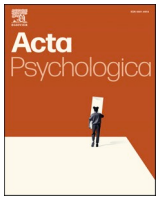
Acta Psychologica 218 (2021) 103359

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



The role of valence in word processing: Evidence from lexical decision and emotional Stroop tasks

Ethan Crossfield, Markus F. Damian *

University of Bristol, United Kingdom

ARTICLE INFO

Keywords:
Visual word processing
Valence
Lexical decision
Emotional Stroop
Mouse tracking

ABSTRACT

It is widely accepted that the valence of a word (neutral, positive, or negative) influences lexical processing, yet data from the commonly used lexical decision and emotional Stroop tasks has yielded inconsistent findings regarding the direction of this influence. One critical obstacle to investigating the independent effects of valence is the matching of emotional and neutral stimuli on the lexical, sublexical, and conceptual characteristics known to influence word recognition. The second obstacle is that the cognitive processes which lead to a lexical decision and a colour naming response are unobservable from the response latency measures typically gathered. The present study compiled a set of neutral, positive, and negative words matched triplet-wise on 26 influential characteristics. The novel "mouse tracking" technique was used to analyse the development of responses to these materials in variants of the lexical decision and emotional Stroop task. A conventional key-press emotional Stroop task is also reported. Results revealed a significant processing advantage for positive words over negative and neutral words in the lexical decision task, whereas valence alone did not produce any significant effects in the emotional Stroop task. The discrepancy between the effects of valence across these different tasks is discussed. We also suggest that previous conflicting findings may be confounded by unmatched emotional and neutral stimuli, thus inflating the potential effects of valence.

1. Introduction

Emotion is central to the human experience and is closely coupled with our cognition (Dolan, 2002), determining our thoughts, perception, and interaction with the world (Zajonc, 1984). Language is a mechanism for communicating and perceiving our own and others' emotions (Jonczyk et al., 2016). Indeed, all words of a given language can be characterised according to their emotional valence, whether they be negative (e.g. poison), positive (e.g. sunshine) or neutral (e.g. torch). Researchers have systematically gathered valence ratings for thousands of words (e.g., Warriner et al., 2013), thus permitting investigations into the effects of valence on word processing. Two experimental tasks which have been prominently featured to explore the potential impact of word valence are the lexical decision task (participants categorise singly presented stimuli as words or non-words, e.g., Vinson et al., 2013), and the emotional Stroop task (participants name or categorise valenced words according to the colour in which they are presented, e.g., Williams et al., 1996). While it is generally accepted that valence plays a role in lexical processing, what this role is remains unclear. As we will show below in our review of the literature, extant findings are inconsistent within as well as across tasks, probably due to the difficulty of identifying sets of words which are optimally matched on all aspects other than their valence. In the experiments reported here, we directly compare performance on lexical decisions and on the emotional Stroop task on the same set of matched word stimuli, allowing us to resolve at least some of the empirical inconsistencies in the literature.

1.1. Effects of valence in lexical decision and emotional Stroop tasks

A crucial question in the emotion literature is which human motivational system is dominant: the approach/appetitive system attuned to positive stimuli, and/or the withdrawal/aversive system attuned to negativity (Bradley, 2000). The model of motivated attention and affective states (Lang et al., 1990) proposes that emotional stimuli, regardless of polarity, capture attention to a greater extent than neutral stimuli due to the survival-related salience which both positive (e.g. food), and negative (e.g. threat) stimuli convey. Such a model might predict a reported processing advantage for emotional words,

* Corresponding author at: University of Bristol, School of Psychological Science, 12a Priory Road, Bristol BS8 1TU, United Kingdom.
E-mail address: m.damian@bristol.ac.uk (M.F. Damian).

https://doi.org/10.1016/j.actpsy.2021.103359
Received 11 August 2020; Received in revised form 18 June 2021; Accepted 22 June 2021
Available online 29 June 2021

irrespective of their polarity, over neutral words in the lexical decision task (LDT; Evaitar & Zaidel, 1991; Kousta et al., 2009). Further, emotional words are often more memorable than neutral words in recall tasks (Anooshian & Hertel, 1994; Ferre et al., 2010), an effect that can be attributed to the heightened salience of emotional words, leading to a stronger memory for them. However, other studies which also used the LDT identified a processing advantage for positive words only, but not for negative words (Wentura et al., 2000; Kissler & Koessler, 2011; Chen et al., 2015; see also Kanske & Kotz, 2007 for relevant EEG results). This might imply a dominant approach system, whereby the brain prioritises processing of positive stimuli to exploit resources that may aid survival (Mergen & Kuruoglu, 2017).

Irrespective of the inconsistency in the findings (is the effect of valence independent of polarity, or does it mainly affect positive words?), the dominant interpretation of valence effects in word processing tasks such as the lexical decision task is in terms of an 'early' mechanism of attentional capture which is presumably domain- and task-general (Gaillard et al., 2006; Zeelenberg et al., 2006; Kousta et al., 2009) and should therefore not be specific to lexical decisions. An alternative possibility is that effects of valence in lexical processing tasks may instead arise because emotional words are 'semantically richer' than neutral ones. According to Yap and Seow (2014), valence effects in tasks such as the LDT could arise from two possible sources (or from both). First, an early, preconscious and task-general effect of valence could be attributed to attentional capture such as those summarised above. Second, a later task-specific effect could emerge which manifests itself in lexical decisions due to enhanced 'semantic richness' (Pexman et al., 2008). Semantic richness is conceptualised as a multidimensional construct of variables including imageability, body-object interaction, and number of senses, and is assumed to influence lexical access via feedback from the semantic to the lexical level. If valence constitutes one of the dimensions of semantic richness (with valenced word being richer than neutral ones), then this mechanism could also account for effects of valence on performance. Based on distributional analysis of response times, Yap and Seow argued that valence effects were mainly due to semantic richness, rather than to the attentional mechanisms postulated previously.

A separate literature on the role of valence in word processing stems from the so-called emotional Stroop task (EST), which is assumed to access an early lexical level of processing (Aycicegi-Dinn & Caldwell-Harris, 2009; Winskel, 2013), and is thus considered potentially more sensitive to the effects of valence. Based on the original colour naming task (Stroop, 1935), the EST requires participants to state or categorise the colour of singly presented emotional and neutral words. In this task, the characteristic finding is that negative words slow down colour processing latencies more than neutral and positive words, a phenomenon termed the emotional Stroop effect (MacKay et al., 2004; Williams et al., 1996). This effect is typically attributed to the threatening connotations of negative words, which disrupt processing of the task-relevant properties of the stimulus, such as its colour (Ben-David et al., 2012). In contrast to the model of motivated attention (Lang et al., 1990), the automatic vigilance hypothesis posits that humans preferentially attend to negative stimuli due to the greater time urgency required when dealing with a negative event, such as a threat, compared to positive events, such as feeding (Pratto & John, 1991). In EST, the survival-related salience of negative stimuli consequentially makes it difficult to disengage attention from the stimulus and to process the relevant aspects of the task (Fox et al., 2001), leading to a delay in colour naming when faced with negative valence.

The interference effect of negative words in colour naming or categorisation was originally described as a "within-trial" effect, reflecting the automatic attentional capture of the emotional content of words (Williams et al., 1996). However, more recent research has identified both 'fast' and 'slow' emotional Stroop effects (Frings et al., 2010). The 'fast' effect refers to the direct influence of valence on the current trial *N*, whereas the 'slow' effect reflects potential carry-over effects of valence from the previous trial $N-1$. For instance, Frings et al. characterised each trial in an EST with regard to whether the target word was neutral or negative, as well as whether trial $N-1$ was neutral or negative. A 'fast' effect of valence is captured by comparing responses on negative vs. neutral trials N which all follow a neutral trial $N-1$. By contrast, a 'slow' valence effect is explored by analysing only trials *N* with neutral valence, but dependent on whether the previous trial $N-1$ was negative or neutral. Frings et al. reported a 'fast' effect (participants responded slower to negative than neutral trials *N*) as well as a 'slow' effect (participants responded slower to neutral stimuli when the previous trial was negative than when it was neutral). The authors argue that the 'fast' and 'slow' effects may be driven by the same process that begins in the current trial and persists until the next. According to the automatic vigilance hypothesis, the difficulty in disengaging attention from negative stimuli (Algom et al., 2004), delays processing of relevant stimulus information such as colour (Öhman et al., 2001), resulting in the 'fast' effect. The same process may also interfere with performance on the subsequent trial, resulting in the 'slow' effect of negative valence in the EST (McKenna & Sharma, 2004).

The term 'emotional Stroop effect' typically only refers to the observed delay in colour naming/categorisation for negative words (McKenna & Sharma, 1995). Indeed, many emotional Stroop investigations did not include positive words (e.g., Richards et al., 1992), and those that did often found no significant differences between latencies on neutral and positive trials (Eilola et al., 2007; Eilola & Havelka, 2011). Nevertheless, some researchers hypothesise that positive valence may have an opposite facilitating effect to the interference effect of negative valence. If the threatening aspects of negative stimuli narrow our attentional scope to focus on the valence stimulus dimension (Lazarus, 1991), then the lack of threat positive stimuli pose may not result in the automatic vigilance which disrupts task performance. A recent EST identified a fast and slow effect of positive words, which facilitated performance on current and subsequent trials (Liu et al., 2018). These findings may be interpreted as positive valence promoting attentional reorientation (Johnson et al., 2010), allowing attention to shift from the irrelevant (valence), to the relevant stimulus dimension (colour).

In summary, the literature on the potential role of valence in word processing tasks presents a complex array of partially contradictory findings. In studies featuring the LDT, valence is generally considered relevant but it is unclear whether the effect of valence is monotonic (i.e., negative > neutral > positive; e.g., Kuperman et al., 2014; Larsen et al., 2008), takes an inverted-U form (negative < neutral > positive; e.g., Kousta et al., 2009) or whether there is a processing advantage for positively valenced stimuli only (e.g., Chen et al., 2015; Kissler & Koessler, 2011). In the EST, negative valence generally appears to exert an inhibitory force (colour naming/classification responses are slower for negative than for positive words) but many studies did not include positively valenced stimuli, for which valence might exert a facilitatory effect (Liu et al., 2018).

1.2. Isolating effects of valence in word stimuli

A potential explanation for the inconsistencies in research on LDT and EST stems from the fact that these studies generally involve the comparison of different words across the valence conditions. The last few decades of research on word processing have identified a host of relevant conceptual and linguistic variables (e.g., Ferrand et al., 2018) and many of these are confounded with valence. For instance, a meta-analysis conducted on the words used in 32 published emotional Stroop studies revealed that the emotional words tended to be longer and lower in frequency of use compared to the neutral words (Larsen et al., 2006). Larsen et al. noted that these lexical differences between the negative and neutral words used are in the direction predicted to slow down reaction time on negative trials. Longer (Frederiksen & Kroll, 1976), and lower frequency (Monsell et al., 1989) words are often

processed slower, suggesting that some of the reported valence effects in emotional Stroop research may be due to other, non-emotional aspects of the stimuli which influence word processing (Kahan & Hely, 2008).

An alternative to finding valenced and non-valenced words which are matched on all aspects other than valence is to conduct multiple linear regression analyses on large sets of words and their reaction times, and to see whether valence makes an independent contribution. In a reply to Larsen et al., Estes and Adelman (2008) carried out several regression analyses of 1011 words with the available lexical decision and naming data from the English Lexicon Project (Balota et al., 2007). The analyses revealed that when the influences of variables including word length, frequency, orthographic neighbourhood size, contextual diversity, and arousal were controlled for, a small valence effect of slower negative than positive word recognition remained (and interestingly, the effect was categorical, such that the extremity of a word's valence appeared irrelevant). However, Estes and Adelman did not include influential variables such as subjective familiarity (Connine et al., 1990), imageability (Sadoski & Paivio, 2001), concreteness (Paivio, 1971, 2013), and age-of-acquisition (AoA; Morrison & Ellis, 1995), in their analyses. A set of analyses reported by Kuperman et al. (2014) probably constitutes the most advanced attempt to explore the role of valence (and of arousal) via a multiple regression approach, with word length, neighbourhood density, frequency, contextual diversity, and age of acquisition included as additional predictors in their full analysis of LDT and naming times of 12,658 words. A further analysis included body-object interaction, imageability, no. of senses, semantic diversity, and semantic experience on a subset of 1083 words for which these were available; this approach was motivated by proposals that highlight as important variables the sensory experience which a word evokes (Juhasz et al., 2011), the extent to which we can physically interact with the word's referent (body-object interaction; Siakaluk, Pexman, Aguilera et al., 2008), and the variation in meaning across all the contexts in which the word appears (semantic diversity; Hoffman et al., 2013). Overall, their results suggested orthogonal monotonic effects of both valence and arousal, with both effects more pronounced for low- than for high-frequency words.

