Attention and Orienting: Sensory and Motivational Processes* (pp. 97–135). Hillsdale, NJ.

Larsen, R. J., Mercer, K. A., & Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion*, *6*(1), 62–72. https://doi.org/10.1037/1528-3542.6.1.62

Larsen, R. J., Mercer, K. A., Balota, D. A., & Strube, M. J. (2008). Not all negative words slow down lexical decision and naming speed: Importance of word arousal. *Emotion*, *8*(4), 445–452. https://doi.org/10.1037/1528-3542.8.4.445

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan*. Chapman & Hall/CRC Press.

Monaghan, J., & Ellis, A. W. (2002). What exactly interacts with spelling–sound consistency in word naming? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*, 183–206.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*, 1105–1107.

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Palazova, M., Mantwill, K., Sommer, W., & Schacht, A. (2011). Are effects of emotion in single words non-lexical? Evidence from event-related brain potentials. *Neuropsychologia*, *49*, 2766–2775.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, *12*, 765–805.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*(3), 380–391.

R development core team. (2017). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Redondo, J., Fraga, I., Comesaña, M., & Perea, M. (2005). Estudio normativo del valor afectivo de 478 palabras españolas. *Psicológica*, *26*(2), 317–326.

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, *39*, 600–605.

Ricketts, J., Davies, R. A. I., Masterson, J., Stuart, M., & Duff, F. J. (2016). Evidence for semantic involvement in regular and exception word reading in emergent readers of English. *Journal of Experimental Child Psychology*, *150*, 330–345.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Sebastián-Gallés, N., Carreiras, M., Cuetos, F., & Martí, M. A. (2000). *LEXESP. Léxico informatizado del español*. Barcelona: Publicacions UB.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work: Crowdsourcing research can balance discussions, validate findings and better inform policy. *Nature*, *526*, 189–191.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive Psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, *49*, 111–123.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1140–1154.

Vinson, D., Ponari, M., & Vigliocco, G. (2014). How does emotional content affect lexical processing? *Cognition and Emotion*, *28*(4), 737–746. https://doi.org/10.1080/02699931.2013.851068

Woollams, A. M., Lambon-Ralph, M. A., Plaut, D. C., & Patterson, K. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review*, *114*, 316–339.

Yao, Z., Yu, D., Wang, L., Zhu, X., Guo, J., & Wang, Z. (2016). Effects of valence and arousal on emotional word processing are modulated by concreteness: Behavioral and ERP evidence from a lexical decision task. *International Journal of Psychophysiology*, *110*, 231–242.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, *15*, 971–979.

Table 1. Summary of descriptive statistics for the critical psycholinguistic variables

	Mean	SD	Minimum	Maximum
Word form frequency	41.7	154.2	0.0	4909.4
Contextual Diversity	9.4	15.3	0.0	98.7
Age-of-Acquisition	4.0	1.2	1.2	6.9
Valence	5.3	1.4	1.2	8.7
Arousal	5.2	1.0	1.9	8.3
Familiarity	5.3	1.0	1.6	7.0
Imageability	4.7	1.2	1.7	7.0
Concreteness	4.7	1.0	2.0	6.8
Letters	6.5	1.8	3.0	10.0
Phonemes	6.4	1.8	2.0	11.0
Syllables	2.7	0.8	1.0	5.0
Length	6.5	1.8	2.5	10.5
Orthographic neighbourhood size	5.0	6.6	0.0	40.0
Phonological neighbourhood size	11.0	12.8	0.0	91.0
Levenshtein distance	1.8	0.6	1.0	3.8
N-size	8.0	9.5	0.0	59.5
Mean bigram frequency	5654.5	3838.7	2.9	30545.9

Table 2. Summary of pairwise (Pearson) correlations between psycholinguistic variables

