Emotion
2006, Vol. 6, No. 1, 62–72

1528-3542/06/$12.00 DOI: 10.1037/1528-3542.6.1.62

Lexical Characteristics of Words Used in Emotional Stroop Experiments

Randy J. Larsen, Kimberly A. Mercer, and David A. Balota
Washington University

Validity of the emotional Stroop task hinges on equivalence between the emotion and the control words in terms of lexical features related to word recognition. The authors evaluated the lexical features of 1,033 words used in 32 published emotional Stroop studies. Emotion words were significantly lower in frequency of use, longer in length, and had smaller orthographic neighborhoods than words used as controls. These lexical features contribute to slower word recognition and hence are likely to contribute to delayed latencies in color naming. The often-replicated slowdown in color naming of emotion words may be due, in part, to lexical differences between the emotion and control words used in the majority of such studies to date.

The emotion Stroop task is widely used in the study of attentional bias to emotional words, as well as in the study of individual differences in those biases. In this task, participants are asked to name the color of words—both emotional and control words—while ignoring the semantic meaning of the words. Although ostensibly similar to the original color Stroop task (Stroop, 1935), the emotional Stroop task (for a review, see Williams, Mathews, & MacLeod, 1996) differs in several important respects. The color Stroop task uses color words printed in colors, whereas the emotional Stroop task uses words with some emotional connotation, such as threatening words, printed in colors (Williams et al., 1996). To assess attentional bias toward emotional words, researchers calculate the mean reaction time to name the colors of the emotional words and subtract from it the mean reaction time to name the colors of unemotional control words (Pratto & John, 1991).[1] Researchers use the term *interference* to describe the cognitive process that occurs within the subject during the emotional Stroop task.[2] The emotional content of the word is said to interfere with color naming or to grab attention, causing the person to be slower to name the colors of those words compared with the neutral control words (Wentura, Rothermund, & Bak, 2000).

One common interpretation of both the original color Stroop and the emotional Stroop involves the notion that people cannot ignore the semantic meaning of isolated words. That is, when a word appears on a screen, people recognize it as a word and automatically access its semantic meaning. Although there is some debate whether the emotional and color Stroop tasks fulfill all the conditions of automatic word processing (Besner & Stolz, 2000), there is no doubt that the color Stroop has been very influential in helping researchers understand the processes involved in selection of color pathways over more prepotent lexical pathways (see MacLeod, 1991, for a review).

Another important difference between the emotional and color Stroop tasks lies in how interference is measured. The measurement model used should reflect the theoretical meaning of the underlying construct. With the color Stroop, the measurement model is straightforward and face valid. In this task, the words are all color words, such as BLUE, GREEN, or RED. The display color is varied across trials, so that sometimes the trial is congruent (e.g., RED written in red ink) and sometimes it is incongruent (e.g., RED written in blue ink). Interference is measured by subtracting congruent from incongruent reaction times. The participants must ignore the same set of words (RED, BLUE, GREEN) across these trial types but must name different colors. In other words, in the original color Stroop, the words that are presumably unintentionally read are identical across congruent and incongruent trial types. This ensures that any differences between congruent and incongruent trials cannot be due to lexical differences in the stimulus words (because they are the same). Therefore, any ob-

Randy J. Larsen, Kimberly A. Mercer, and David A. Balota, Department of Psychology, Washington University.

The research reported in this article was supported by Grant RO1-MH63732 from the National Institute of Mental Health to Randy J. Larsen.

Correspondence concerning this article should be addressed to Randy J. Larsen, Department of Psychology, Campus Box 1125, One Brookings Drive, Washington University, St. Louis, MO 63130. E-mail: rlarsen@wustl.edu

[1] Another major difference between the emotional and color Stroop tasks is that, for the color Stroop, interference can be calculated at the level of the item, whereas for the emotional Stroop task, interference is calculated at the level of the list. As discussed by Algom, Chajut, and Lev (2004), imagine the item *blue*, which could appear in a congruent color (blue) or an incongruent color (red). The difference in color-naming speed between these two possible colors would yield an *item-specific* interference effect. For the emotional Stroop, item-specific interference cannot be calculated because there are no congruent conditions. For example, the word *murder* is not associated with, or semantically congruent with, any particular color. Instead, mean reaction time to name the colors of a list of emotion words are compared with the mean reaction time to name the colors of a list of control words. One important implication of this is that the semantic conflict (in the incongruent condition) and agreement (in the congruent condition) that form the basis of the color Stroop effect is absent from the emotional Stroop phenomenon. In this sense, the emotional Stroop task is not really a Stroop task at all.

[2] Searching through the 32 studies used in this report, we found that all but 4 explicitly used the term *interference* to refer to the difference in reaction time between emotional and control words. We found that a number of synonyms were also used, such as *attentional bias, attentional capture, automatic vigilance, inhibition, selective processing, disruption of performance*, *selective attention, intrusion, hypervigilance, impaired color naming,* and *color-naming decrement.*

served interference must be due to differences in the color's being congruent or incongruent with the semantic meaning of the word, because the same words appear in the congruent and incongruent lists.

The characteristics of the emotional Stroop task are quite different, making the interpretation of interference scores ambiguous.[3] Words in the emotion and control lists are *never* the same, yet the calculation of interference proceeds according to the same measurement model, by subtracting the response times to the control words from response times to the emotional words. The words in the emotion and control lists are *always* different, thereby making the emotional Stroop task a quasi-experimental paradigm. Due to the quasi-experimental nature of the task, it is crucial that the emotional and control words be carefully matched on all lexical features that influence word recognition. If, for example, the emotional words were longer, then any slowing in reaction time to name their colors might be due, in part, to the additional visual processing time imposed by the more complex stimulus words.