In the experiments reported below, we attempted to assess potential effects of valence in a direct comparison of LDT and EST. A direct comparison of results from these two tasks, performed on a carefully matched of valenced stimuli, will allow us to exclude confounds from imperfectly matched stimuli which were presumably present in previous studies on this issue, and to focus exclusively on issues of differential task characteristics and demands with regard to valence. Doing so necessitates the identification of words which are matched on most variables other than valence (we are not aware of studies which would have attempted to tackle valence effects in emotional Stroop tasks via multiple regression). From the literature summarised above, it is clear that great care must be taken to identify and control potentially confounding variables.

1.3. Exploring valence effects using 'mouse tracking'

Research on word processing is usually done via experiments in which participants make a key press decision on a given stimulus. Even if all known lexical and sublexical variables are controlled for, the reaction times measures typically used in the EST and LDT only provide data on when participants have reached their decision, but not the processes that lead to it (Chen et al., 2015). While some researchers have applied physiological techniques to study how valence influences our physical responses to stimuli (Eilola & Havelka, 2011), they do little to uncover how valence impacts our online processing of words. Motor responses, such as pressing a computer key to convey a response, are continuously updated during the processing of a stimulus and the decision-making process (Abrams & Balota, 1991). Recently, experimental techniques have been explored in which responses are given not via a key press, but rather via a dynamic movement, such as pointing towards a response zone (e.g., Erb et al., 2016) or carrying out the response via a computer mouse movement (e.g., Freeman & Ambady, 2010). With these techniques, variables and processes which impact on decision making emerge in the properties of the dynamic response movement trajectories (see Erb et al., 2021, and Schoemann et al., 2019, for recent overviews).

In a typical "mouse tracking" study (e.g. Barca & Pezzulo, 2015), participants begin a trial by clicking on a field located at the bottom-centre of the computer screen. A target stimulus is shown in the centre of the screen and responses are indicated by moving the mouse and clicking on one of the two response button boxes, typically located at the top left and top right of the screen. Mouse position is continuously recorded during trials and measures including initiation time (interval in ms between the participant clicking the START box field and the onset of mouse movement), reaction time (interval in ms between stimulus presentation and response), and the curvature of the mouse trajectory itself are computed. Recording the mouse movement trajectory allows measurement of the 'attraction' to the non-selected response (Freeman & Ambady, 2010) and a 'curvature' towards the incorrect response presumably reflects the competition between simultaneously active action plans. This technique has been successfully used in various areas of cognitive research, such as social psychology (Faust et al., 2019), development (Krueger & Storkel, 2020), and psycholinguistics (e.g., Tomlinson et al., 2013).

Recent work has begun to use mouse tracking to uncover the characteristics of lexical processing in LDT (e.g., Barca et al., 2017), and Stroop tasks (Incera et al., 2013). For instance, in a LDT, Barca and Pezzulo (2012; see also Barca & Pezzulo, 2015) showed that responses to high-frequency words were carried out with a stronger curvature of the response movement from start to finish than responses to low-frequency words. Furthermore, "nonword" responses to pseudoword letter strings were carried out with more curvature than to random letter strings. In both cases, the curvature appeared to reflect the 'attraction' of an ultimately correct response towards the response alternative. Hence, response trajectories reflect the dynamics of decision making as it takes place, and a lexical variable, in this case word frequency, clearly emerged in the "word" trajectories. We are aware of a single study in which mouse tracking has been used in conjunction with a manual Stroop task (Yamamato et al., 2016), although in a "reversed" task (responses to words rather than colours). Despite the relative lack of evidence, it appears that mouse tracking can be used in conjunction with word processing tasks, and has the potential to provide insight into the variables and processes which lead to a decision (perhaps more so than data derived from conventional key presses do).

1.4. The present study

As summarised above, the possibility that the emotional valence of a word could affect its processing has been investigated in two largely separate streams of research. Work on single word processing, typically using the LDT, suggests that valence is indeed an important predictor of processing speed, but it remains unclear what form the effect takes, and specifically, whether or not there is a difference between positively or negatively valenced words. Studies using the EST have mainly focused on the comparison of neutral to negatively valenced words, and again the exact way in which valence might affect colour classification latencies is unresolved. As pointed out previously (e.g., Larsen et al., 2006) and also above, the main difficulty in advancing the issue is that valence is confounded with a plethora of other conceptual, lexical, and sublexical variables. This makes it difficult to match stimuli on all characteristics other than their valance, and it is clear that most previous studies in both literatures did not meet that criterion.

In the current study, we aimed to generate a set of neutral, negatively, and positively valenced words which were matched as optimally as possible on variables other than valence. We then used this set of words both in a LDT (Experiment 1) and ESTs (Experiments 2 and 3). Using the same set of words across both experimental tasks promises to

potentially unify the so-far largely separate streams of research on the role of valence reviewed above. We selected sets of neutral, negative, and positive words which were statistically matched on more than two dozen of relevant variables. These of course included 'standard' properties such as word length, frequency of occurrence, and age of acquisition, but we also made a concerted effort to match stimuli on recently suggested 'conceptual' variables such as sensory experience (Juhasz et al., 2011), body-object interaction (Siakaluk, Pexman, Sears et al., 2008), and semantic diversity (Hoffman et al., 2013). Words across the three valence categories were also matched on arousal (the extent to which a word is calming or exciting) as it has been shown (Kuperman et al., 2014) that valence and arousal make largely independent contributions to word processing times (we note that many previous studies, such as the EST reported by Frings et al., 2010, did not hold arousal constant when varying valence). Kuperman et al. (2014) also highlight the importance of keeping word frequency constant in valence research, as while valence and arousal exert independent effects on lexical processing, both variables interact with word frequency. Indeed, the effects of higher-level conceptual variables such as imageability and AoA are stronger among low frequency words than high frequency words (Cortese & Schock, 2013), therefore our stimuli were matched on multiple measures of word frequency to avoid this previously ignored and influential confound (Larsen et al., 2006).

One residual variable which is difficult if not impossible to disentangle from valence is 'dominance', or the extent to which a word denotes an entity which is weak/submissive or strong/dominant. In Warriner et al.'s (2013) ratings of almost 14,000 English words, valence and dominance were strongly correlated, $r = 0.72$ (p. 1196). Such a strong association makes it unlikely that stimuli could be identified which can be varied on valence but held constant on dominance (plus all the other potentially relevant variables). Our stimuli therefore differed not only on valence but also on dominance and hence potential effects attributed to valence could instead have been caused by dominance. This aspect will be revisited in the General Discussion.

As briefly summarised above, there is growing use of mouse tracking measurements in lexical decision (Barca et al., 2017; Barca & Pezzulo, 2012, 2015) and standard Stroop research (Incera et al., 2013; Incera & McLennan, 2016). These existing studies validate the use of mouse tracking to explore lexical processing, but to our knowledge this approach has not been used to investigate the role of valence. Hence, Experiments 1 (LDT) and 2 (EST) used mouse tracking as the response mode. An additional key-press EST (Experiment 3) is also reported to allow a more direct comparison to previous research. If valence does influence word processing independent of other lexical and sublexical variables, then in our LDT, latencies as well as mouse trajectories should be affected by the valence manipulation. The exact pattern of findings is of major importance here, as in the past a generic processing advantage for emotional over neutral words has been proposed (Kanske & Kotz, 2007; Kousta et al., 2009) but it is also possible that valence exerts a monotonic effect, with latencies (and in our case, potentially trajectories) following a negative > neutral > positive pattern (Kuperman et al., 2014). Regarding our ESTs, based on the results reported by Frings et al. (2010) we expected both 'fast' and 'slow' effects of negative valence to emerge. If effects of valence in the emotional Stroop effect are really constrained to negative valence, then we would expect little influence of positively valenced words in this task. However, as discussed above, many reported valence effects are confounded by unmatched emotional and neutral stimuli sets (Larsen et al., 2006), and improved matching reduces the size of this effect (Estes & Adelman, 2008). Therefore, the null finding of no differences between neutral and emotional word processing should not be ruled out.

2. Experiment 1

2.1. Method

2.1.1. Participants

Forty-six undergraduate students (13 male) from the University of Bristol participated in the experiment as part of a course requirement. The mean age was 20.5 ($SD = 3.48$) years. All participants reported (corrected to) normal vision and were comfortable navigating a computer mouse with their right hand. Participants also confirmed that they were native English speakers and not fluent in any other language. Informed consent was obtained from all participants.

2.1.2. Materials and design

The stimuli included 87 words of three different valence categories (29 positive, 29 negative, and 29 neutral words) which were selected from the affective ratings for valence, arousal, and dominance for over 13,000 English words (Warriner et al., 2013). Words were chosen to cluster around the lower range of valence ratings for negative words ($M = 2.09$, $SD = 0.27$, range 1.5–2.5), around the middle range for neutral words ($M = 5.13$, $SD = 0.26$, range 4.5–5.5), and towards the higher range for the positive words ($M = 7.44$, $SD = 0.37$, range 7.0–8.5). Ratings differed significantly between the conditions, $F(2, 84) = 2225$, $p < .001$, with all conditions differing from one another; $ps < .001$, using Tukey-corrected follow-up tests. Dominance ratings from the Warriner et al. norms for these words also significantly increased from negative ($M = 3.50$, $SD = 0.58$, range = 2.6–5.0) to neutral ($M = 5.14$, $SD = 0.72$, range = 3.6–6.2) to positive ($M = 6.14$, $SD = 0.74$, range = 5.0–7.9), $F(2, 84) = 110.9$, $p < .001$, with all conditions differing from one another, $ps < .001$. Apart from valence and dominance, stimuli were statistically matched triplet-wise on all other lexical and sublexical variables listed in Table 1. See Appendix A for a full list of materials. 87 orthographically legal and pronounceable non-word stimuli were obtained from the English Lexicon Project database (https://elexicon.wustl.edu/; Balota et al., 2007) and were matched pairwise to the word stimuli in terms of length (number of letters).

We collected familiarity, imageability, sensory experience, and body-object interaction ratings for the word stimuli (Crossfield & Damian, unpublished data), following the Bristol norms procedure (Stadthagen-Gonzalez & Davis, 2006). Thirty-six (7 male) participants provided familiarity and imageability ratings (mean age = 19.5 years, $SD = 1.09$), and a different sample of 32 (7 male) participants provided sensory experience and body-object interaction ratings (mean age = 20.2 years, $SD = 2.65$). Ratings were collected online on a 1 (low) to 7 (high rating) scale via Gorilla (https://gorilla.sc/), a commonly used platform for online behavioural tasks (Anwyl-Irvine et al., 2020). Participants rated a list of 140 words for each variable separately (e.g. providing 140 familiarity ratings, then providing 140 imageability ratings) by selecting the number on the 1–7 scale which corresponded to their rating. The 1–7 scale was presented underneath each target word for the duration of the studies. To improve the validity of ratings, 20 control words were added for each variable (e.g. 10 low imageability, and 10 high imageability words were added for the imageability ratings) in order to represent the entire range of the rating scale.

2.1.3. Procedure and apparatus

Participants were tested in a university laboratory in groups of no more than 30. Word and non-word stimuli were presented in a black lowercase Arial font, size 28, on a white background using the software MouseTracker (Freeman & Ambady, 2010). MouseTracker collected the raw data of each mouse trajectory, recording x and y coordinates of the trajectory of the mouse movement every 16 ms. Trials were viewed from a comfortable distance (approximately 60 cm) on a 23 inch Dell P2319H monitor with screen resolution 1920 × 1080. Participants were instructed to begin each trial by using the computer mouse to click on the grey START box located at the bottom-centre of the screen. The

Table 1
Lexical and sublexical characteristics of the word stimuli. Averages per condition (standard deviations in parentheses).