	CD	Frequency	AoA	Familiarity	Imageability	Concreteness	Valence	Arousal	N-O	N-P	LevN	Nsize	Letters	Phonemes	Syllables	Length
Context Distinctiveness (CD)																
Word form frequency (frequency)	**0.73**															
Age-of-Acquisition (AoA)	**-0.44**	**-0.25**														
Familiarity	**0.43**	**0.23**	**-0.67**													
Imageability	**0.05**	0.01	**-0.56**	**0.38**												
Concreteness	**-0.10**	**-0.07**	**-0.37**	**0.16**	**0.72**											
Valence	**0.15**	**0.12**	**-0.19**	**0.18**	**0.05**	-0.01										
Arousal	0.01	-0.02	**0.14**	-0.03	**-0.14**	**-0.19**	**-0.41**									
Orthographic neighbourhood size (N-O)	**0.19**	**0.11**	**-0.29**	**0.15**	**0.19**	**0.09**	0.02	**-0.11**								
Phonological neighbourhood size (N-P)	**0.21**	**0.14**	**-0.30**	**0.15**	**0.19**	**0.09**	**0.04**	**-0.10**	**0.90**							
Levenshtein distance (LevN)	**-0.24**	**-0.13**	**0.37**	**-0.23**	**-0.17**	-0.03	**-0.07**	**0.06**	**-0.71**	**-0.69**						
N-size	**0.21**	**0.13**	**-0.30**	**0.16**	**0.19**	**0.09**	0.04	**-0.11**	**0.96**	**0.99**	**-0.71**					
Letters	**-0.25**	**-0.17**	**0.36**	**-0.14**	**-0.28**	**-0.21**	**-0.05**	**0.12**	**-0.63**	**-0.65**	**0.73**	**-0.66**				
Phonemes	**-0.25**	**-0.17**	**0.37**	**-0.14**	**-0.30**	**-0.23**	-0.04	**0.12**	**-0.62**	**-0.66**	**0.72**	**-0.66**	**0.98**			
Syllables	**-0.25**	**-0.17**	**0.35**	**-0.15**	**-0.23**	**-0.14**	-0.01	**0.05**	**-0.50**	**-0.50**	**0.66**	**-0.51**	**0.82**	**0.82**		
Length	**-0.25**	**-0.17**	**0.36**	**-0.14**	**-0.29**	**-0.22**	**-0.04**	**0.12**	**-0.62**	**-0.65**	**0.73**	**-0.66**	**1.00**	**1.00**	**0.82**	
Mean bigram frequency	**0.20**	**0.14**	**-0.19**	**0.13**	**0.06**	-0.01	0.01	-0.03	**0.28**	**0.23**	**-0.30**	**0.25**	**-0.17**	**-0.17**	**-0.25**	**-0.17**

Significant correlations ($p < .05$) are presented in bold, to avoid visual clutter; CD = Context distinctiveness; Frequency = Lexical frequency; AoA = Age-of-Acquisition; length in Letters, Phonemes or Syllables; Orth N-size = orthographic neighbourhood size;

Table 3. Summary of linear mixed-effects models of the cross-task data, including lexical decision and word naming data. Table shows baseline, baseline plus monotonic valence, baseline plus categorical positive-negative valence, and baseline plus emotional-neutral valence models.

Baseline model						
Fixed effects	Estimate	SE	df	t	p	
Intercept	507.0	8.2	75	61.6	<0001	***
Task	-29.7	12.6	68	-2.4	0.0210	*
Context Distinctiveness	-60.7	2.9	4040	-21.0	<0001	***
Curvilinear Context Distinctiveness	159.2	8.4	4025	19.0	<0001	***
AoA	5.5	1.3	4070	4.3	<0001	***
Curvilinear AoA	5.1	1.4	4336	3.6	0.0003	***
Familiarity	-15.1	1.0	4999	-14.9	<0001	***
Curvilinear familiarity	7.1	1.2	4461	5.8	<0001	***
Imageability	1.1	0.6	4254	1.8	0.0681	.
Length	-7.0	2.2	4216	-3.2	0.0012	**
Curvilinear Length	19.0	2.3	4223	8.2	<0001	***
Syllables	1.9	0.9	4259	2.2	0.0280	*
Neighbourhood Size	-10.2	2.4	4204	-4.3	<0001	***
Curvilinear neighbourhood size	15.1	4.0	4228	3.8	0.0002	***
Bigram frequency	2.2	0.5	4200	4.2	<0001	***
Valence						
Task × Context Distinctiveness	42.0	3.3	138000	12.9	<0001	***
Task × Curvilinear Context Distinctiveness	-107.2	9.4	138000	-11.4	<0001	***
Task × AoA	-5.6	1.5	138200	-3.9	0.0001	***
Task × Curvilinear AoA	-0.8	1.6	138600	-0.5	0.5989	
Task × Familiarity	10.9	1.1	139300	9.6	<0001	***
Task × Curvilinear familiarity	-6.1	1.4	138800	-4.5	<0001	***
Task × Imageability	0.9	0.6	138500	1.3	0.1780	
Task × Length	6.3	2.4	138500	2.6	0.0098	**
Task × Curvilinear Length	-10.7	2.6	138500	-4.1	<0001	***
Task × Syllables	1.6	1.0	138600	1.7	0.0924	.
Task × Neighbourhood Size	1.7	2.6	138500	0.6	0.5297	
Task × Curvilinear neighbourhood size	-3.8	4.5	138500	-0.8	0.4023	
Task × Bigram frequency	-1.5	0.6	138600	-2.5	0.0129	*
Task × Valence						
Random effects	Variance	SD				
Words - Intercepts	219.3	14.8				
Participants - Intercepts	2205.5	47.0				
Residual	8280.2	91.0				
	AIC	BIC				
	1666793	1667552				