Burt (2002) demonstrated that a critical variable in word recognition tasks—word frequency—also influences color-naming latency. In particular, low-frequency words produce longer color-naming latencies than high-frequency words do. Burt (2002) viewed this finding as consistent with a capacity model in which the slowdown to process the low-frequency words spills over to color-naming performance. Thus, it appears that variables that influence lexical processing can indeed modulate color-naming performance. The purpose of the present study is to evaluate words used in the emotional Stroop literature to investigate whether they do indeed conform to this assumption of lexical equivalence across word categories.

Word Recognition and the Lexical Features of Words

Psychologists have long been studying the lexical features of words that contribute to word recognition (Cattell, 1886). Many lexical properties of words have been evaluated, such as word frequency, familiarity, and length in letters (see Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 1994). In word recognition research, two tasks have become the gold standard against which to study the effects of lexical characteristics on word recognition. These tasks are lexical decision latency and word-naming speed. In lexical decision tasks, participants are presented with a string of letters, and their task is to decide as quickly as possible whether that string represents a word or a nonword. In speeded naming, participants are simply presented with a word and asked to say the word aloud as quickly as possible. These tasks presumably provide a window into the processes involved in word recognition, and consequently, any lexical feature of the word that influences word recognition (e.g., its length in letters, its frequency of use) will influence lexical decision speed and/or naming speed. Such features also influence the color-naming latency of words (e.g., Burt, 1999, 2002). For this reason, the word recognition literature is important for understanding the emotional Stroop phenomenon. Anything that influences word recognition speed could potentially influence color-naming speed, and hence influence the magnitude of interference detected with the emotional Stroop procedure.

Which lexical characteristics of words most strongly influence word recognition? In predicting lexical decision time, the frequency with which a word is used in discourse appears to be a strong correlate (Balota et al., 2004). There are many ways to assess frequency of use, and several researchers have published norms based on various criteria. However, there is considerable variability across the different indicators of word frequency in terms of their ability to predict lexical decision times (Burgess & Livesay, 1998; Zevin & Seidenberg, 2002). One of the most commonly used measures of word frequency is the set of norms published by Kučera and Francis (1967). These norms are derived from a corpus of 1,014,000 words from a wide variety of American English texts. We used the Kučera and Francis norms in our study as one indicator of frequency. However, these norms are almost four decades old, and the Kučera and Francis norms showed the smallest correlations with lexical decision speed among a set of frequency norms examined by Balota et al. (2004).

Lund and Burgess (1996) have provided a more recent set of frequency norms, used in their work on the Hyperspace Analogue to Language (HAL). These norms are based on approximately 131 million words gathered across 3,000 Usenet newsgroups in February 1995. The HAL norms are stronger predictors of lexical decision time than the Kučera and Francis norms (Balota et al., 2004), and so in this study we also used the HAL norms as an index of word frequency. Because the HAL norms are not normally distributed, we also used the log transform of this index. In summary, frequency of use is an important feature to consider because infrequently used words take longer to recognize than frequently used words. Consequently, if emotional words used in Stroop studies are more infrequent than the control words, then estimates of interference would be spuriously high for those words.

Another lexical feature of words that influences recognition speed is length in letters (Balota et al., 2004). Although word length is a stronger predictor of word-naming latency than lexical decision speed is, word length nevertheless appears to play an important and obvious role in word recognition. In general, longer words take more time to process than shorter words. If emotional words used in emotional Stroop tasks are longer than control words, then we would expect spuriously high interference estimates for those words because of this lexical difference.

A third lexical feature related to word recognition speed is orthographic neighborhood size. This feature refers to the number of words into which a single word can be transformed by changing one letter in the word while preserving the identity and position of the other letters (Coltheart, Davelaar, Jonasson, & Besner, 1977). Words that have larger orthographic neighborhoods tend to produce faster response latencies in some studies using a lexical decision task (Andrews, 1989, 1992; Forster & Shen, 1996), al-

[3] There is some disagreement about the nature of interference effects found with the emotional Stroop task. Although all researchers agree that generic slowing can be found on threatening compared with control words, some argue that the effect is due to an automatic vigilance mechanism that captures cognitive resources when threat is detected in the perceptual stream (Algom et al., 2004). Other researchers imply that the generic slowdown is due specifically to cognitive conflict or response inhibition caused by the threat value of the word (e.g., see review in Williams et al., 1996). For any theoretical explanation to have credibility, however, researchers must rule out the alternative explanation that the threat words differ in other ways (i.e., lexical features related to word recognition) from the control words.

though this effect appears moderated by word frequency (see Balota et al., 2004). Turning to naming latencies, there are more consistent facilitatory effects of orthographic neighborhood size, and this effect again is larger for low-frequency words (Andrews, 1997). One hypothesized explanation is that words with large orthographic neighborhoods contain more semantic links in memory, thereby facilitating processing speed. Consequently, we would predict that if emotional words used in emotional Stroop tasks have smaller orthographic neighborhoods than the control words, then higher emotional interference estimates would be obtained.