	Condition		
	Negative	Neutral	Positive
Valence[a]	2.09 (0.3)	5.13 (0.3)	7.44 (0.4)
Dominance[a]	3.50 (0.6)	5.14 (0.7)	6.14 (0.7)
Arousal[a]	4.89 (0.9)	4.66 (0.7)	4.75 (1.1)
Concreteness[b]	3.60 (0.9)	3.96 (1.0)	3.55 (0.9)
Familiarity[c]	4.34 (0.7)	4.15 (0.8)	4.50 (0.7)
Imageability[c]	4.53 (0.8)	4.46 (1.3)	5.01 (1.1)
Age of acquisition[d]	7.97 (1.7)	7.69 (2.0)	7.43 (2.2)
Contextual diversity[e]	3.12 (0.1)	2.88 (0.1)	3.15 (0.8)
Semantic diversity[f]	1.57 (0.2)	1.54 (0.3)	1.62 (0.2)
Number of senses[g]	3.59 (3.2)	3.83 (3.8)	4.17 (2.7)
Sensory experience[c]	3.30 (0.9)	3.01 (0.9)	3.24 (0.9)
Body-object interaction[c]	3.55 (0.8)	3.73 (1.3)	3.28 (0.9)
Number of letters[h]	6.21 (1.4)	6.76 (1.7)	6.66 (1.9)
Number of phonemes[h]	5.07 (1.7)	5.41 (1.6)	5.76 (1.7)
Number of syllables[h]	2.03 (0.9)	2.10 (0.9)	2.31 (0.8)
Number of morphemes[h]	1.07 (0.3)	1.17 (0.4)	1.17 (0.4)
Orthographic neighbourhood[h]	2.21 (3.5)	2.10 (3.8)	2.48 (5.1)
Phonological neighbourhood[h]	4.90 (6.6)	5.14 (8.3)	2.72 (5.0)
Orthographic similarity (OLD20)[i]	2.12 (0.6)	2.39 (0.8)	2.21 (0.6)
Orthographic frequency[j]	37.5 (75)	22.2 (33)	27.7 (44)
Semantic neighbourhood density[j]	0.58 (0.08)	0.54 (0.12)	0.58 (0.08)
Semantic neighbourhood size[j]	3129 (3055)	2249 (2712)	3070 (2679)
Celex frequency[k]	34.1 (64)	17.5 (33)	20.1 (29)
Celex written frequency[k]	35.4 (48)	17.5 (33)	21.0 (30)
Celex spoken frequency[k]	16.6 (23)	18.3 (61)	9.30 (16)
SUBTLEX-UK frequency[l]	4.02 (0.6)	3.77 (0.6)	4.09 (0.7)
Bigram frequency type[m]	50 (35)	47 (33)	62 (57)
Bigram frequency token[m]	914 (450)	683 (505)	1262 (2352)

[a] From Warriner et al. (2013). Valence, dominance, and arousal ratings on a 1–9 scale.

[b] From Brysbaert et al. (2014). The degree to which a word refers to a tangible entity. Ratings on a 1–5 scale.

[c] Ratings conducted for this study (see section "Materials and design"). Ratings on a 1–7 scale.

[d] From Kuperman et al. (2012). Ratings of the age (in years) at which participants thought they had learned the word.

[e] From SUBTLEX-UK (Van Heuven et al., 2014). Contextual diversity denotes the number of contexts in which a word appears (Adelman et al., 2006). Here, contextual diversity represents the log total number of television programmes in SUBTLEX-UK in which a word occurred.

[f] From Hoffman et al. (2013). The variation in meaning across all the contexts in which the word appears. Latent semantic analysis based on the written text portion of the British National Corpus (BNC; British National Corpus Consortium, 2007).

[g] From WordNet (Princeton University, 2010). Denotes the number of different meanings a single word can have.

[h] From the English Lexicon Project (Balota et al., 2007).

[i] From Yarkoni et al. (2008). Orthographic Levenshtein distance denotes the number of insertions, deletions, and substitutions needed to generate one string of elements from another. OLD20 represents the mean LD from a word to its 20 closest orthographic neighbours.

[j] From Shaoul and Westbury (2010). Orthographic frequency denotes the averaged frequency (per million) of a word's orthographic neighbours. Semantic neighbourhood density refers to the density of words in a given semantic neighbourhood, and semantic neighbourhood size denotes the amount of words in this neighbourhood. Based on the Hyperspace analog to language (HAL) high-dimensional model of semantic space which uses global co-occurrence frequency of words in a large corpus of text.

[k] From Baayen et al. (1995). Counts per million in the CELEX Lexical database.

[l] From Van Heuven et al. (2014). Log of count per million in the SUBTLEX-UK corpus.

[m] From N-Watch (Davis, 2005). Type bigram frequency denotes the number of words in which a letter bigram occurs in a given position; token frequency represents the sum of frequencies of the types. Computed on the basis of the COBUILD/CELEX word frequency corpus.

START box disappeared when clicked and the stimulus (word or non-word) immediately appeared in the centre of the screen for 2500 ms or until the participant completed their response, upon which the START box would return to its original position. As each trial began through clicking the START box, participants could self-administer breaks during the experimental session by delaying the onset of the next trial.

Responses were made by moving the mouse and clicking on one of the two black response button areas on the screen. Participants were instructed to respond as quickly and accurately as possible by clicking on the top right area of the screen for word stimuli, and the top left area for non-word judgments (see Fig. 1, top panel, for an example trial screen). After 10 practice trials (5 word, 5 non-word), the 174 experimental stimuli (87 word, 87 non-word) were singly presented in a different random order for each participant. The experimental session lasted approximately 25 min.

2.2. Results

Data were processed in *R* (R Core Team, 2020) using the package *mousetrap* (Kieslich et al., 2019) and results were statistically analysed using the packages *afex* (Singmann et al., 2020) and *emmeans* (Lenth, 2020). Trials with nonwords were discarded from further analysis. For each trial with a word stimulus, we computed response accuracy, initiation time (the time at which a participant initiated the mouse movement, calculated as the first time sample within a trial on which the mouse cursor was moved outside the "Start" region), movement duration (the time interval between initiation time and clicking on the response button), and response latency (the time interval between onset of the target display and clicking on the response button). We

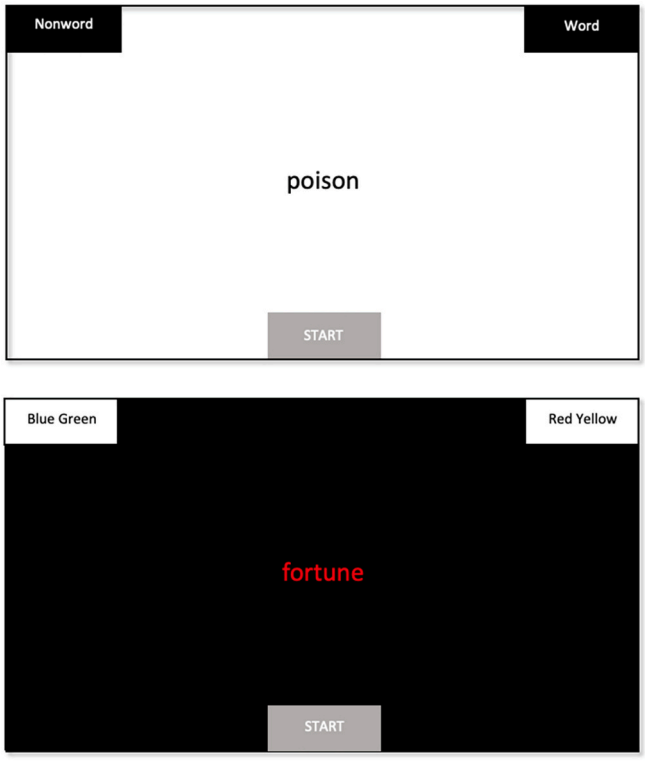


Fig. 1. Example trial screen for lexical decision task (top panel; Experiment 1) and EST (bottom panel; Experiment 2). Once participants pressed the START button, a random experimental stimulus would appear in the centre of the screen, and participants categorised the stimuli using the computer mouse to click on a response button, according to lexical status (Experiment 1; "nonword" vs. "word") or according to colour (Experiment 2; "blue/green" or "red/yellow"). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

furthermore computed two measures of 'curvature' which are widely used in the literature on mouse tracking (Freeman & Ambady, 2010). The first was area-under-curve (AUC), which was defined as the geometric area between the actual and an idealised trajectory which proceeded in a straight line from Start to Response region. The second was maximum deviation (MAD), i.e., the largest perpendicular deviation between the actual and the idealised trajectory. AUC was measured in in cm^2, and MAD was measured in cm. Trajectories with greater AUC and/or MAD indicate greater attraction to the non-selected response alternative (Freeman & Ambady, 2010). Measures were then aggregated within and across participants. Trials on which participants had made an error were excluded from the analysis of the other dependent variables (1.4%). We further excluded data from trials on which participants had made no response (1.1%), as well as from trials on which a participant's response latency was above or below 2.5 standard deviations of the participants conditional mean (2.5%).

Fig. 2 shows the results. The right side of the figure shows averaged time-normalised[1] mouse movement trajectories. These show an effect of valence on curvature, with average trajectories in the "neutral" condition most curved, trajectories in the "positive" condition straightest, and trajectories in the "negative" condition in between. The inset panels on the left hand side of the figure present the six main dependent variables of interest, namely error rate, initiation time, duration, response latencies, and the two curvature measures (AUC and MAD). Repeated measures one-way analyses of variance (ANOVAs) were conducted on each measure. For detailed statistics see Appendix B in the Supplementary materials. Valence significantly affected all measures other than initiation times. For the former, Tukey-corrected follow-up tests showed that the positive condition differed significantly from the baseline neutral condition on all measures (errors, durations, and RTs: $p < .0001$; AUC: $p = .039$; MAD: $p = .0002$). The negative condition differed significantly from the neutral condition on errors; $p = .011$; marginally significantly on durations; $p = .092$, and RTs, $p = .073$, but not on AUC; $p = .707$, nor on MAD; $p = .206$.

We additionally conducted items analyses on errors and response latencies. For errors, the effect of valence was marginally significant, $p = .059$, with a significant difference between the negative and the neutral condition, $p = .048$, but no difference between the positive and the neutral condition, $p = .285$. For response latencies, there was a highly significant overall effect of valence; $p = .009$, a highly significant difference between the positive and the neutral baseline condition; $p = .0068$, but no significant difference between the negative and the neutral condition; $p = .462$.

Finally, as described in the Introduction, in the literature on the EST it is customary to try to identify "slow" effects of valence, i.e., those which arise not as a result of processing in a given trial N, but rather carrying over from the preceding trial N − 1 with a valenced stimulus. Such an analysis is more difficult to perform in the LDT because words and nonwords are randomly intermixed. To explore similar effects,

we conducted an analysis on nonwords when preceded by valenced words, and found a marginally significant effect on response latencies, $F_1(2, 90) = 2.93$, $p = .059$, with average nonword latencies of 1060 ms, 1083 ms, and 1078 ms when preceded by positive, neutral, and negative words. However, a corresponding items analysis was not significant, $F_2(2, 238) = 2.32$, $p = .101$, and neither were effects on any of the other dependent variables (errors; initiation times; durations; MAD; AUC). We conclude that in the LDT, "slow" effects of valence are difficult to establish.[2]

[1] X and y coordinates were sampled every 16 ms until a response was made. Time normalisation involves the interpolation of the raw coordinates from each response into 101 equally sized timesteps.

[2] We would like to thank a reviewer for suggesting this analysis.

2.3. Discussion

Valence of response words in the LDT significantly affected all dependent measures other than initiation times. The null finding concerning initiation times is predicted on the assumption that in the mouse tracking paradigm, cognition and action are not 'serial' and hence a response movement is initiated *before* a decision has been fully completed (e.g., Freeman et al., 2011). With regard to the specific pattern evoked by valenced stimuli, compared to words from the neutral condition *positively* valenced words were processed more accurately (indexed by error rate), faster (indexed by response latency and movement duration), and more efficient (indexed by AUC and MAD). By contrast, the impact of *negative* valence was much more limited: although the average trajectories for the negative condition were slightly less curved than the 'neutral' trajectories (see right panel of Fig. 2), effects on measures of curvature (AUC and MAD) were not significant, and effects on response durations and latencies were only marginally significant. An effect of negative valence did arise in the error rates, however. Concerning response latencies, these were 38 ms faster for the positive condition, and 15 ms faster for the negative condition, than for the neutral condition. This renders the role of negative valence in word processing somewhat ambiguous, but it appears that the bulk of the effect of valence comes from the positively valenced words. A similar processing advantage for positive words only was found by Chen et al. (2015), though these authors exerted considerably less control over the lexical and sublexical characteristics of their stimuli, therefore this effect may have been driven by variables other than valence. One study with considerable control also reported positive words to be processed faster than neutral words, aligning with our present findings (Kousta et al., 2009). However, negative words also displayed a processing advantage in Kousta et al., something the present study did not find. This discrepancy may be due to the greater control of lexical variables in the present study, as previous research suggests that increased control reduces the interference effect of negative words in ESTs (Estes & Adelman, 2008; Larsen et al., 2006). Our results suggest that the influence of negative valence is similarly reduced in LDTs with increased variable control, while positive valence continues to facilitate processing.