Monotonic continuous valence						
Fixed effects	Estimate	SE	df	t	p	
Intercept	506.7	8.2	74	61.5	<0001	***
Task	-29.5	12.6	68	-2.3	0.0217	*
Context Distinctiveness	-61.0	2.9	4056	-21.1	<0001	***
Curvilinear Context Distinctiveness	160.4	8.3	4041	19.2	<0001	***
AoA	4.8	1.3	4086	3.7	0.0002	***
Curvilinear AoA	5.5	1.4	4355	3.9	0.0001	***
Familiarity	-14.9	1.0	5022	-14.7	<0001	***
Curvilinear familiarity	7.2	1.2	4481	5.9	<0001	***
Imageability	0.9	0.6	4271	1.5	0.1271	
Length	-7.4	2.2	4235	-3.4	0.0006	***
Curvilinear Length	19.3	2.3	4242	8.4	<0001	***
Syllables	2.2	0.9	4280	2.6	0.0101	*
Neighbourhood Size	-9.9	2.3	4222	-4.2	<0001	***
Curvilinear neighbourhood size	14.4	4.0	4246	3.6	0.0003	***
Bigram frequency	2.1	0.5	4218	4.1	<0001	***
Valence	-2.9	0.5	4135	-6.4	<0001	***
Task × Context Distinctiveness	42.2	3.3	138100	12.9	<0001	***
Task × Curvilinear Context Distinctiveness	-108.0	9.4	138100	-11.4	<0001	***
Task × AoA	-5.1	1.5	138200	-3.5	0.0005	***
Task × Curvilinear AoA	-1.1	1.6	138700	-0.7	0.4809	
Task × Familiarity	10.7	1.1	139400	9.5	<0001	***
Task × Curvilinear familiarity	-6.2	1.4	138900	-4.6	<0001	***
Task × Imageability	1.0	0.6	138600	1.5	0.1248	
Task × Length	6.6	2.4	138500	2.7	0.0069	**
Task × Curvilinear Length	-10.9	2.6	138600	-4.2	<0001	***
Task × Syllables	1.4	1.0	138600	1.4	0.1497	
Task × Neighbourhood Size	1.5	2.6	138600	0.6	0.5781	
Task × Curvilinear neighbourhood size	-3.3	4.5	138600	-0.7	0.4646	
Task × Bigram frequency	-1.4	0.6	138700	-2.5	0.0135	*
Task × Valence	2.1	0.5	138300	4.0	0.0001	***
Random effects	Variance	SD				
Words - Intercepts	215.7	14.7				
Participants - Intercepts	2205.6	47.0				
Residual	8279.2	91.0				
	AIC	BIC				
	1666756	1667534				