The goal of our research is to evaluate words used in emotional Stroop studies in terms of these three important lexical characteristics: word frequency, length, and orthographic neighborhood. To do this, we took a large and systematic sample of words from published emotional Stroop studies and compared the different word categories (e.g., negative/threatening, disorder specific, positive, and neutral) in terms of these three characteristics. If indeed the word categories were found to differ, especially the negative words and the control words (which make up the bulk of emotional Stroop studies), then we controlled for the lexical characteristics to see if any differences in lexical decision latency and naming speed remained once lexical differences were accounted for.

An important question to consider is, are tasks that require word recognition (lexical decision, naming) fundamentally different from the color-naming task used in the emotional Stroop paradigm? Most researchers (reviewed in Williams et al., 1996) endorse the belief that the emotional Stroop task works, to the extent that it works (i.e., produces emotional interference) at all, because words are recognized, that is, that semantic meaning is accessed during the task. If the emotional meaning of a word slows down color naming, participants are accessing the semantic meaning during the task. Now, this process may be automatic (i.e., outside of awareness, fast, and without control), but nevertheless the semantic meaning must get into the cognitive system (through word recognition) in order for the emotional Stroop task to have an effect. Because the emotional Stroop task relies on word recognition, traditional word recognition tasks (lexical decision, naming) can be used as proxies. This point was demonstrated by Algom, Chajut, and Lev (2004), who found interference effects from the same set of words in color-naming, word-reading, and lexical decision speed within the same set of experiments.

Our study of the lexical characteristics of words used in this literature is important for several reasons. First, it provides a detailed lexical evaluation of the word stimuli on which the emotional Stroop phenomenon is based. The validity of the emotional Stroop phenomenon, especially the calculation of interference scores, hinges crucially on the lexical equivalence between the emotional words and the control words used. If the word lists are unbalanced in a way that promotes longer recognition time for the emotional words than the control words, then interference scores are contaminated by invalid variance. Even if the lexical effects are small, they will accumulate in the calculation of interference effects. The standard way of assessing interference in this literature (as the sum of the differences between reaction times to emotional words minus reaction times to the control words) aggregates any reaction time differences in recognition speed (some of which may be due lexical differences) into the interference score.

One way to think about this is that the interference score contains two components, one due to any true or reliable reaction time difference between emotional and control words and one due to the lexical difference between emotional and control words. The component related to lexical differences between the word lists would represent invalid variance that accumulates in the interference score. Because interference scores may be thus contaminated, any estimate of the emotional effect based on unbalanced word lists would be spuriously high.

A second reason why this study is important is that even if emotional and control words differ on crucial lexical characteristics related to word recognition, we can nevertheless estimate the true size of the emotional Stroop effect by controlling for those lexical characteristics and investigating whether any response speed or accuracy differences remain. Finally, the study is important because it points the way for future researchers to construct emotional Stroop word lists in ways that may eliminate this threat to the internal validity of their experiments.

Method

Selection of Words Used in Emotional Stroop Studies

We conducted a literature search of the PsycINFO database, using the keywords *emotion* and *Stroop*, and limited our results to peer-reviewed journal articles in the English language that reported empirical research. In addition, we conducted an ancestry search to obtain any articles missed by the search engine. This resulted in a total of 72 empirical publications that covered the emotional Stroop task. Of these, 32 of the articles provided lists of the actual words used in the research. A list of these 32 papers is provided in the Appendix.

From these 32 papers, a total of 1,401 words were used in the reported research. However, some words were repeated across studies. After eliminating redundancies, we obtained a total of 1,033 unique words. These words formed the basis of the data set used in this study.

We coded the words on a number of dimensions. Most important, we wanted to divide words into categories relevant to this research area. Most of the studies have a negative word list and a neutral or control word list. Some studies include positive words as well as control words. Consequently, for each word we simply adopted the authors' designation of positive, neutral, or negative. Seven of the 32 articles investigated disorder-specific words in special populations, and so included words that were unusual. Consequently, in these few cases, the words used may not have a negative valence for all participants. For example, in a study of snake phobia, words such as *copperhead* and *cottonmouth* were deemed negative by the authors. Another study of rape victims used trauma-specific words, such as *intercourse* and *vagina*. In these cases, where words were selected for disorder-specific purposes, we assigned a fourth word code: *disorder specific*. In our total sample of 1,033 words, 322 were coded as negative, 393 were coded as neutral, 240 were coded as positive, and 78 were coded as disorder specific. We will provide a list of these words on request.

Lexical and Reaction Time Characteristics for the Emotional Stroop Words

Balota and colleagues (Balota et al., 2002) have assembled a large, searchable database containing lexical characteristics and both naming and lexical decision time for a large corpus of words. This project is called the English Lexicon Project (ELP) and is available online at http://elexicon.wustl.edu/default.asp. The ELP provides normative data for visual speeded naming and lexical decision latency for 40,481words obtained across a large group of participants at six universities. At the time of our use, the ELP database was based on 2,752,698 reaction time measurements obtained from 816 participants in the lexical decision reaction time portion of the project. Also collected were 1,120,820 experimental measurements

from 442 participants in the naming reaction time part of the project. The database also contains descriptive characteristics for each of the words.

We submitted the 1,033 Emotional Stroop words to the ELP search engine, which found exact matches on all of our original words. The word, word category code (i.e., negative, disorder specific, positive, and neutral), and study information from our emotional Stroop database were then merged with the lexical and reaction time data on each of these words. This list of 1,033 words and their associated lexical characteristics and reaction time data form the data set used in this study.