In the following, we used the same words as in Experiment 1, but now embedded in an EST. Words were presented in one of four colours, and participants classified the colour of the word (while instructed to ignore the word itself) into one of two responses indicated by a computer mouse movement. In line with previous research (e.g., Frings et al., 2010) the experiment was designed such that we could explore both 'fast' (i.e., on-line effects on a given trial N) and 'slow' (effects emerging from the previous trial N − 1) effects of valence.

3. Experiment 2

3.1. Method

3.1.1. Participants

Twenty-six (3 male) undergraduate students from the University of Bristol participated in the study to fulfil a course requirement. The mean age was 19.5 years ($SD = 1.42$). All participants reported (corrected to) normal vision, including the absence of colour blindness. Participants also confirmed that they were native English speakers and were comfortable using a computer mouse with their right hand. Informed consent was obtained from all participants.

3.1.2. Materials and design

The 87 word stimuli (29 positive, 29 negative, 29 neutral) were those used in Experiment 1. For the EST, the standard colour palette on Microsoft Word was used to generate prototypical red, yellow, blue, and green versions of each word. Each word measured approximately 1.5 cm in height, and between 1 cm and 5 cm in width when presented in a trial.

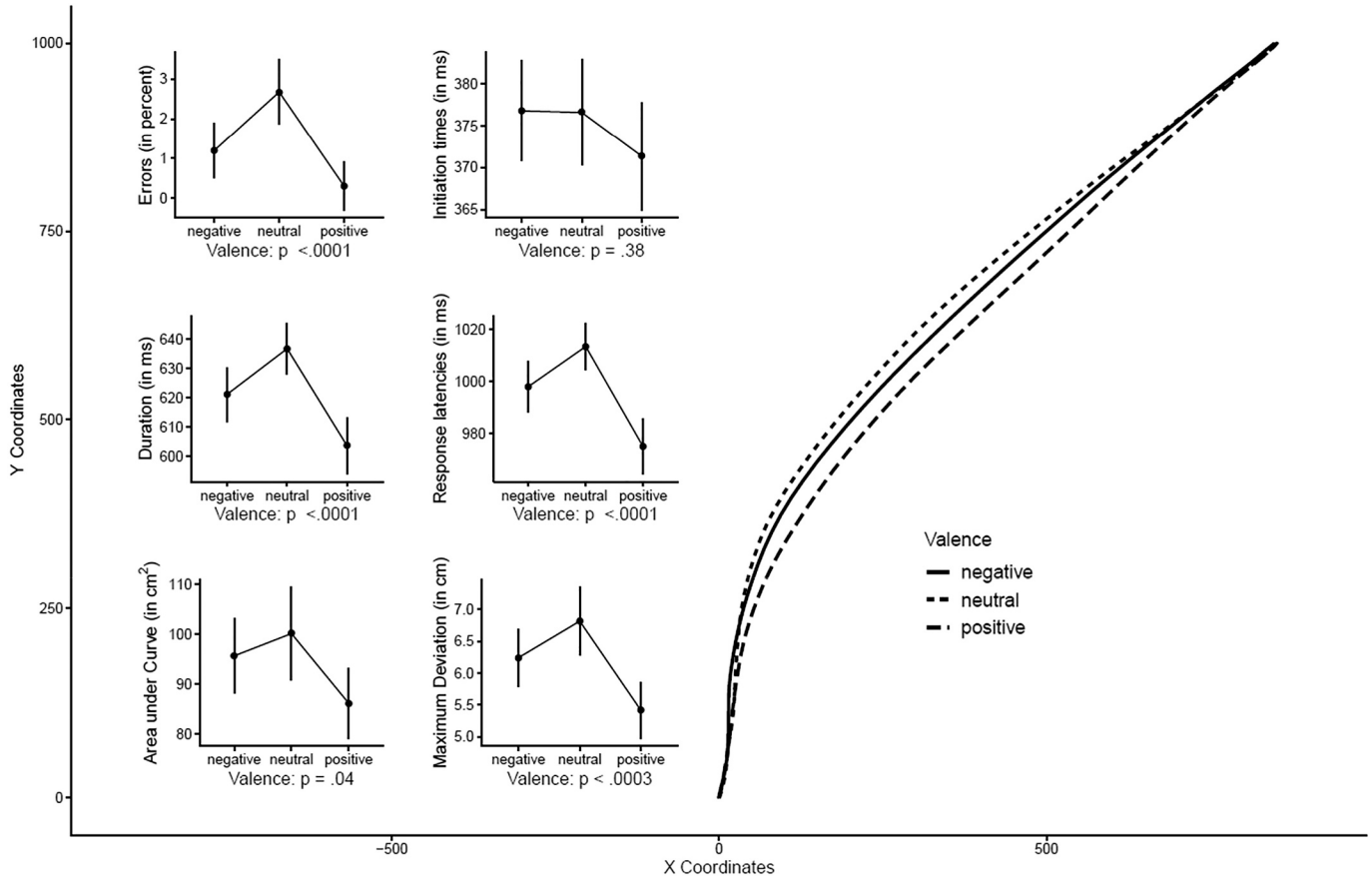


Fig. 2. Lexical decision task, dependent on word valence (negative; neutral; positive). Right side of figure: average time-normalised trajectories. Left side of figure: dependent variables (errors; movement initiation times; movement durations; response latencies; area under curve (AUC); and maximum deviation (MAD)). Error bars in inset panels reflect 95% within-participants confidence intervals (Morey, 2008). The *p* value reported underneath each inset panel is from a repeated-measures one-way analysis of variance.

The experiment was divided into 4 blocks of 87 trials, resulting in each participant completing 348 experimental trials. Stimuli were presented in a different random order generated for each block and participant. Each word was presented once per block in a different ink colour between blocks, therefore participants were exposed to all words in all available colours. This was designed to avoid any potential word-colour associations which may facilitate certain responses (e.g. the colour red may have stronger associations with the word 'injury' than 'seashore'; see Mohammad, 2011).

The overall design included valence on trial N (negative, neutral, positive) as well as Valence on trial N − 1 (negative, neutral, positive) as within-participant factors. Following the approach by Frings et al. (2010), we specifically tested for the presence of 'fast' and 'slow' effects as follows: 'fast' effects compared neutral to negative valence on trial N, while trial N − 1 was neutral. 'Slow' effects compared neutral to negative valence on trial N − 1, while trial N was neutral.

3.1.3. Procedure and apparatus

Testing took place in a university laboratory, using the same Apparatus as in Experiment 1. As in Experiment 1, following a mouse click on the START box, a stimulus immediately appeared in the centre of the screen for 2500 ms or until the participant completed their response. Participants were instructed to categorise the colour of the word as quickly and accurately as possible while trying to ignore the word itself. Responses were made by using the computer mouse to click on one of the two black response button areas on the screen (top left area: blue or green responses, top right area: red or yellow responses). The response button areas were labelled for the duration of the experiment (see Fig. 2, bottom panel, for an example trial screen). After 12 practice trials (3 words per ink colour), the four experimental blocks were presented, and participants had the opportunity for a short rest between blocks. The experimental session lasted approximately 25 min.

3.2. Results

Trials on which participants had made an error were excluded from the analysis of the other dependent variables (1.1%). We further excluded data from trials on which participants had made no response (1.9%), as well as from trials on which a participant's conditional mean on response latencies was above or below 2.5 standard deviations of the participants conditional mean (2.3%).

Fig. 3 shows the results. The right side of the figure shows averaged time-normalised mouse movement trajectories, indicating no clear effect of either valence, or valence $n - 1$. The inset panels on the left hand side of the figure present the six main dependent variables of interest, with results from a repeated measures ANOVA underneath with the factors valence and valence $n - 1$. Neither variable significantly affected any of the measures, nor was there a significant interaction. For detailed statistics see Appendix C in the Supplementary materials.

Table 2 presents response latencies and error rates for all combinations. In the context of the emotional Stroop literature, the focus is usually on the neutral vs. negative comparison; "fast effects" are captured by neutral and negative trials for which trial $n - 1$ was neutral, whereas "slow" effects are captured by neutral trials on which the valence on trial $n - 1$ was either neutral, or negative. Both critical comparisons are bolded in the Table. Regarding the fast effect, average

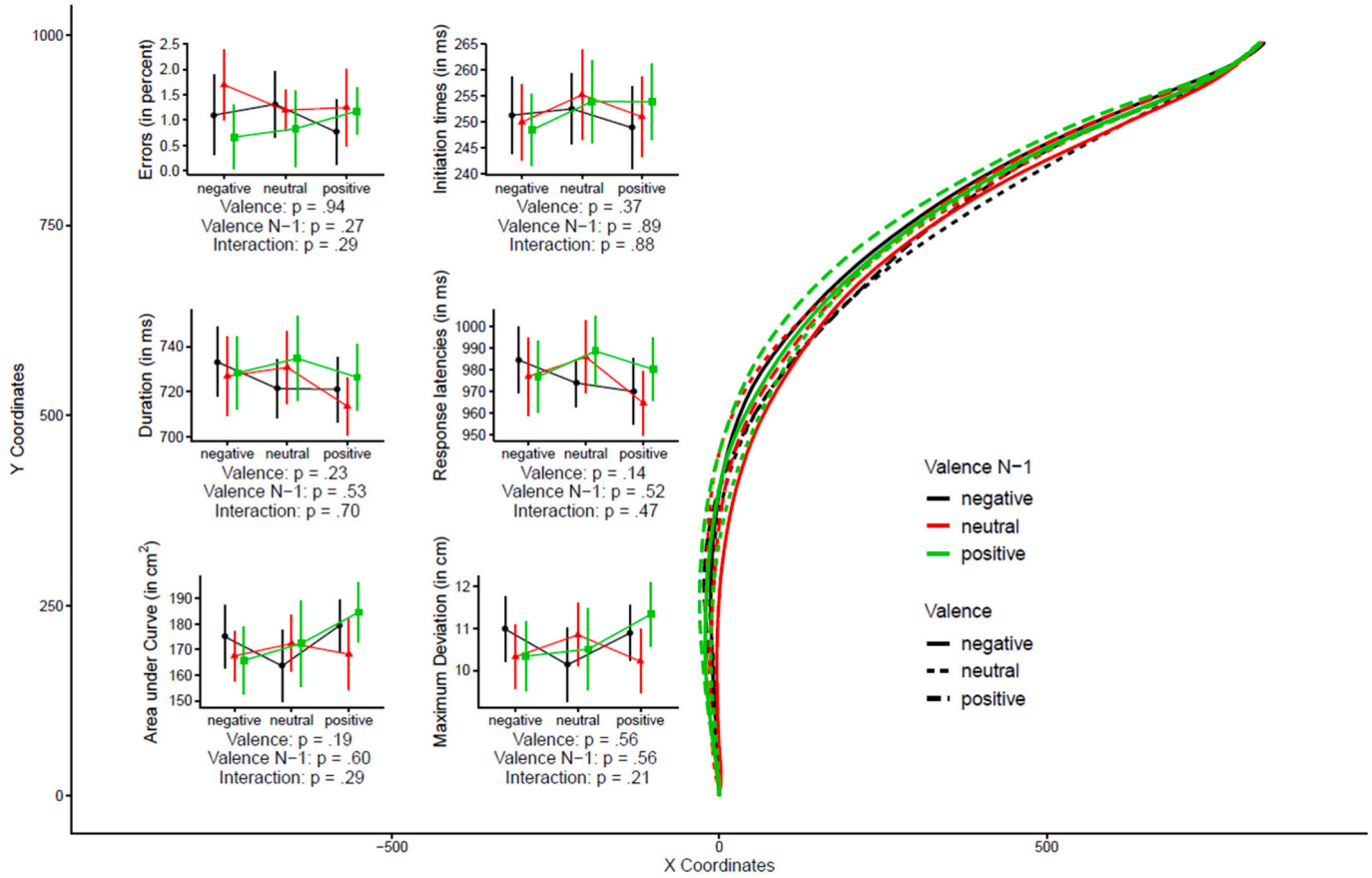


Fig. 3. Emotional Stroop task, dependent on word valence (negative; neutral; positive) and word valance on previous trial (Valence $n - 1$; negative; neutral; positive). Right side of figure: average time-normalised trajectories. Left side of figure: dependent variables (errors; movement initiation times; movement durations; response latencies; area under curve (AUC); and maximum deviation (MAD)). Error bars in inset panels reflect 95% within-participants confidence intervals (Morey, 2008). The *p* values reported underneath each inset panel are from a repeated-measures ANOVA.