Positive-Negative categorical valence						
Fixed effects	Estimate	SE	df	t	p	
Intercept	510.5	8.3	75	61.8	<0001	***
Task	-31.5	12.6	68	-2.5	0.0148	*
Context Distinctiveness	-60.9	2.9	4059	-21.1	<0001	***
Curvilinear Context Distinctiveness	160.0	8.3	4044	19.2	<0001	***
AoA	4.9	1.3	4091	3.8	0.0002	***
Curvilinear AoA	5.3	1.4	4359	3.8	0.0001	***
Familiarity	-14.9	1.0	5024	-14.7	<0001	***
Curvilinear familiarity	7.0	1.2	4483	5.8	<0001	***
Imageability	0.8	0.6	4272	1.5	0.1439	
Length	-7.4	2.2	4237	-3.5	0.0006	***
Curvilinear Length	19.3	2.3	4244	8.4	<0001	***
Syllables	2.1	0.9	4282	2.5	0.0142	*
Neighbourhood Size	-10.4	2.3	4224	-4.4	<0001	***
Curvilinear neighbourhood size	15.3	4.0	4248	3.8	0.0001	***
Bigram frequency	2.1	0.5	4220	4.1	<0001	***
Valence	-5.7	1.0	4308	-5.9	<0001	***
Task × Context Distinctiveness	42.1	3.3	138100	12.9	<0001	***
Task × Curvilinear Context Distinctiveness	-107.5	9.4	138000	-11.4	<0001	***
Task × AoA	-5.3	1.5	138200	-3.6	0.0003	***
Task × Curvilinear AoA	-1.0	1.6	138700	-0.6	0.5389	
Task × Familiarity	10.8	1.1	139400	9.5	<0001	***
Task × Curvilinear familiarity	-6.1	1.4	138900	-4.5	<0001	***
Task × Imageability	1.0	0.6	138600	1.5	0.1326	
Task × Length	6.5	2.4	138500	2.7	0.0074	**
Task × Curvilinear Length	-10.9	2.6	138600	-4.2	<0001	***
Task × Syllables	1.5	1.0	138600	1.6	0.1177	
Task × Neighbourhood Size	1.8	2.6	138500	0.7	0.5017	
Task × Curvilinear neighbourhood size	-3.9	4.5	138600	-0.9	0.3885	
Task × Bigram frequency	-1.4	0.6	138700	-2.5	0.0135	*
Task × Valence	2.9	1.1	138600	2.7	0.0072	**
Random effects	Variance	SD				
Words - Intercepts	215.3	14.7				
Participants - Intercepts	2205.6	47.0				
Residual	8279.8	91.0				
	AIC	BIC				
	1666762	1667541				

Emotional-neutral categorical valence						
Fixed effects	Estimate	SE	df	t	p	
Intercept	507.3	8.3	75	61.5	<0001	***
Task	-29.4	12.6	68	-2.3	0.0225	*
Context Distinctiveness	-60.9	2.9	4041	-20.9	<0001	***
Curvilinear Context Distinctiveness	159.6	8.4	4026	19.0	<0001	***
AoA	5.5	1.3	4070	4.2	<0001	***
Curvilinear AoA	5.1	1.4	4338	3.6	0.0003	***
Familiarity	-15.1	1.0	4999	-14.9	<0001	***
Curvilinear familiarity	7.1	1.2	4461	5.9	<0001	***
Imageability	1.0	0.6	4254	1.8	0.0723	.
Length	-7.0	2.2	4217	-3.2	0.0012	**
Curvilinear Length	19.0	2.3	4223	8.2	<0001	***
Syllables	1.9	0.9	4259	2.2	0.0268	*
Neighbourhood Size	-10.2	2.4	4205	-4.3	<0001	***
Curvilinear neighbourhood size	15.2	4.0	4229	3.8	0.0002	***
Bigram frequency	2.2	0.5	4201	4.2	<0001	***
Valence	-0.5	1.0	4119	-0.5	0.5881	
Task × Context Distinctiveness	41.8	3.3	137900	12.7	<0001	***
Task × Curvilinear Context Distinctiveness	-106.7	9.5	137900	-11.3	<0001	***
Task × AoA	-5.6	1.5	138100	-3.9	0.0001	***
Task × Curvilinear AoA	-0.8	1.6	138600	-0.5	0.5963	
Task × Familiarity	10.9	1.1	139200	9.6	<0001	***
Task × Curvilinear familiarity	-6.1	1.4	138700	-4.5	<0001	***
Task × Imageability	0.9	0.6	138400	1.3	0.1865	
Task × Length	6.3	2.4	138400	2.6	0.0100	*
Task × Curvilinear Length	-10.7	2.6	138400	-4.1	<0001	***
Task × Syllables	1.6	1.0	138500	1.7	0.0891	.
Task × Neighbourhood Size	1.7	2.6	138400	0.6	0.5333	
Task × Curvilinear neighbourhood size	-3.7	4.5	138400	-0.8	0.4097	
Task × Bigram frequency	-1.4	0.6	138500	-2.5	0.0132	*
Task × Valence	-0.6	1.1	138100	-0.5	0.5993	
Random effects	Variance	SD				
Words - Intercepts	219.2	14.8				
Participants - Intercepts	2205.5	47.0				
Residual	8280.2	91.0				
	AIC	BIC				
	1666796	1667574				