Results

The means and standard deviations, calculated across the 1,033 words, are presented in Table 1 for all the lexical and reaction time data associated with the words in each of the four word categories. Analyses proceeded by analyzing the reaction time data to see if the word categories differed on these parameters. Of primary interest was testing whether the negative words produced slower lexical decision and/or naming latencies than the control words, a pattern recently reported by Algom et al. (2004). Next we tested whether the word categories differed with respect to important lexical characteristics. Finally, we examined word category differences in the behavioral data (lexical decision and naming speed) after controlling for the effects of all significant lexical differences among the word categories. This last analysis allowed us to estimate the true size of emotional Stroop effects after imposing lexical equivalence through the use of analysis of covariance (ANCOVA).

Behavioral Differences Associated With Word Valence

We used a series of one-way analyses of variance (ANOVAs) to test the hypothesis that the positive, negative, neutral, and disorder-specific words would be different on the reaction time and accuracy measures of lexical decision and naming speed. These analyses revealed overall differences among the word types on the following characteristics: lexical reaction time, $F(3, 1029) = 5.53$, $p < .01$; lexical decision accuracy, $F(3, 1029) = 5.84$, $p < .01$; and naming reaction time, $F(3, 1029) = 4.31$. Speeded naming accuracy did not differ as a function of word category.

We next used Tukey's least significant difference (LSD) test to test specific planned contrasts. With regard to lexical decision speed, the negative words produced significantly slower reaction times than the neutral words produced, replicating Algom et al. (2004) and conceptually replicating the generic slowdown in processing negative words found in the literature on the emotional Stroop (e.g., Pratto & John, 1991). In addition, the disorder-specific words produced significantly slower reaction times than all other word categories in terms of lexical decision speed. As for lexical accuracy, the disorder-specific words had significantly lower accuracy rates than all the other word categories. Regarding naming speed, the negative words took significantly longer to name than the neutral words, and the disorder-specific words produced significantly slower naming speeds than all other word categories. Taken together, these findings, which are from an independent database with regard to lexical decision speed and naming speed, are consistent with the original studies that typically found color-naming speed differences between negative and control words. Our results support Algom et al.'s (2004) conclusion that negative words are associated with a generic slowdown in processing, regardless of whether that slowdown is assessed with color naming, lexical decision, or speeded naming. However, as we argued in the introduction, such findings could be due, at least in part, to lexical differences among the word categories.

Lexical Differences Among the Word Categories

We now report the results from a series of one-way ANOVAs to test the hypothesis that the positive, negative, neutral, and disorder-specific words are different on a variety of lexical characteristics. These analyses revealed overall differences among the word categories on the following characteristics: length, $F(3, 1029) = 4.90$, $p < .01$; three measures of frequency, including frequency as normed according to Kučera and Francis (1967), $F(3,$

Table 1
Means (and Standard Deviations) on Lexical Characteristics and Behavioral Data on 1,033 Unique Words Used in 32 Published Emotion Stroop Studies, Broken Down by Word Type

Measure	Negative words	Neutral words	Positive words	Disorder-specific words[a]
Lexical RT (in ms)	675 (89)	664 (85)	664 (90)	706 (108)
Lexical accuracy (% correct)	.95 (.07)	.94 (.10)	.97 (.05)	.91 (.14)
Naming RT (in ms)	685 (79)	672 (74)	675 (78)	703 (91)
Naming Accuracy (% correct)	.98 (.04)	.97 (.05)	.98 (.04)	.97 (.06)
Length (in letters)	6.69 (2.17)	6.23 (1.87)	6.82 (2.11)	6.56 (2.28)
Frequency (K-F)	27.76 (45.67)	61.97 (149.80)	46.21 (91.57)	22.43 (37.11)
Frequency (HAL)	10,839 (19,111)	30,935 (103,312)	23,585 (81,632)	9,426 (21,842)
Frequency (log HAL)	8.15 (1.69)	8.69 (1.95)	8.78 (1.57)	7.67 (1.82)
Orthographic neighborhood	2.38 (3.98)	3.02 (4.26)	2.28 (3.38)	2.53 (4.03)

Note. RT = reaction time; K-F = normed according to the method of Kučera and Francis (1967); HAL = normed according to the Hyperspace Analogue to Language project (see Lund & Burgess, 1996).
[a] Includes words from 7 of the 32 reports: (a) phobia-specific words used in study samples of persons with that phobia, such as snake words (*cobra, viper*), (b) trauma-specific words, such as sex words for rape victims (*penis, rape, intercourse*), (c) sexual abuse words (*incest, cruel*), (d) illness words (*vomit, bleed, virus*), (e) depression words (*defeat, reject, hopeless*), (f) speech anxiety words (*stutter, audience, ridicule*), and (f) body-dysmorphic words (*ugly, repulsive, disfigured*).

962) = 6.75, $p < .01$; frequency as normed by Lund and Burgess's (1996) HAL, $F(3, 1029) = 4.85$, $p < .01$, and log frequency HAL, $F(3, 1029) = 11.96$, $p < .01$; and orthographic neighborhood, $F(3, 1029) = 2.36$, $p < .05$.

A series of planned comparisons were computed in order to explicate these overall differences. For the lexical characteristic of length, a Tukey's LSD test showed that both the negative and positive words used were significantly longer in length than the neutral words. With regard to frequency, all three of the frequency indicators showed that the negative words used in this literature were significantly rarer than the neutral words. In addition, all three frequency indicators showed that the disorder-specific words used were significantly rarer than the neutral words. Finally, in terms of orthographic neighborhood, the LSD comparisons showed that the negative and positive words used have significantly smaller neighborhoods than the neutral words. All of these lexical differences are in the direction hypothesized to spuriously slow down reaction time to negative emotional words compared with control words.