Table 2
Response times (in ms) and error percentages (in parentheses), dependent on valence on trial $n - 1$, and on trial *n*. The two critical comparisons have been bolded: the "fast" valence effect is in the vertically aligned two bolded cells; the "slow" effect is in the horizontally aligned bolded cells.

	Valence *n* – 1			
Valence	Negative	Neutral	Positive	Overall
Negative	984 (1.1)	977 (1.7)	977 (0.7)	979 (1.2)
Neutral	974 (1.3)	986 (1.2)	989 (0.8)	983 (1.0)
Positive	970 (0.8)	965 (1.3)	980 (1.2)	972 (1.1)
Overall	976 (1.1)	976 (1.3)	982 (0.9)	

RTs were 9 ms faster in the negative (977 ms) than in the neutral (986 ms) condition; $t(25) = 0.58$, $p = .565$. Regarding the slow effect, average RTs were 12 ms faster in the negative (974 ms) than in the neutral (986 ms) condition; $t(25) = 0.86$, $p = .398$. Note that both effects numerically are opposite to what is typically reported in the literature (e.g., Frings et al., 2010).

The analyses described above represent a global null finding: neither valence on trial *n*, nor on trial $n - 1$, appeared to affect any of the dependent measures. To assess the evidence for the null, we additionally computed Bayes factors, using the package *BayesFactor* (Morey & Rouder, 2018). These are also reported in Appendix C. For the factor "valence", all BF_{01} were $\geq$6.9; for the factor "valence $n - 1$", all BF_{01} were $\geq$6.3, and for the interaction, all BF_{01} were $\geq$123. Hence, there is considerable evidence that valence does not affect performance in the Stroop task.

In summary, as opposed to the LDT featured in Experiment 1 in which valence affected responses (mainly driven by positive words), in the EST used in the current experiment, valence had no effect. This finding is certainly unexpected, based on a number of previous studies which had reported effects of valence in this task (mainly as interference from negative words; Williams et al., 1996; MacKay et al., 2004). One possibility which would explain this null finding is that for some reason (and contrary to conventional tasks with key press responses) participants were able to ignore the identity of the words when carrying out their colour classification responses. Perhaps some characteristic of the specific display which we used for our mouse tracking experiment, with a START button at the bottom of the screen, and response buttons at the upper edges of the screen, allowed participants to categorise the colour of a word while avoiding lexical access. If so, the null finding obtained in the results of the current experiment would not speak to whether valence really affects performance in the EST.

There is a straightforward way to test this possibility: across the three valence conditions, words were statistically matched on a range of properties (other than valence, and dominance) but there was still substantial variability on these variables. If some of these variables could be shown to affect performance in our task, this would constitute rather strong evidence that participants did indeed process the words. We tackled this analysis with two separate sets of item-based multiple

regression analyses, first for the LDT, and then for the emotional Stroop. For each response word (item), we computed average response latency, as well as a measure of response trajectory curvature (AUC), and we included a subset of the variables on which words had been matched triplet-wise across valence condition in a stepwise linear multiple regression analysis (choosing a subset of variables was necessary because some of the predictors were highly collinear, particularly so the various measures of word frequency). We included arousal, dominance, length in number of letters, orthographic N, SUBTLEX frequency, bigram frequency (type and token), age-of-acquisition, familiarity, imageability, concreteness, contextual diversity, semantic diversity, number of senses, sensory experience, and body-object interaction in the analyses. A correlation matrix for these predictors is presented in Appendix D in the Supplementary materials.

3.2.1. Lexical decision task

Item-based RTs and AUC correlated strongly with one another, $r = 0.65$, $t(85) = 7.90$, $p < .001$, $BF_{10} > 1000$, indicating that they captured at least partially overlapping variability. A stepwise linear multiple regression analysis conducted on response latencies identified as the key variables SUBTLEX frequency, familiarity, and contextual diversity, $F(3, 83) = 37.17$, $p < .001$, with a combined adjusted R^2 of 0.57. We additionally conducted a Bayesian regression analysis on these variables and found "extreme" evidence for the impact of SUBTLEX frequency and familiarity ($BF_{10} > 1000$ and 141 respectively) but only "anecdotal" evidence for contextual diversity ($BF_{10} = 1.4$). A parallel analysis conducted on AUC identified familiarity, sensory experience, and arousal as significant predictors, $F(3, 83) = 8.73$, $p < .001$, with an adjusted R^2 of 0.21. Bayesian analysis showed "extreme" evidence for familiarity ($BF_{10} = 133$), "moderate" evidence for sensory experience ($BF_{10} = 5.7$) and only "anecdotal" evidence for arousal ($BF_{10} = 1.6$). Overall, familiarity strongly affects both response latencies and AUC; frequency appears to selectively affect latencies, whereas sensory experience affects curvature.

3.2.2. Emotional Stroop

RTs and AUC correlated strongly with one another, $r = 0.34$, $t(85) = 3.35$, $p = .001$, $BF_{10} = 33.8$. In the stepwise linear regression performed on RTs, arousal and semantic diversity emerged as significant predictors, $F(2, 81) = 8.72$, $p < .001$, with an adjusted R^2 of 0.16. Bayesian analysis showed "very strong" evidence for arousal ($BF_{10} = 77.2$) but merely "anecdotal" evidence for semantic diversity ($BF_{10} = 1.7$). Specifically, RTs and arousal are negatively correlated ($r = -0.34$) such that higher arousal induces faster RTs. In the stepwise analysis conducted on AUC, arousal, imageability, semantic diversity, and number of senses were significant, $F(4, 79) = 6.22$, $p < .001$, with an adjusted R^2 of 0.20. Bayesian analysis showed "moderate" evidence for arousal ($BF_{10} = 5.9$), imageability ($BF_{10} = 9.98$), and semantic diversity ($BF_{10} = 6.97$) but merely "anecdotal" evidence for number of senses ($BF_{10} = 2.27$). AUC and arousal are negatively correlated ($r = 0.25$) such that higher arousal induces lower (straighter) trajectories.

Overall, arousal is the one predictor which clearly affected both RTs and AUC, with the latter additionally affected by imageability and semantic diversity. The broader implication of this analysis is that participants clearly were unable to suppress processing of the target words.

3.3. Discussion

Contrary to a number of previous studies which had featured the EST, in our study the valence (neutral, positive, negative) of the response words did not affect performance. This contrasts with the results from the Experiment 1 in which the valence of the same response words had clearly affected performance, mainly driven by positively valenced words. At the same time, the null finding in the EST could not be attributed to participants successfully being able to ignore the words when carrying out the colour classification: the regression analysis conducted on RTs and response trajectory curvature (AUC) showed clear evidence of a subset of the predictors affecting performance. We conclude that the possibility that valence is not a relevant variable in the EST should be taken seriously.

It could be argued that our study, which used "mouse tracking", used a novel methodology which differed in important ways from the previous studies, which had all used key presses as the dependent variable. For instance, the RTs measured with our methodology took on average 978 ms, which is substantially longer than the average of 750 ms in a comparable key-press EST reported by Frings et al. (2010). For this reason, Experiment 3 replicated Experiment 2, but now with a conventional key-press response rather than mouse-generated responses.

4. Experiment 3

4.1. Method

4.1.1. Participants

Forty-four participants (16 male) participated in the study online for a monetary reward. The mean age was 32.5 ($SD = 13.6$) years old. All participants reported (corrected to) normal vision, including the absence of colour blindness. Participants also confirmed that they were native English speakers and not fluent in any other languages. Informed consent was obtained from all participants.

4.1.2. Materials and design

The 87 word stimuli (29 positive, 29 negative, 29 neutral) used were identical to those of Experiment 1 and 2, with the standard colour palette on Microsoft word used to generate prototypical red, yellow, blue, and green ink colour copies of each word. Each word measured approximately 2 cm in height, and between 1 cm and 12 cm in width when viewed in a trial. The design of this study was also identical to Experiment 2, whereby participants viewed four blocks of 87 trials in a random order and were exposed to all words in all available colours. Analysis of fast and slow emotional Stroop effects was again enabled by adopting a 2×2 factorial design, varied within participants. The first factor was the valence in trial n (negative vs. neutral), and the second factor was valence in trial $n - 1$ (negative vs. neutral).

Data were collected via the Prolific Academic online platform (https://www.prolific.co/). Prolific is a recommended online participant recruitment platform, as participants acquired using this platform are considered to provide accurate data due to the pre-screening function which reduces the risk of dishonest participants (for reviews, see Palan & Schitter, 2017; Peer et al., 2017). This allowed us to recruit the target population accurately (e.g., only native English speakers could sign up), and restrict participation to desktop users only, thus controlling for monitor size as accurately as possible.

4.1.3. Procedure

After signing up on Prolific, participants were given information regarding the experiment, and completed the first half of a short questionnaire regarding factors that may influence cognitive performance in the task (see Appendix E in the Supplementary materials). They were then redirected to the EST, which was presented using Gorilla (https://gorilla.sc/). After completing the EST, participants finished the short questionnaire.

Participants were instructed to ignore the meaning of the presented word and report the colour as quickly and accurately as possible by pressing the 'Q' key for blue or green words, and the 'P' key for red or yellow responses. The stimuli were presented in the centre of a white background until the participant completed their response, upon which the next stimulus immediately appeared. After 12 practice trials (3 words per ink colour), the 4 experimental blocks were presented, including scheduled breaks at the end of each block which also provided a reminder of the keypress response mappings. The experimental session lasted approximately 20 min.

4.2. Results

Three participants were excluded from data analysis due to excessive error rates (>20%). One additional participant was excluded due to an average response latency close to 2 s. For the remaining 40 participants, trials on which participants had made an error were excluded from the response time analysis (3.7%). We further excluded response latencies above 2000 ms or below 200 ms (1.3%), as well as latencies above or below 2.5 standard deviations of a participant's conditional mean (3.5%).

Table 3 presents response times and errors. Repeated-measures ANOVAs conducted on the errors showed no significant effect of Valence, $F_1(2, 78) = 1.59$, $p = .21$; $F_2(2, 84) = 1.73$, $p = .18$, of Valence N − 1, $F_1(2, 78) = 1.77$, $p = .18$; $F_2(2, 168) = 1.37$, $p = .26$, nor a significant interaction, $F_1(4, 156) = 0.68$, $p = .61$; $F_2(4, 168) = 0.81$, $p = .52$. Similarly, ANOVAs conducted on response latencies showed no significant effect of Valence, $F_1(2, 78) = 0.09$, $p = .92$; $F_2(2, 84) = 0.11$, $p = .90$, of Valence N − 1, $F_1(2, 78) = 0.51$, $p = .60$; $F_2(2, 168) = 0.25$, $p = .78$, nor a significant interaction, $F_1(4, 156) = 0.78$, $p = .54$; $F_2(4, 168) = 1.04$, $p = .39$.

As for Experiment 2, we specifically tested for "fast effects" as captured by neutral and negative trials for which trial $n - 1$ was neutral, and "slow" effects are captured by neutral trials on which the valence on trial $n - 1$ was either neutral, or negative. Both critical comparisons are bolded in the Table. The "fast" effect (7 ms) was not significant, $t_1(39) = 0.84$, $p = .41$; $t_2(56) = 0.04$, $p = .97$, and neither was the "slow" effect (12 ms), $t_1(39) = 1.22$, $p = .23$; $t_2(28) = 0.80$, $p = .43$.[3]

A stepwise linear regression was performed on RTs in a manner parallel to those performed for the results from Experiment 2, with arousal, dominance, length in number of letters, orthographic N, SUBTLEX frequency, bigram frequency (type and token), age-of-acquisition, familiarity, imageability, concreteness, contextual diversity, semantic diversity, number of senses, sensory experience, and body-object interaction included in the analyses. Concreteness emerged as the single significant predictor, $F(1, 85) = 13.22$, $p < .001$, with an adjusted R^2 of 0.12. Bayesian analysis showed "very strong" evidence for an effect of concreteness on RTs ($BF_{10} = 41.4$). RT and Concreteness were positively correlated ($r = 0.37$) such that higher concreteness induced slower RTs; in other words, the more concrete a word was, the slower its colour classification responses.

Table 3
Response times (in ms) and error percentages (in parentheses), dependent on valence on trial $n - 1$, and on trial n. The two critical comparisons have been bolded: the "fast" valence effect is in the vertically aligned two bolded cells; the "slow" effect is in the horizontally aligned bolded cells.