*** p < 001; ** p < 01; * p < 05; 140,306 observations, 2555 words, 61 participants

Table 4. Summary of the final graded valence model

Fixed effects	Estimate	SE	df	t	p	
Intercept	506.4	8.6	80	59.0	<.0001	***
Task	-29.5	13.0	71	-2.3	0.0261	*
Context Distinctiveness	-61.8	3.7	2377	-16.8	<.0001	***
Curvilinear Context Distinctiveness	163.7	10.7	2374	15.3	<.0001	***
AoA	4.8	1.7	1058	2.7	0.0062	**
Curvilinear AoA	5.8	1.7	2495	3.3	0.0010	***
Familiarity	-15.0	1.4	751	-11.0	<.0001	***
Curvilinear familiarity	7.4	1.5	2543	4.9	<.0001	***
Imageability	1.0	0.8	392	1.3	0.1896	
Length	-8.1	2.9	1331	-2.8	0.0049	**
Curvilinear length	20.3	2.9	2453	7.1	<.0001	***
Syllables	2.1	1.2	364	1.8	0.0764	.
Neighbourhood size	-9.8	2.9	2411	-3.4	0.0008	***
Curvilinear neighbourhood size	13.8	5.0	2455	2.8	0.0057	**
Bigram frequency	2.2	0.7	523	3.3	0.0010	***
Valence	-3.4	1.4	2255	-2.4	0.0168	*
Arousal	0.6	1.4	2079	0.4	0.6923	
Valence x arousal	0.4	0.5	2389	0.7	0.4699	
Context Distinctiveness x Valence	-0.7	3.3	2415	-0.2	0.8274	
Curvilinear context distinctiveness x Valence	0.0	9.9	2397	0.0	0.9982	
Context distinctiveness x arousal	3.1	3.4	2428	0.9	0.3629	
Curvilinear context distinctiveness x arousal	-10.8	10.3	2412	-1.1	0.2912	
Task x Context distinctiveness	42.7	4.3	3668	10.0	<.0001	***
Task x Curvilinear context distinctiveness	-110.1	12.5	3737	-8.8	<.0001	***
Task x AoA	-4.9	2.1	538	-2.3	0.0204	*
Task x Curvilinear AoA	-1.5	2.0	3887	-0.7	0.4629	
Task x Familiarity	11.0	1.7	338	6.6	<.0001	***
Task x Curvilinear familiarity	-6.5	1.7	3940	-3.7	0.0002	***
Task x Imageability	0.8	0.9	161	0.9	0.3662	
Task x Length	7.4	3.5	741	2.1	0.0371	*
Task x Curvilinear length	-12.0	3.3	3836	-3.6	0.0003	***
Task x Syllables	1.5	1.5	160	1.0	0.3143	
Task x Neighbourhood size	1.5	3.4	3468	0.4	0.6545	
Task x Curvilinear neighbourhood size	-2.9	5.8	3826	-0.5	0.6119	
Task x Bigram frequency	-1.5	0.8	206	-1.9	0.0643	.
Task x Valence	4.3	1.7	3108	2.6	0.0089	**
Task x Arousal	-0.1	1.7	2192	-0.1	0.9550	
Task x Valence x arousal	-1.1	0.6	3749	-1.9	0.0515	.
Task x Context distinctiveness x Valence	4.0	3.8	3786	1.1	0.2925	
Task x Curvilinear context distinctiveness x Valence	-10.4	11.5	3764	-0.9	0.3681	
Task x Context distinctiveness x arousal	-1.8	3.9	3722	-0.5	0.6383	
Task x Curvilinear context distinctiveness x arousal	4.9	11.9	3772	0.4	0.6781	