Word Category Differences in Lexical Decision and Naming Speed After Controlling for the Effects of Lexical Characteristics

The lexical characteristics that most differentiate negative from neutral words are all associated with decreased word recognition speed on the part of negative words. It is still possible, however, that some of the slowing in lexical decision and speeded naming to negative words or disorder-specific words may be due to the threat value of the words, even after we controlled for the lexical differences among the word categories. To examine this possibility, we performed a series of ANCOVAs, covarying out the influence of word length, frequency (using the log frequency HAL method of measurement), and orthographic neighborhood on reaction times and accuracy in the two behavioral tasks: lexical decision latency and speeded naming time.

Lexical Decision Latency

For lexical decision latency, we analyzed the impact of word category (positive, negative, neutral, and disorder specific) on mean reaction times, after covarying out the impact of the lexical characteristics associated with each word (word length, frequency, orthographic neighborhood). This analysis revealed that the effect of word category on reaction time remained significant after covarying out the effects of lexical characteristics, $F(6, 1026) = 3.37$, $p < .01$. The overall model, including the lexical characteristics and the word category, accounted for a large proportion of variance in lexical reaction time (full-model $\eta^2 = .553$). However, only a small proportion of variance in lexical reaction time was attributable to word category after removing the effects of lexical characteristics (partial $\eta^2 = .01$). Covarying out the lexical characteristics resulted in an estimated mean for the negative words that was equivalent to the neutral words. However, the estimated mean for the disorder-specific words remained significantly higher (i.e., longer reaction time) than the neutral or any other word category. These types of disorder-specific words appeared to take longer to process in a lexical decision task even after controlling for lexical differences between them and neutral words.

For the lexical decision task, we computed the same covariance analyses for mean accuracy. The results indicated that word category influenced accuracy in completing the task, even after controlling for lexical differences among the word categories, $F(3, 1026) = 3.78$, $p < .01$. The overall model, including the lexical characteristics and the word category, accounted for a modest proportion of variance in lexical decision accuracy (full-model $\eta^2 = .234$). However, only a small proportion of variance in lexical decision accuracy was attributable to word category after removing the effects of lexical characteristics (partial $\eta^2 = .011$). Covarying out the lexical characteristics resulted in an estimated mean for the negative words that indicated a higher degree of accuracy than the estimated mean for the neutral words. However, the estimated mean accuracy for the disorder-specific words remained significantly lower (i.e., less accurate) than the neutral word category. Once again, these types of disorder-specific threatening words appeared to produce more inaccuracies in a lexical decision task even after controlling for lexical differences between the disorder-specific words and the neutral words.

Speeded Naming Task

We replicated the same covariance analyses for naming speed, analyzing the effects of word category (positive, negative, neutral, and disorder specific) on naming speed and accuracy after controlling for the influence of the lexical characteristics of the words (length, frequency, orthographic neighborhood).

For naming speed, we found that the effect of word category was no longer significant after covarying out the effects of lexical characteristics, $F(3, 1026) = 1.08$. The overall model, including the lexical characteristics and the word category, accounted for a large proportion of variance in naming speed (full-model $\eta^2 = .40$). However, only an insignificant proportion of variance in naming speed was attributable to word category after removing the effects of lexical characteristics (partial $\eta^2 = .003$). Therefore, covarying out the lexical characteristics almost entirely eliminated any naming speed differences among the word categories.

Regarding the accuracy of naming speed, very similar results were found. That is, whereas the full ANCOVA accounted for a significant—$F(6, 1026) = 25.91$, $p < .01$—proportion of variability in accuracy (full-model $\eta^2 = .13$), the inclusion of word category above and beyond lexical characteristics accounted for an insignificant amount of variance. In other words, the lexical characteristics that differed across the word categories completely accounted for all variability in word-naming speed.

In summary, there was some evidence that there is an influence of word category above and beyond the lexical characteristics, but this appears to be quite small in terms of effect size and limited to lexical decision performance. This pattern was consistent with the observation that the lexical decision task is more sensitive to semantic variables than the naming task is (see Balota et al., 2004).

Summary of the Data Aggregated by Study

It can be argued that a different pattern of results might be found if the data were first aggregated by study and then summarized. This is due to the fact that different studies use word lists of different length. For example, some studies have word lists as short as 3 words each, whereas others have word lists as long as 50

words each. Such differences could potentially skew results based on an analysis at the level of words (or vice versa). For example, if only a few studies used vastly unbalanced word lists but the word lists used in these studies were quite large, then an analysis by words would overrepresent the degree of lexical unbalance in this literature. As such, we aggregated the data within each of the 32 studies.[4] These aggregated data are presented in Table 2.

We know from the analyses reported so far that word frequency is the most potent variable influencing reaction time and the measure that shows the largest difference between emotional and neutral word categories. In Table 2, if we pick one frequency measure, say the HAL index, we find that the majority of studies show an imbalance in favor of the neutral words being more common than the negative words. If we simply count how many studies exhibit an imbalance in that direction, we find that 21 of the 32 studies (or 66%) show a pattern of the negative words' being rarer than the positive words. Although most researchers used word lists that are unbalanced with regard to frequency, some of the researchers in this area *do* successfully balance for word frequency between their emotional and neutral words (e.g., Arntz et al., 2000; Becker et al., 2001; Williams & Nulty, 1986).