	Valence N − 1			
Valence	Negative	Neutral	Positive	Overall
Negative	634 (3.9)	636 (3.9)	630 (4.0)	633 (3.9)
Neutral	641 (3.1)	629 (3.2)	636 (3.4)	635 (3.2)
Positive	632 (3.2)	630 (3.7)	642 (4.7)	635 (3.8)
Overall	636 (3.4)	632 (3.6)	636 (4.0)	

4.3. Discussion

Experiment 3 replicated the null finding concerning the potential effect of valence in an EST obtained in Experiment 2, but now in a conventional key-press task rather than the mouse tracking procedure employed in Experiment 2. It appears that when positively, negatively, or neutrally valenced words which are carefully matched on variables other than valence are used as the stimuli in the EST, valence has no effect on the speed and accuracy of responses. As was the case in Experiment 2, the multiple regression analysis performed on the item means allowed us to exclude the possibility that participants were genuinely able to prevent processing of the word. In accordance with the results from the analysis conducted on the item means of Experiment 1 (LDT) and 2 (EST), it appears that the predictors which powerfully affect word processing in LDTs (predominantly: familiarity, frequency, and contextual diversity) do not affect responses in the EST in the same way. In Experiment 2, the variables which strongly affected responses were arousal and semantic diversity; by contrast, in the present experiment, concreteness emerged as the only strong predictor. The reason for this discrepancy is at present unclear; however, the main point of this analysis was to demonstrate that participants had indeed processed the stimulus words. We conclude that valence by itself has no effect on performance in the EST.

5. General discussion

Our results displayed contrasting valence effects of the same response words in different experimental paradigms. In the LDT, a processing advantage for valenced over neutral words emerged, with much of the effect driven by the positively valenced words. By contrast, the valence of the same set of positive, negative, and neutral words did not affect responses in our ESTs. Regression analyses performed on item averages showed that responses in the LDT were dominated by the 'classic' lexical variables such as frequency, familiarity, etc., and conceptual variables such as contextual diversity, sensory experience, etc. had a weaker effect. Similar regressions performed on responses in the ESTs showed no effects of lexical variables, but exclusively conceptual variables to be relevant (arousal, semantic diversity, concreteness, etc.).

Our study explored the potential effects of valence by directly comparing LDT and the EST, something which is not typically done in the literature. This comparison affords insight into the origin of valence effects. As summarised in the Introduction, according to Yap and Seow (2014) valence effects in the LDT could arise either from an early, preconscious and task-general mechanism (e.g., 'automatic vigilance'; Pratto & John, 1991), or due to enhanced 'semantic richness' of valenced compared to neutral words which generates additional feedback from the semantic to the lexical level and should therefore be specific to tasks which involve lexical access. In our own results, valence had task-specific effects, i.e. it affected lexical decisions but not colour categorisations. Early attentionally-based effects could be expected to be task-general and so should have emerged in both tasks. Because this was not the case, in line with Yap and Seow we attribute the valence effects in our LDT (in our case mainly driven by positively valenced words) to feedback from the semantic to the lexical level where strings are

[3] We would like to thank an anonymous reviewer for highlighting the following aspect of our results. In the results of Experiment 2 (Table 2) it appears that responses on which in trials N − 1 and N the valence category is the same (positive-positive, etc.) appear slower than responses on which the valence category mismatched. To capture this potential aspect, we re-analysed Experiment 2 and 3 by coding for each trial whether on trial N − 1, valence was "matching" or "mismatching". In Experiment 2, responses were 10 ms slower for "matching" than for "mismatching" responses, a difference which was marginally significant, $F(1, 25) = 3.65$, $p = .068$. The difference was also present in the two measures of curvature of movement trajectories (AUC: $F(1, 25) = 4.58$, $p = .042$; MAD: $F(1, 25) = 7.71$, $p = .010$). However, in Experiment 3 (Table 3) latencies for "matching" and "mismatching" trials were 635 and 634 ms and hence virtually identical, $F < 1$, $p = .93$. An explanation for the effect of valence category match/mismatch is currently lacking, and it is not clear why the match/mismatch effect emerges in mouse tracking but not in a key press experiment.

ultimately classified as words or nonwords. By contrast, the colour classification responses carried out in the EST would not benefit from such feedback, and hence valence is irrelevant in this task.

The multiple regression analyses conducted on item averages for both tasks were originally intended to demonstrate that participants in the EST had indeed processed the words. However, the results offer some additional insight into the (null) effects of valence. Predictably, responses in the LDT were dominated by the classic lexical properties of frequency and familiarity; this is expected as effects of this type have been widely documented in the literature (e.g., Brysbaert et al., 2018). Additionally, there was weak evidence for a potential effect of sensory experience and arousal on response curvatures. In combination, these results suggest a powerful effect of lexical variables, and a weaker effect of semantic variables (among them, valence) which as argued above, likely arises from feedback from the semantic to the lexical level. By contrast, the multiple regression analysis of responses in the emotional Stroop showed no lexical effects. This is not unexpected in a Stroop task which requires manual rather than verbal colour classification. For instance, in the computational model of Stroop effects by Roelofs (2003), colours activate corresponding conceptual codes whereas words activate lexical entries. In Stroop tasks with verbal responses, colour and word representations compete for lexical selection and this competition accounts for the classic Stroop colour-word interference. By contrast, representations corresponding to manual responses are accessed directly from conceptual codes, i.e., without lexical involvement. If so, this accounts for our observation that 'lexical' variables such as frequency or familiarity did not affect colour classification responses in our EST. Presumably, participants accessed response codes corresponding to the stimulus colour directly and without "verbal mediation" (i.e., without access to the lexical entries of the colour responses). For this reason, the 'classic' lexical variables of word processing (frequency, familiarity, etc.) which emerged powerfully in the regression analyses performed on the LDT did not affect colour classification responses in the EST.

Instead, we found a substantial effect of arousal in the regression analysis of item means in the EST reported in Experiment 2, but arousal had little or no effect on performance in the LDT. Valence and arousal were confounded in many if not most earlier studies, but Kuperman et al.'s (2014) seminal contribution suggested that valence and arousal exert orthogonal effects on lexical decision (and word naming) performance measures. This constellation is puzzling: arousal – the extent to which a word is calming or exciting – could constitute one facet of a word's 'semantic richness', and as such could affect word processing in the same way that valence presumably does (see above). However, in this case effects of arousal should be task-specific to the LDT but they should not emerge in the emotional Stroop, which is the opposite of what we found. The null finding of arousal in the LDT might be a statistical artefact, however: the correlation matrix for our predictor variables (Appendix D) shows that frequency, familiarity, and arousal form a tight cluster, with $r = 0.59$ for familiarity-SUBTLEX frequency, $r = 0.39$ for arousal and SUBTLEX frequency, and $r = 0.22$ for arousal-familiarity (all $ps < .001$). As these predictors are collinear, frequency and/or familiarity dominates responses in the LDT and mask the effect of arousal. By contrast, in the EST, frequency/familiarity are irrelevant and so arousal emerges instead as a powerful statistical predictor. Hence, arousal might affect both tasks in a similar fashion (but its effect is masked in the LDT by collinearity with frequency/familiarity). If this is correct, an account could be that arousal might be attention-capturing, in the sense that was originally hypothesised for valence, e.g., in the automatic vigilance hypothesis (Pratto & John, 1991). This admittedly speculative possibility could be explicitly tested in further research, via an EST in which arousal is varied while holding valence, and the other influential variables constant.

In the LDT, the effect of valence was stronger for positive than for negative words. If our account of valence in terms of feedback from the semantic to the lexical level (see above) is accurate, a further question is why the effects of valence in the LDT are largely confined to positively valenced words. According to a 'semantic richness' view, positive and negative words alike are associated with more semantic information than neutral words, and so both should elicit more semantic feedback in lexical decisions. Positive valence has been found to facilitate word processing compared to neutral and negative valence in previous LDTs (Chen et al., 2015; Scott et al., 2014; Wentura et al., 2000). This processing advantage for positive words is typically ascribed to the general slowdown in processing associated with negative stimuli (Algom et al., 2004), which would predict delayed responses on negative compared to neutral trials. Our results did not follow this pattern and instead suggested a specific advantage for positive words, which may reflect a general positivity bias in processing (Walker et al., 2003). Such a bias could arise from the posited lower response threshold of positive stimuli, as they are less threatening than negative stimuli (Kuperman et al., 2014). These authors further note that as lexical decisions are not always made after full processing of a stimulus, this lower response threshold may facilitate responses to positive words in the LDT. Some authors have also argued that the generally positive mood of healthy participants may facilitate the processing of positive stimuli (Erickson et al., 2005). Indeed, one experimental manipulation of mood in a LDT found faster processing of positive words when participants were in an elated mood (Challis & Krane, 1988). However, these effects of mood manipulation are not always observed (Clark et al., 1983), and do not provide an explanation for the absence of a negative valence effect in Experiment 1. An alternative explanation of the positive advantage identified in Experiment 1 refers to the argument that moving one's hand towards a mouse and subsequently moving the mouse is an approach behaviour (much like a key-press response; Kissler & Koessler, 2011), and due to the interplay between perception and action (Spivey, 2007), this approach behaviour may facilitate responses to positive words in the LDT (Neumann & Strack, 2000).

Kousta et al. (2009) reported processing advantages of positively and negatively valenced words compared to neutral words. Despite controlling for several influential lexical and sublexical variables, these authors did not account for other factors including contextual diversity, which emerged as a significant predictor of responses in our LDT. This is not to say that contextual diversity specifically confounded the findings of Kousta et al., rather to point out that increased control of variables that influence word processing has previously been found to reduce the size of negative valence effects in emotional Stroop research (Estes & Adelman, 2008). Comparison of our Experiment 1 findings with Kousta et al. suggests that increased variable control also diminishes the influence of negative valence in the LDT, highlighting the importance of matched stimuli sets when investigating valence effects (Larsen et al., 2006), regardless of the experimental design.

The emotional Stroop effect is considered a robust phenomenon (Williams et al., 1996), with numerous studies documenting an interference effect of negative valence (Algom et al., 2004; MacKay et al., 2004; McKenna & Sharma, 1995). This interference was not observed in the present study, as valence had neither 'fast' nor 'slow' effects on responses in our key-press or mouse tracker versions of the EST. The absence of any valence effects in Experiments 2 or 3 may again be due to our extensive control of influential lexical and sublexical characteristics (Estes & Adelman, 2008), suggesting that the effects of valence are not powerful enough to generate the emotional Stroop effect once other confounding variables are taken into account. Thus, the present findings contribute to the continued debate on whether valence (or arousal) alone can produce the emotional Stroop effect (Schimmack, 2005; Vogt et al., 2008). Interestingly, the neutral words used in several documented emotional Stroop effects were less arousing than the emotional words (Eilola & Havelka, 2011; Frings et al., 2010; Sutton et al., 2007), and the regression analysis of Experiment 2 revealed arousal as a significant predictor of responses. Indeed, in an analysis of 12,658 words for which lexical decision and naming latency data were available from the English Lexicon Project (Balota et al., 2007), calming words were recognised faster than arousing words (Kuperman et al., 2014).

Therefore, emotional Stroop effects derived from stimuli sets which were unmatched on arousal may be confounded by this variable, resulting in faster processing of neutral (typically more calming) words compared to negative (typically more arousing) words.

Some findings of an emotional Stroop effect involved participants providing valence and/or arousal ratings for the experimental stimuli prior to colour naming (Frings et al., 2010; Frings & Wuhr, 2012). Providing these word ratings requires a deep level of conscious processing (Aycicegi-Dinn & Caldwell-Harris, 2009), which contrasts to the EST in which participants are instructed to ignore word meaning (MacKay et al., 2004). It could be argued that previous exposure and processing of the emotional words in rating tasks makes it more challenging to ignore their meaning when presented in the EST, as emotional stimuli are typically more memorable and salient than neutral stimuli (Colombel, 2000; Kensinger et al., 2002), and rating tasks increase the personal relevance of the words (Williams et al., 1996). Of course, it is unlikely that meaning is ever truly ignored in the EST, as task-irrelevant aspects of a stimulus (e.g. meaning) are automatically processed without the explicit intention to do so (Frings & Wuhr, 2012). Further, the regression analyses of Experiment 2 and 3 revealed several predictors to significantly influence performance, despite the absence of any valence effects, indicating that participants were unable to avoid processing the words. However, exposure to experimental stimuli in rating tasks may have a priming effect on colour naming, as Warren (1972, 1974) found that a semantically related prime preceding an emotional Stroop trial slows down response latency. The deeper processing of stimuli in a rating task (Winskel, 2013) may result in such priming effects when the same word is encountered in an experimental trial. Therefore, these potential priming and memory influences may make word meaning more accessible (Frings et al., 2010), thus inflating the size of valence effects in studies which incorporate a valence rating task.