Random effects	Variance	SD
Items on intercepts	83.1	9.1
Items on task	399.6	20.0
Participants on intercepts	2274.0	47.7
Participants on CD x arousal	0.7	0.8
Participants on CD x valence	0.0	0.0
Participants on valence x arousal	0.0	0.0
Participants on arousal	0.6	0.8
Participants on valence	0.0	0.2
Participants on bigram frequency	1.0	1.0
Participants on neighbourhood	1.3	1.2
Participants on syllables	9.4	3.1
Participants on length	40.3	6.3
Participants on imageability	2.7	1.6
Participants on familiarity	10.5	3.2
Participants on AoA	13.4	3.7
Participants on CD	1.2	1.1
Residual	8108.0	90.0

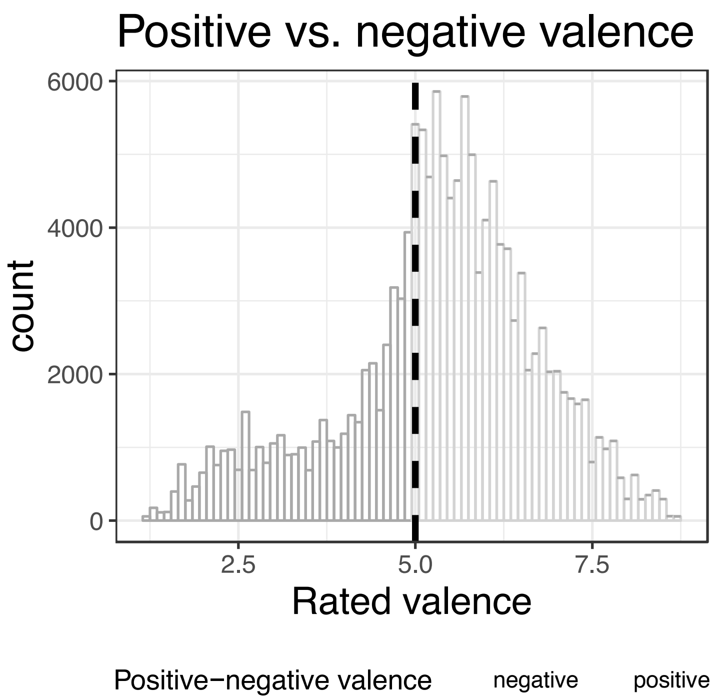
*** if $p < .001$; ** if $p < .01$; * if $p < .05$; 136,688 observations, 2555 words, 61 participants

Table 5. Summary of the task-specific model for responses from each task estimated separately.