The studies presented in Table 2 also differ substantially in whether they control for word length, though only about half lean in the direction of favoring a larger interference effect (i.e., negative words being longer). To reiterate the results described above, orthographic neighborhood was only weakly related to longer lexical and naming reaction time. At the level of the study, slightly more than half the studies (17 of 32) had an imbalance in orthographic neighborhood size that would bias interference scores (negative words having smaller orthographic neighborhood sizes, and thus slower processing than neutral control words).

The last four rows of Table 2 present the means calculated across the study averages. To get an idea about differences in results due to conducting the analyses at the level of the individual words versus at the level of the study, the reader only needs to compare these rows in Table 2 with the data in Table 1. In Table 1, where results were based on averaging over the 1,033 words in this study, there were large differences among the word categories in terms of length, frequency, and orthographic neighborhood, all of which were in the direction of bias in favor of increased interference to negative words. When analyzing the data by first averaging within studies, then averaging across those studies (see Table 2), the differences among word categories appear much smaller. However, this discrepancy is due to differences in the number of words used in the individual studies. In the analysis at the level of the study, each mean from each study is given equal weight, regardless of the number of words that went into each study's means. A few studies might have used a large number of unbalanced words, yet these studies are given as much weight as studies that used only a few words that were balanced. This makes the interpretation of the means in Table 2 problematic. The utility of Table 2 lies mostly in the identification of particular studies that were or were not well-balanced on lexical characteristics. On the other hand, the utility of Table 1 is that, in a sense, it portrays the results one might expect if one conducted a large-scale emotional Stroop experiment utilizing all 1,033 words that are found in this literature. The results from this level of analysis (see Table 1) clearly show that the collection of words used in this literature is unbalanced, especially with respect to frequency, and that a small interference effect is likely to be found after correcting for lexical characteristics in lexical decision speed.

Discussion

With this article we hope to convince readers of the importance of several points regarding the emotional Stroop phenomenon. The first point is the importance of lexical equivalence among word categories in studies using the emotional Stroop task. If researchers want to infer that any slowdown in response latencies to negative words (compared with neutral words) is due to the emotional content of the words, then it is absolutely crucial that the emotional and neutral words be matched on lexical features known to influence word recognition, especially frequency of word use. The second important point is that researchers must keep in mind that all Stroop tasks operate as they do precisely because people recognize the words and access the semantic meanings of isolated words (see Burt, 2002). While most researchers are aware of this fact, they may not appreciate the consequences, which are that any word feature that differentially influences word recognition between the comparison word lists (e.g., between emotional and control words) will produce spurious interference estimates. The third point is that the emotional Stroop task is fundamentally different from the color Stroop task, and the measurement model used in the color Stroop may not be directly adaptable to the emotional Stroop. To put it simply, in the emotional Stroop, there are never any congruent trials. That is, the words used on the emotion trials are always different from the words used on the control trials. This makes the emotional Stroop a quasi-experimental design, whereas the original color Stroop is a true experimental design. In the color Stroop task, the same words appear on both the congruent and incongruent trials, which leaves users of this paradigm with no concerns regarding lexical equivalence between congruent and incongruent trials. However, in the emotional Stroop, because the words on the emotional and control trials are always different, the issue of lexical equivalence among word categories becomes crucial.

In this study we obtained 1,033 unique words used in published emotional Stroop reports. We first analyzed them for lexical decision latency and naming speed and replicated the general slowing to the negative compared with the neutral words (Algom et al., 2004). This finding also held for lexical decision accuracy, with more mistakes being made on negative than neutral word trials.

When we turned to an analysis of the lexical equivalence among word categories, we found a striking pattern of nonequivalence. That is, negative words were significantly longer in length, more rare in terms of frequency of use, and had smaller orthographic

[4] We did not conduct a formal meta-analysis, which would involve calculating an effect size for each experiment and predicting that from lexical features. We decided that there is simply too much variability across studies: in the samples studied (some are clinical samples, some are preselected on personality variables, some are single gender, etc.), in the word lists themselves (i.e., 1,000 unique words), and in the procedures (i.e., computer presentation vs. card presentation). Although this variability precludes a full meta-analysis of the emotional interference effect sizes, it does not preclude an analysis of the words themselves, especially since our lexical and naming reaction time data were obtained from separate experiments on the individual words.