The response dynamics provided by the mouse tracking software (Freeman & Ambady, 2010) permitted analysis of the real-time development of lexical decision (Barca et al., 2017) and colour naming responses (Incera et al., 2013). This novel response modality could be used to explain the discrepancy between present findings and previous lexical decision (Kousta et al., 2009), and emotional Stroop investigations (Frings et al., 2010). Indeed, mouse tracker responses are typically longer than the key-press equivalent (see Experiment 2 and 3). However, the similar absence of valence effects in both our key-press and mouse tracking versions of the EST suggests that the response modality does not change how words are processed in this task (Incera & McLennan, 2016; Sugg & McDonald, 1994). Further, the effects of variables such as word frequency on LDTs have been found to produce consistent results of a processing advantage for high frequency words, regardless of key-press (Kuchinke et al., 2007; Scarborough et al., 1977) or mouse tracker responses (Barca & Pezzulo, 2012). Based on our current and these previous findings, we argue that while consideration of response modality is important when comparing valence research, responding with a computer mouse does not alter the processing demands of the lexical decision or EST. Instead, recording mouse movement trajectories allowed us to investigate the qualitatively different way in which positive words were processed in Experiment 1, and provided six measures of evidence indicating null valence effects in the EST.

According to the 'polarity correspondence principle' (Proctor & Cho, 2006), participants in two-choice tasks tend to code the stimulus and response alternatives according to positive or negative polarity based on their relative salience. Performance is best for a mapping that maintains correspondence of the respective code polarities. For instance, de la Vega et al. (2012) asked participants to perform a valence judgement on single words, and varied the assignment of response hand to valence category. They found an interaction such that positive words were judged faster with the right than the left hand, and the reverse for negative words. This finding could potentially be explained by the polarity principle (positive words, and right hand responses, are coded as +, and the coinciding polarity leads to better performance; but see Song et al., 2017, for counter-evidence). In our LDT (Experiment 1) word and nonword responses were consistently mapped to the right and left response fields respectively, and all participants operated the computer mouse with their right hand. Could the polarity principle explain our findings regarding the effects of valence? De la Vega et al. in their first experiment reported results from a LDT (rather than from explicit valence judgments used in subsequent experiments) and found no interaction between response hand and word valence, making it unlikely to us that polarity is relevant in LDT. Furthermore, our participants confirmed in their consent procedure that they were "comfortable with operating a computer mouse with their right hand" (although we did not explicitly select for right-handers). Because most individuals are not comfortable with operating a computer mouse with their left hand, it would be difficult or impossible to rotate response hand and category assignment. We acknowledge that there is a possibility that polarity may have influenced the results, and that mouse tracking is less than ideal to investigate this issue.

A reviewer pointed out that some of our stimuli might be semantically related to one another (e.g., slave-racism) and that for this reason, semantic priming could have impacted the results, particularly for combinations within a particular valence category. To explore this possibility, we computed measures of semantic similarity for all possible stimulus combinations. At present semantic overlap is most precisely captured by prediction-based distributional models that are trained on large text corpora such as subtitles from popular film and television series (Mandera et al., 2017). We used the authors' interface (http://meshugga.ugent.be/snaut-english/) with the default (optimal) settings, and identified 83 out of 7482 pairwise stimulus combinations (1.1%) which could be considered outliers (i.e., having relatedness values outside 1.5 times the interquartile range above the upper quartile or below the lower quartile). Reanalysis of our data with these instances excluded rendered statistically equivalent results to the original analysis, giving no reason to suspect that semantic priming could have influenced the results.

As briefly alluded to in the Introduction, although our stimuli were statistically matched on a variety of potentially relevant lexical and conceptual variables, they were unmatched on 'dominance'; the confound between valence and dominance is so powerful (Warriner et al., 2013) that it is impossible to break. The potential importance of dominance is generally underplayed in the literature; it is for instance not mentioned in Kousta et al. (2009) nor in Frings et al. (2010). In our reading, a proper psychological account of dominance is currently lacking. Notwithstanding, the effects which we attributed to valence in the LDT might have also arisen from dominance. It is currently unclear how to resolve this issue. Fig. 4 shows the relationship between ratings of dominance and valence in our materials. As can be seen, valence and dominance are strongly correlated ($r = 0.85$) but within each level of the variable valence, there is still considerable residual variability on dominance. As a preliminary step to evaluate whether dominance could have affected the results, we conducted linear regressions on item-based average RTs, separately for each level of 'valence' (neutral, negative, positive). We included SUBTLEX frequency, familiarity, and critically, dominance as predictors. Under neither level of 'valence' did 'dominance' have a residual significant effect, nor did it when the analyses were conducted on average AUCs rather than RTs. Further research is required to disentangle potential effects of valence from those of dominance.

Considering the use of identical materials in our three experiments, it appears that valence exerts a different influence on word processing depending on the experimental paradigm. The EST is thought to engage an early automatic lexical level of processing (Aycicegi-Dinn & Caldwell-Harris, 2009; Winskel, 2013), yet the existence of the emotional Stroop effect suggests that task-irrelevant stimulus aspects such as word meaning and valence are also accessed (Williams et al., 1996), despite semantic processing being a higher order component of comprehension (Demonet et al., 1992). By contrast, semantics are a

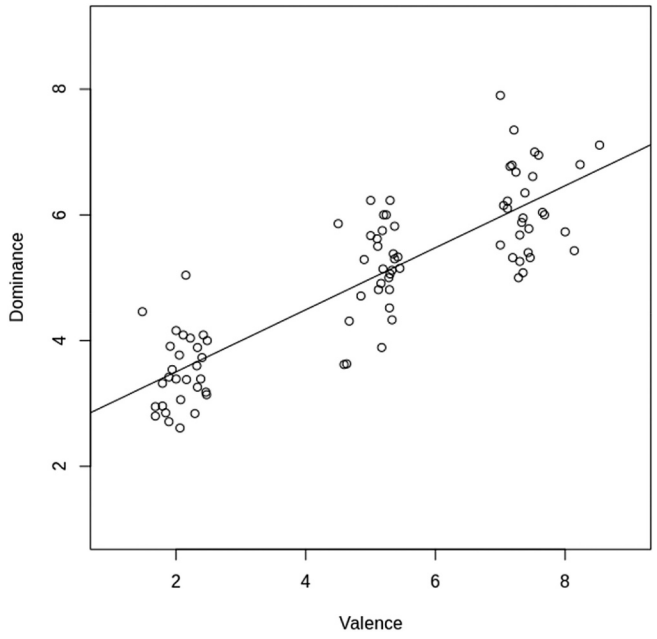


Fig. 4. Relationship between Dominance and Valence ratings in the materials used in Experiments 1–3. Both dimensions were rated on a 1–9 scale.

crucial indicator of whether the stimulus is a word or non-word in LDTs, therefore conscious access to meaning provides a useful source of information when judging the 'wordness' of a stimulus (Ratcliff et al., 2004). These differences in task demands may explain the discrepancy between the valence effects in the present study, and highlight the caution required when comparing lexical processing across different tasks. Our pattern of findings suggests that when word meaning is a task-relevant dimension of the stimulus, as in LDTs (Ratcliff et al., 2016), valence alone produces the effect of a processing advantage mostly for positive words. However, as word meaning is task-irrelevant in the EST (Pratto & John, 1991), valence alone does not produce any significant effects on word processing. Indeed, several authors have concluded that valence has a larger effect on lexical decision than naming responses (Kuperman et al., 2014; Larsen et al., 2008), which may be due to the deeper semantic access and feedback involved in LDTs (Yap & Seow, 2014). Our data align with this conclusion, and we postulate that significant effects of valence alone are constrained to tasks where valence is a relevant dimension for task success.

CRediT authorship contribution statement

Ethan Crossfield: Conceptualisation, Methodology, Formal analysis, Writing - Original draft, Writing - Review & editing, Funding acquisition. **Markus Damian:** Conceptualisation, Formal analysis, Writing - Original draft, Writing - Review & editing.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

Acknowledgment

This project was partly supported by research grant RPG-2019-054 by the Leverhulme Trust to the second author.

Appendices B-E. Stimuli used in all experiments

Neutral: audition, bone, cactus, chauffeur, crucial, doorbell, employee, fanatic, gasoline, groin, hairdo, highway, hose, kilt, mechanic, millennium, pastor, plot, radius, risk, roommate, situation, streak, stretch, territory, ticket, torch, underpants, volcano
Negative: amputation, anxiety, attack, bomb, coma, coward, criminal, crisis, death, deathbed, disease, failure, funeral, grave, greed, grief, headache, injury, intruder, jail, morgue, nausea, poison, pollution, prison, racism, sewage, sick, slave
Positive: celebration, champion, comrade, daydream, delicacy, dream, epic, fortune, hero, hug, mate, meditation, miracle, oasis, paradise, picnic, romance, rose, sanctuary, seashore, spa, summer, sunrise, sunset, sunshine, treasure, vacation, victory, wedding.

Appendices B-E. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2021.103359.

References

Abrams, R., & Balota, D. (1991). Mental chronometry: Beyond reaction time. *Psychological Science, 2*, 153–157.
Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *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.
Anooshian, L., & Hertel, P. (1994). Emotionality in free recall: Language specificity in bilingual memory. *Cognition and Emotion, 8*, 503–514.
Anwyl-Irvine, A. L., Massonnie, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioural experiment builder. *Behaviour Research Methods, 52*, 388–407.
Aycicegi-Dinn, A., & Caldwell-Harris, C. L. (2009). Emotion memory effects in bilingual speakers: A levels of processing approach. *Bilingualism: Language and Cognition, 12*, 291–303.
Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM). Linguistic Data Consortium.* Pennsylvania, USA: University of Pennsylvania.
Balota, D. A., Yap, M. J., Cortese, M. J., Hutchinson, K. A., et al. (2007). The English lexicon project. *Behaviour Research Methods, 39*, 455–459.
Barca, L., & Pezzulo, G. (2012). Unfolding visual lexical decision in time. *PLoS One, 7*(4), Article e35932. https://doi.org/10.1371/journal.pone.0035932
Barca, L., & Pezzulo, G. (2015). Tracking second thoughts: Continuous and discrete processed during visual lexical decision. *PLoS One, 10*(2), Article e0116193. https://doi.org/10.1371/journal.pone.0116193
Barca, L., Pezzulo, G., Ouellet, M., & Ferrand, L. (2017). Dynamic lexical decisions in French: Evidence for a feedback inconsistency effect. *Acta Psychologia, 180*, 23–32.
Ben-David, B. M., Chajut, E., & Algom, D. (2012). The pale shades of emotion: A signal detection theory analysis of the emotional Stroop task. *Psychology, 3*, 537–541.
Bradley, M. M. (2000). Emotion and motivation. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 602–642). New York: Cambridge University Press.
British National Corpus Consortium. (2007). In BNC XMLth (Ed.), *British National Corpus version 3*. Oxford, U.K.: Oxford University Computing Services.
Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911.
Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science, 27*(1), 45–50.
Challis, B. H., & Krane, R. V. (1988). Mood induction and the priming of semantic memory in a lexical decision task: Asymmetric effects of elation and depression. *Bulletin of the Psychonomic Society, 26*(4), 309–312.
Chen, P., Lin, J., Chen, B., Lu, C., & Guo, T. (2015). Processing emotional words in two languages with one brain: ERP and fMRI evidence from Chinese-English bilinguals. *Cortex, 71*, 34–48.
Clark, D. M., Teasdale, J. D., Broadbent, D. E., & Martin, M. (1983). Effect of mood on lexical decisions. *Bulletin of the Psychonomic Society, 21*, 175–178.
Colombel, F. (2000). The processing of emotionally positive words according to the amount of accessible retrieval cues/Traitement de concepts à connotation émotionnelle positive selon la quantité d'indices de récupération accessibles. *International Journal of Psychology, 35*, 279–286.
Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(6), 1084–1096.
Cortese, M. J., & Schock, J. (2013). Imageability and age of acquisition effects in disyllabic word recognition. *The Quarterly Journal of Experimental Psychology, 66*(5), 946–972.
Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37*, 65–70.
de la Vega, I., De Filippis, M., Lachmair, M., Dudschig, C., & Kaup, B. (2012). Emotional valence and physical space: Limits of interaction. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 375–385.