	Lexical decision						Word naming					
Fixed effects	Estimate	SE	df	t	p		Estimate	SE	df	t	p	
Intercept	505.5	8.4	49	60.2	<.0001	***	477.4	10.2	27	46.7	<.0001	***
Context Distinctiveness	-60.3	3.7	2244	-16.5	<.0001	***	-19.5	2.6	2441	-7.4	<.0001	***
Curvilinear Context Distinctiveness	150.8	10.1	2256	15.0	<.0001	***	58.2	8.2	2468	7.1	0.0000	***
AoA	4.7	1.7	683	2.7	0.0069	**	-0.1	1.3	120	-0.1	0.9446	
Curvilinear AoA	5.4	1.7	2378	3.2	0.0013	**	4.3	1.2	2498	3.6	0.0003	***
Familiarity	-14.6	1.4	367	-10.6	<.0001	***	-4.1	1.0	129	-3.9	0.0001	***
Curvilinear familiarity	7.2	1.5	2434	4.9	<.0001	***	0.9	1.0	2498	0.8	0.4005	
Imageability	1.0	0.8	195	1.3	0.2121		1.9	0.6	41	3.2	0.0025	**
Length	-8.1	3.0	839	-2.7	0.0062	**	-0.8	2.2	229	-0.4	0.7133	
Curvilinear length	20.2	2.9	2338	7.0	<.0001	***	8.3	2.0	2488	4.2	<.0001	***
Syllables	2.0	1.1	256	1.8	0.0749	.	3.6	1.1	37	3.3	0.0020	**
Neighbourhood size	-9.7	2.9	2331	-3.3	0.0010	***	-8.2	2.1	1673	-4.0	0.0001	***
Curvilinear neighbourhood size	13.7	5.0	2341	2.7	0.0065	**	11.0	3.6	2471	3.1	0.0020	**
Bigram frequency	2.2	0.7	242	3.2	0.0016	**	0.7	0.5	65	1.6	0.1127	
Valence	-3.4	1.5	1791	-2.4	0.0183	*	0.9	1.0	2461	0.9	0.3603	
Arousal	0.6	1.4	1774	0.4	0.6859		0.5	1.0	757	0.5	0.6261	
Valence x arousal	0.4	0.5	391	0.7	0.4631		-0.8	0.3	2470	-2.2	0.0265	*
Context Distinctiveness x Valence	-0.7	3.3	2298	-0.2	0.8292		3.4	2.3	2475	1.4	0.1509	
Curvilinear context distinctiveness x valence	0.0	9.4	2279	0.0	0.9959		-11.1	7.5	2471	-1.5	0.1376	
Context distinctiveness x arousal	3.1	3.4	2296	0.9	0.3573		1.3	2.4	2309	0.5	0.5871	
Curvilinear context distinctiveness x arousal	-10.3	9.7	2294	-1.1	0.2867		-6.4	7.7	2463	-0.8	0.4029	
Random effects												
	Variance	SD					Variance	SD				
Items on intercepts	462.9	21.5					84.5	9.2				
Participants on intercepts	2147.0	46.3					2490.0	49.9				
Participants on CD x arousal	0.3	0.5					1.0	1.0				
Participants on CD x valence	0.0	0.0					0.0	0.0				
Participants on valence x arousal	0.1	0.4					0.0	0.0				
Participants on arousal	0.2	0.5					0.9	0.9				
Participants on valence	0.1	0.2					0.0	0.0				
Participants on bigram frequency	1.7	1.3					0.1	0.3				
Participants on neighbourhood	0.0	0.0					3.1	1.8				
Participants on syllables	5.2	2.3					15.2	3.9				
Participants on length	47.4	6.9					27.9	5.3				
Participants on imageability	3.3	1.8					2.1	1.4				
Participants on familiarity	12.8	3.6					6.8	2.6				
Participants on AoA	13.0	3.6					13.6	3.7				
Participants on CD	2.0	1.4					0.0	0.1				
Residual	9246.0	96.2					6653.0	81.6				

*** if p < .001; ** if p < .01; * if p < .05; in decisions, 79616 observations, 36 participants; in naming, 60690 observationsm, 25 subjects; in both, 2555 words

Figure 1. Distribution of standardized valence values, showing the sub-division of observations (word naming and lexical decision) into responses to negative or positive valence words (upper plot) or to emotional or neutral words (lower).



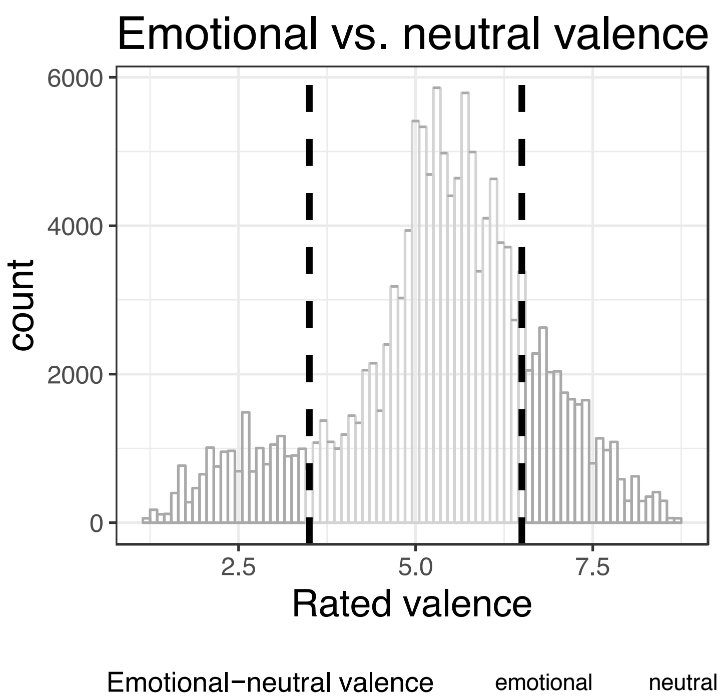


Figure 2. Scatterplot showing the relationship between word recognition RT and rated valence, for each task. Points in grey show trial-level latencies. Black lines show loess smoothers.