Table 2
Mean Word Characteristics Averaged Within Publication by Word Type

Study	Word type	Length[a]	K-F[b]	HAL[c]	Log(HAL)[d]	Ortho-N[e]	Lex-RT[f]	Name-RT[g]
Richards et al., 1992	Negative	7.05	28.00	10,919	8.48	1.45	649.95	667.55
	Neutral	7.28	119.18	62,789	9.67	1.68	655.95	668.63
	Positive	7.32	93.53	55,666	9.42	2.00	626.58	662.63
Pratto & John, 1991	Negative	6.85	12.94	7,749	7.43	2.23	701.64	703.08
	Positive	7.20	32.65	14,815	8.45	1.50	682.52	686.09
Dalgleish, 1995	Negative	6.10	43.30	11,078	8.17	2.68	649.06	685.42
	Neutral	6.08	28.38	10,675	7.68	1.88	711.55	705.53
	Positive	6.11	31.17	14,439	8.60	2.56	640.56	653.17
Mathews & Sebastian, 1993	Negative	6.21	50.95	25,773	9.41	3.42	665.00	649.16
	Neutral	6.52	38.84	8,329	7.87	3.29	651.71	643.29
	Specific	6.16	7.69	2,389	6.98	2.68	759.16	735.68
White, 1996	Negative	5.83	42.22	14,234	8.89	1.48	649.30	657.35
	Neutral	5.96	69.36	23,052	9.36	2.24	616.64	651.52
	Positive	6.00	44.29	12,743	8.89	1.43	645.87	651.00
van den Hout et al., 1995	Negative	6.00	92.86	28,417	9.28	4.13	640.25	669.75
	Neutral	5.43	67.57	33,637	9.77	2.14	639.57	647.43
McKenna & Sharma, 1995	Negative	5.03	50.31	20,156	9.39	4.77	610.28	640.13
	Neutral	5.08	50.36	32,975	9.44	5.31	644.23	646.62
Compton et al., 2000	Negative	4.31	74.31	37,475	10.18	9.62	609.23	630.00
	Neutral	4.36	90.55	65,438	10.35	6.27	606.09	634.45
	Positive	4.33	107.00	57,963	10.19	6.67	587.75	626.00
Gilboa-Schechtman et al., 2000	Negative	6.83	20.17	9,054	8.73	1.50	634.67	645.33
	Neutral	6.83	10.80	9,491	8.67	0.50	704.33	677.33
	Positive	7.17	19.17	5,274	7.64	1.33	649.67	652.83
Parker et al., 1993	Neutral	6.20	37.60	17,695	9.46	1.60	620.80	639.60
	Specific	5.00	17.25	12,213	9.26	4.20	658.40	688.40
Dijksterhuis & Aarts, 2003	Negative	4.60	94.38	34,091	9.49	6.80	629.47	642.73
	Positive	4.93	118.53	58,517	10.38	5.07	577.20	629.47
Kitayama & Ishii, 2002	Negative	5.05	34.67	18,944	9.54	3.76	611.38	635.95
	Neutral	5.21	30.86	16,736	9.46	3.64	626.14	640.29
	Positive	5.00	39.69	20,371	9.73	2.92	607.92	627.38
Mathews & Klug, 1993	Negative	6.75	40.13	1,150	8.90	3.25	637.50	653.88
	Positive	7.38	35.88	8,850	8.72	1.38	649.88	675.69
Myers & McKenna, 1996	Negative	6.60	10.50	8,498	8.25	1.10	696.40	737.40
	Neutral	6.60	17.40	37,460	8.57	1.90	667.70	686.50
Williams & Nulty, 1986	Negative	6.20	29.80	9,817	8.50	3.40	635.60	650.00
	Neutral	6.00	31.20	9,837	8.77	4.60	669.40	654.00
Segerstrom, 2001	Negative	7.20	20.80	6,682	8.08	1.20	655.30	659.70
	Neutral	7.16	43.93	13,217	7.88	1.47	663.42	693.05
	Positive	7.20	35.80	11,282	8.57	0.30	660.10	664.90
Schwartz et al., 1996	Negative	4.89	45.22	20,911	9.31	5.17	603.44	644.67
	Neutral	4.61	38.67	20,389	9.01	7.06	628.06	633.11
	Positive	4.89	44.72	22,484	9.22	4.56	628.89	635.56
de Houwer & Hermans, 1994	Negative	5.67	23.00	4,398	8.11	1.67	639.67	674.33
	Positive	4.67	17.00	10,550	9.11	4.67	574.67	582.67
Wentura et al., 2000	Negative	8.15	19.45	7,969	7.35	1.65	727.30	730.85
	Neutral	8.05	32.00	9,697	7.91	0.57	715.48	717.76
	Positive	7.77	57.45	36,898	8.88	1.77	685.85	690.23

Table 2 (*continued*)