Demonet, J. F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J. L., et al. (1992). The anatomy of phonological and semantic processing in normal subjects. *Brain, 115*, 1753–1768.
Dolan, R. J. (2002). Emotion, cognition, and behaviour. *Science, 298*, 1191–1194.
Eilola, T. M., & Havelka, J. (2011). Behavioural and physiological responses to the emotional and taboo Stroop tasks in native and non-native speakers of English. *International Journal of Bilingualism, 15*, 353–369.
Eilola, T. M., Havelka, J., & Sharma, D. (2007). Emotional activation in the first and second language. *Cognition and Emotion, 21*, 1064–1076.
Erb, C. D., Moher, J., Sobel, D. M., & Song, J. H. (2016). Reach tracking reveals dissociable processes underlying cognitive control. *Cognition, 152*, 114–126.
Erb, C. D., Smith, K. A., & Moher, J. (2021). Tracking continuities in the flanker task: From continuous flow to movement trajectories. *Attention, Perception, & Psychophysics, 83*, 731–747.
Erickson, K., Drevets, W. C., Clark, L., Cannon, D. M., Bain, E. E., Zarate, C. A., Jr., Charney, D. S., & Sahakian, B. J. (2005). Mood-congruent bias in affective go/no-go performance of unmedicated patients with major depressive disorder. *American Journal of Psychiatry, 162*(11), 2171–2173.
Estes, Z., & Adelman, J. S. (2008). Automatic vigilance for negative words in lexical decision and naming. Comment on Larsen, Mercer, and Balota (2006). *Emotion, 8*, 441–444.
Evaitar, Z., & Zaidel, E. (1991). The effects of word length and emotionality on hemispheric contribution to lexical decision. *Neuropsychologia, 29*, 415–428.
Faust, N. T., Chatterjee, A., & Christopoulos, G. I. (2019). Beauty in the eyes and the hand of the beholder: Eye and hand movements differential responses to facial attractiveness. *Journal of Experimental Social Psychology, 85*, Article 103884.
Ferrand, L., Méot, A., Spinelli, E., New, B., Pallier, C., Bonin, P., … Grainger, J. (2018). MEGALEX: A megastudy of visual and auditory word recognition. *Behavior Research Methods, 50*, 1285–1307.
Ferre, P., Garcia, T., Fraga, I., Sanchez-Casas, R., & Molero, M. (2010). Memory for emotional words in bilinguals: Do words have the same emotional intensity in the first and second language? *Cognition & Emotion, 24*, 760–785.
Fox, E., Russo, R., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold attention in subclinical anxiety? *Journal of Experimental Psychology: General, 130*, 681–700.
Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 361–379.
Freeman, J., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology, 2*, 59.
Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behaviour Research Methods, 42*, 226–241.
Frings, C., Englert, J., Wentura, D., & Bermeitinger, C. (2010). Decomposing the emotional Stroop. *Quarterly Journal of Experimental Psychology, 63*, 42–49.
Frings, C., & Wuhr, P. (2012). Don't be afraid of irrelevant words: The emotional Stroop effect is confined to attended words. *Cognition and Emotion, 26*, 1056–1068.
Gaillard, R., Del Cul, A., Naccache, L., Vinckier, F., Cohen, L., & Dehaene, S. (2006). Nonconscious semantic processing of emotional words modulates conscious access. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 7524–7529.
Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behaviour Research Methods, 45*, 718–730.
Incera, S., Markis, T. A., & McLennan, C. T. (2013). Mouse-tracking reveals when the Stroop effect happens. *Ohio Psychologist, 60*, 33–34.
Incera, S., & McLennan, C. T. (2016). Mouse tracking reveals that bilinguals behave like experts. *Bilingualism: Language and Cognition, 19*, 610–620.
Johnson, K., Waugh, C. E., & Fredrickson, B. L. (2010). Smile to see in the forest: Facially expressed positive emotions broaden cognition. *Cognition and Emotion, 24*, 299–321.
Jonczyk, R., Boutonnet, B., Musial, K., Hoemann, K., & Thierry, G. (2016). The bilingual brain turns a blind eye to negative statements in the second language. *Cognitive, Affective, and Behavioural Neuroscience, 16*, 527–540.
Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S., & Gullick, M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *Quarterly Journal of Experimental Psychology, 64*, 1683–1691.
Kahan, T. A., & Hely, C. D. (2008). The role of valence and frequency in the emotional Stroop task. *Psychonomic Bulletin & Review, 15*, 956–960.
Kanske, P., & Kotz, S. A. (2007). Concretneess in emotional words: ERP evidence from a hemifield study. *Brain Research, 1148*, 138–148.
Kensinger, E. A., Brierley, B., Medford, N., Growdon, J. H., & Corkin, S. (2002). Effects of normal aging and Alzheimer's disease on emotional memory. *Emotion, 2*, 118–134.
Kieslich, P. J., Henninger, F., Wulff, D. U., Haslbeck, J. M. B., & Schulte-Mecklenbeck, M. (2019). Mouse-tracking: A practical guide to implementation and analysis. In M. Schulte-Mecklenbeck, A. Kühberger, & J. G. Johnson (Eds.), A handbook of process tracing methods (pp. 111–130). New York, NY: Routledge.
Kissler, J., & Koessler, S. (2011). Emotionally positive stimuli facilitate lexical decisions – An ERP study. *Biological Psychology, 86*, 254–264.
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.
Krueger, B. I., & Storkel, H. L. (2020). Childrens response bias and identification of misarticulated words. *Journal of Speech, Language, and Hearing Research, 63*, 259–273.
Kuchinke, L., Vo, M. L. H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decision vary with word frequency but not emotional valence. *Internal Journal of Psychophysiology, 65*, 132–140.
Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General, 143*, 1065–1081.
Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990.
Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review, 97*, 377–395.
Larsen, R. J., Mercer, K. A., & Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion, 6*, 62–72.
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*, 445–452.
Lazarus, R. S. (1991). *Emotion and adaption*. Oxford, UK: Oxford University Press.
Lenth, R. V. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.4.5 https://CRAN.R-project.org/package=emmeans.
Liu, X., Yang, Y., Jiang, S., & Li, J. (2018). The facilitating effect of positive emotions during an emotional Stroop task. *Neuroreport, 29*, 883–888.
MacKay, D. G., Shafto, M., Taylor, J. K., Marian, D. E., Abrams, L., & Dyer, J. R. (2004). Relations between emotion, memory, and attention: Evidence from taboo Stroop, lexical decision, and immediate memory tasks. *Memory and Cognition, 32*, 474–488.
Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.
McKenna, F. P., & Sharma, D. (1995). Intrusive cognitions: An investigation of the emotional Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1595–1607.
McKenna, F. P., & Sharma, D. (2004). Reversing the emotional Stroop effect reveals that it is not what it seems: The role of fast and slow components. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 382–392.
Mergen, F., & Kuruoglu, G. (2017). A comparison of Turkish-English bilinguals processing of emotion words in their two languages. *Eurasian Journal of Applied Linguistics, 3*, 89–98.
Mohammad, S. (2011). Colourful language: Measuring word-colour associations. In *Proceedings of the 2nd Workshop on Cognitive Modelling and Computational Linguistics* (pp. 97–106).
Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General, 118*, 43–71.
Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology, 4*, 61–64.
Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for common designs*. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor.
Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 116–133.
Neumann, R., & Strack, S. (2000). Approach and avoidance: The influence of proprioceptive and exteroreceptive cues on encoding of affective information. *Journal of Personality and Social Psychology, 79*(1), 39–48.
Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130*, 466–478.
Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinchart, and Winston.
Paivio, A. (2013). Dual coding theory, word abstractness, and emotion: A critical review of Kousta et al. (2011). *Journal of Experimental Psychology: General, 142*, 282–287.
Palan, S., & Schitter, C. (2017). Prolific.ac - A subject pool for online participants. *Journal of Behavioural and Experimental Finance, 17*, 22–27.
Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behaviour research. *Journal of Experimental Social Psychology, 70*, 153–163.
Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review, 15*(1), 161–167.
Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*, 380–391.
Princeton University. (2010). About WordNet. Retrieved from https://wordnet.princeton.edu/.
Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin, 132*, 416–442.
R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111*(1), 159–182.
Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*, 260–281.
Richards, A., French, C. C., Johnson, W., Naparstek, J., & Williams, J. (1992). Effects of mood manipulation and anxiety on performance of an emotional Stroop task. *British Journal of Psychology, 83*, 479–491.
Roelofs, A. (2003). Goal-referenced selection of verbal action: modeling attentional control in the Stroop task. *Psychological Review, 110*(1), 88–125.
Sadoski, M., & Paivio, A. (2001). *Imagery and text: A dual coding theory of reading and writing*. Mahwah, NJ: Erlbaum.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 1–17.
Schimmack, U. (2005). Attentional interference effects of emotional pictures: Threat, negativity, or arousal? *Emotion, 5*, 55–66.
Schoemann, M., Lüken, M., Grage, T., Kieslich, P. J., & Scherbaum, S. (2019). Validating mouse-tracking: How design factors influence action dynamics in intertemporal decision making. *Behavior Research Methods, 51*, 2356–2377.
Scott, G. G., O'Donnell, P. J., & Sereno, S. C. (2014). Emotion words and categories: Evidence from lexical decision. *Cognitive Processing, 15*(2), 209–215.
Shaoul, C., & Westbury, C. (2010). *Neighborhood density measures for 57,153 English words*. Edmonton, AB: University of Alberta. downloaded from http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.arcs.ncounts.html.
Siakaluk, P. D., Pexman, P. M., Aguilera, L., Owen, W. J., & Sears, C. R. (2008). Evidence for the activation of sensorimotor information during visual word recognition: The body–object interaction effect. *Cognition, 106*(1), 433–443.
Siakaluk, P. D., Pexman, P. M., Sears, C. R., Wilson, K., Locheed, K., & Owen, W. J. (2008). The benefits of sensorimotor knowledge: Body-object interaction facilitates semantic processing. *Cognitive Science, 32*, 591–605.
Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *afex: Analysis of factorial Experiments*. R package version 0.26-0 https://CRAN.R-project.org/package=afex.
Song, X., Chen, J., & Proctor, R. W. (2017). Role of hand dominance in mapping preferences for emotional-valence words to keypress responses. *Acta Psychologica, 180*, 33–39.
Spivey, M. (2007). *The continuity of mind*. USA: Oxford University Press.
Stadthagen-Gonzalez, H., & Davis, C. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behaviour Research Methods, 38*, 598–605.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.
Sugg, M. J., & McDonald, J. E. (1994). Time course of inhibition in color-response and word-response versions of the Stroop task. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 647–675.
Sutton, T. M., Altarriba, J., Gianico, J. L., & Basnight-Brown, D. M. (2007). The automatic access of emotion: Emotional Stroop effects in Spanish-English bilingual speakers. *Cognition and Emotion, 21*, 1077–1090.
Tomlinson, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language, 69*, 18–35.
Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology, 67*, 1176–1190.
Vinson, D., Ponari, M., & Vigliocco, M. (2013). How does emotional content affect lexical processing? *Cognition and Emotion*, (4), 737–746.
Vogt, J., De Houwer, J., Koster, E. H. W., Van Damme, S., & Crombez, G. (2008). Allocation of spatial attention to emotional stimuli depends upon arousal not valence. *Emotion, 8*, 880–885.
Walker, W. R., Skowronski, J. J., & Thompson, C. P. (2003). Life is pleasant—And memory helps to keep it that way! *Review of General Psychology, 7*, 203–210.
Warren, R. E. (1972). Stimulus encoding and memory. *Journal of Experimental Psychology, 94*, 90–100.
Warren, R. E. (1974). Association, directionality, and stimulus encoding. *Journal of Experimental Psychology, 102*, 151–158.
Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for over 13,915 English lemmas. *Behaviour Research Methods, 45*, 1191–1207.
Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach and avoidance-related social information. *Journal of Personality and Social Psychology, 78*, 1024–1037.
Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*, 3–24.
Winskel, H. (2013). The emotional Stroop task and emotionality rating of negative and neutral words in late Thai-English bilinguals. *International Journal of Psychology, 48*, 1090–1098.
Yamamoto, N., Incera, S., & McLennan, C. T. (2016). A reverse Stroop task with mouse tracking. *Frontiers in Psychology, 7*, 670.
Yap, M. J., & Seow, C. S. (2014). The influence of emotion on lexical processing: Insights from RT distributional analysis. *Psychonomic Bulletin & Review, 21*, 526–533.
Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*, 971–979.
Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist, 39*, 117–123.
Zeelenberg, R., Wagenmakers, E., & Rotteveel, M. (2006). The impact of emotion on perception: Bias or enhanced processing? *Psychological Science, 17*, 287–291.