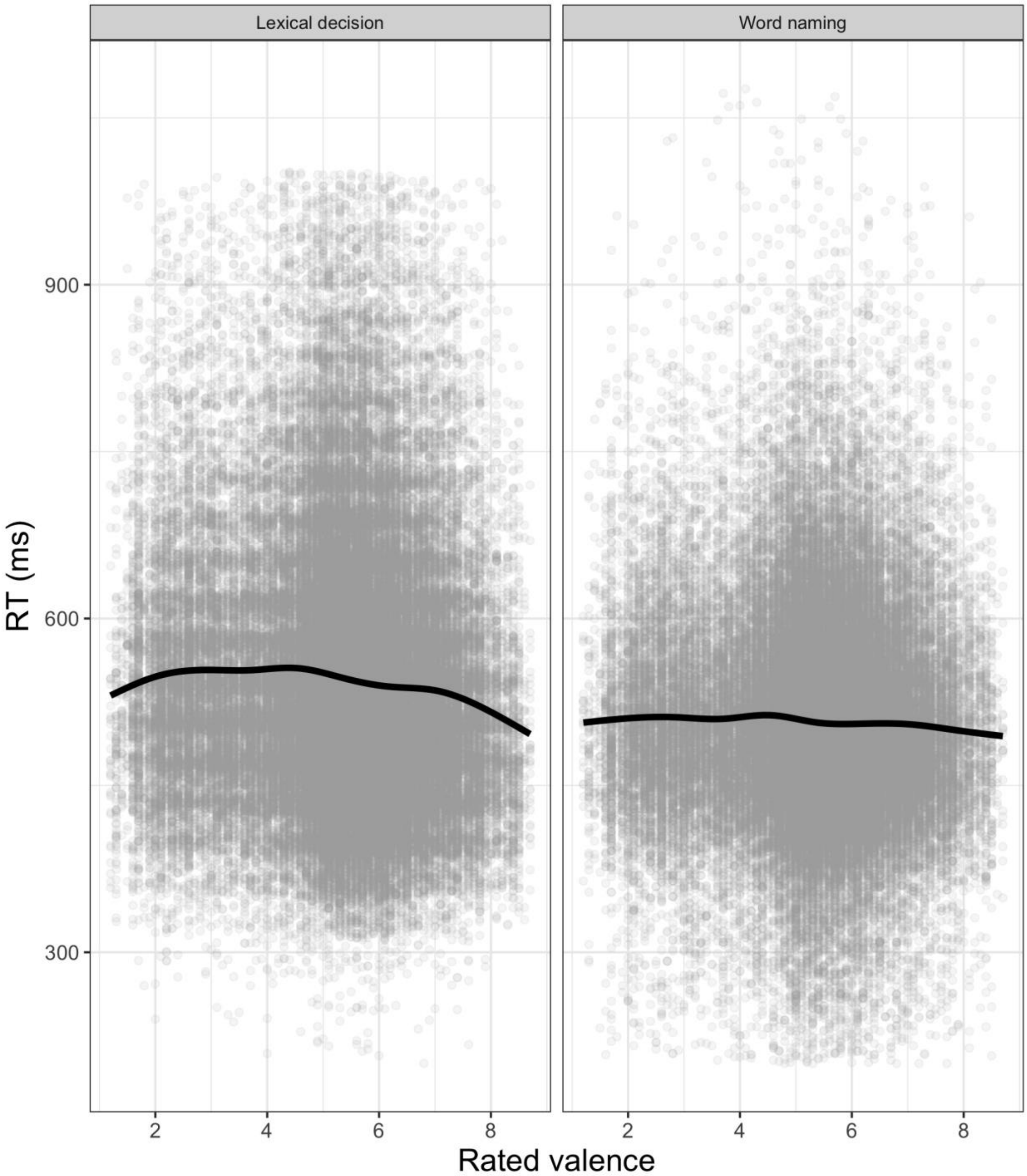


Figure 3. Scatterplot showing the relationship between word recognition RT and rated arousal, for each task. Points in grey show trial-level latencies. Black lines show loess smoothers.