Study	Word type	Length[a]	K-F[b]	HAL[c]	Log(HAL)[d]	Ortho-N[e]	Lex-RT[f]	Name-RT[g]
Price et al., 1998	Negative	7.10	16.17	4,647	7.58	1.40	652.65	688.90
	Neutral	6.76	114.78	29,679	8.79	2.10	663.48	672.14
	Positive	7.22	21.39	8,198	7.88	1.67	695.28	681.89
Mogg & Marden, 1990	Negative	7.00	9.50	4,459	7.81	1.00	695.75	736.25
	Neutral	8.00	3.00	134	4.87	0.00	855.50	836.00
	Positive	6.50	8.00	3,558	8.10	2.50	681.00	727.00
McKenna, 1986	Negative	4.60	94.20	30,804	9.93	6.80	615.20	608.80
	Neutral	4.60	93.60	51,859	10.34	7.00	621.00	609.00
Cassiday et al., 1992	Neutral	6.20	37.60	17,695	9.46	1.60	620.80	639.60
	Positive	6.40	82.60	50,710	9.80	3.80	583.20	606.40
	Specific	6.20	21.50	12,651	8.80	4.50	675.60	685.30
Arntz et al., 2000	Negative	6.00	168.40	45,141	9.92	4.00	606.80	620.20
	Neutral	7.86	70.00	45,070	9.98	1.14	653.86	689.14
	Specific	6.78	34.73	22,564	8.32	2.43	678.30	695.04
Paunovic et al. 2002	Negative	7.42	17.00	9,536	8.42	0.11	682.32	689.47
	Neutral	6.92	37.96	21,635	8.63	2.24	678.12	703.20
	Positive	7.09	29.18	17,438	8.86	1.05	674.27	684.73
Seddon & Waller, 2000	Negative	6.00	27.86	7,541	7.83	2.71	665.14	665.00
	Neutral	5.63	36.00	19,518	8.11	2.25	696.50	684.75
	Positive	5.63	21.13	5,328	7.53	2.00	674.75	665.13
Bentall & Kaney, 1989	Negative	6.20	40.00	16,894	8.60	2.80	678.60	722.60
	Neutral	6.20	101.00	64,485	9.76	3.00	630.40	636.60
	Specific	6.60	23.40	8,948	8.74	1.00	659.80	653.20
Green et al., 1995	Negative	4.67	59.33	22,950	9.62	7.33	605.11	629.33
	Neutral	4.78	33.75	11,953	8.59	6.00	603.44	638.00
Lundh & Simonsson-Sarnecki, 2002	Negative	6.95	21.14	7,414	8.09	3.45	692.36	677.45
	Neutral	5.48	171.38	111,494	8.98	5.48	639.88	673.60
	Specific	5.52	26.61	10,495	7.82	3.43	651.10	679.95
Becker et al., 2001	Negative	7.08	46.83	13,219	8.73	0.50	648.58	667.42
	Neutral	5.83	42.83	13,936	8.93	2.67	634.08	636.67
	Positive	6.50	28.83	12,149	9.17	1.92	620.75	648.50
	Specific	7.00	41.27	13,208	8.38	2.33	668.58	663.67
Buhlman et al., 2002	Negative	6.80	22.40	9,959	8.61	0.00	679.40	680.00
	Neutral	7.00	13.60	4,813	7.85	0.80	705.40	661.00
	Positive	7.40	32.80	19,653	8.61	0.90	658.30	706.60
	Specific	7.50	10.25	3,498	7.07	0.25	741.50	693.50
Miller & Patrick, 2000	Negative	6.53	34.71	14,139	8.89	3.00	640.27	640.27
	Neutral	6.80	17.67	6,966	7.83	1.27	693.67	678.53
	Positive	6.80	24.93	11,707	8.64	2.40	630.87	654.47
Averaged across studies								
	Negative	6.24	42.32	15,483	8.69	2.99	651.50	668.03
	Neutral	6.23	54.48	28,121	8.80	2.01	661.09	668.07
	Positive	6.37	43.90	21,789	8.87	1.56	640.20	658.01
	Specific	6.34	22.83	10,746	8.17	1.46	686.55	686.84

Note. Full references for listed studies are provided in the Appendix.
[a] Length is average word length in letters. [b] K-F is average frequency of use according to the Kučera and Francis (1967) norms. [c] HAL is the average frequency of use according to the Hyperspace Analogue to Language (Lund & Burgess, 1996) norms. [d] Log(HAL) is the average of the log transform of the HAL frequency index. [e] Ortho-N is the average orthographic neighborhood index. [f] Lex-RT is average lexical decision speed in milliseconds. [g] Name-RT is the average naming speed in milliseconds.

neighborhoods than their neutral counterpart words. All of these lexical features contribute to slower word recognition speed (Balota et al., 2004). It may be the case that the generic slowdown observed in color naming of negative words, so often replicated in the literature, is due not so much to the negativity or threat value of the word per se but to the fact that the words used are typically longer, rarer, and have smaller orthographic neighborhoods than the control words typically used.

It might be argued that, because many emotional Stroop studies focus on individual differences, the problem we are reporting may not really be a problem for such studies. That is, if individual differences in emotional interference scores correlate with an external variable (typically a personality variable like neuroticism or trait anxiety), then the interference score is de facto valid. We disagree. Instead, we argue that individual differences in a contaminated variable make such correlational findings ambiguous. That is, if the interference score contains both true emotional variance and error variance due to lexical contamination, then any correlation with it is uninterpretable (i.e., is the correlation due to the emotional component or due to the low-frequency component). Indeed, it may be that persons with trait anxiety or high neuroticism have attentional capture to rare words, and that the individual difference correlation with the interference score may be due more to the frequency component than to the emotional component of the interference score.

A recent debate in the literature on the emotional Stroop task concerns the explanation for the typical finding that people in general are slower to color-name negative words compared with neutral words. Algom et al. (2004) presented the argument that negative words temporarily disrupt cognitive processing by eliciting an automatic vigilance mechanism, thereby producing a generic slowing in concurrent cognitive processes that are not specific to color naming. If this is true, they argued, then the effects of word negativity should be observable on any cognitive process involving word recognition, such as lexical decision latency and word-naming speed. They demonstrated exactly these effects in four separate experiments using negative and control words that were matched on subjective familiarity ratings.[5] The negative and control words they used in their four experiments were not equivalent in terms of the HAL frequency index. Their negative words were *betrayal, crisis, danger, failure,* and *fear*. These five negative words have a mean HAL frequency index of 14,380. The neutral control words used by Algom et al. (2004) were *avenue, field, path, neighborhood,* and *passage*. These five control words have a mean HAL frequency index of 26,584, suggesting that the average neutral control word used in this study is almost twice as frequent in everyday language as the negative words. Although Algom et al. (2004) went on to demonstrate a generic slowing in color naming, lexical decision, and word naming between these negative and neutral words, it is possible that at least some of this effect was due to the objective frequency differences between these specific sets of five negative and five neutral words.

Our point is not to argue that the use of unbalanced word lists completely invalidates the generic slowdown effect routinely observed in color-naming and lexical decision tasks. Instead, our point is that the use of unbalanced word lists can inflate an estimate of any valid effect due to the emotional meaning of the words. The best way to think of this is that there are two components to an interference effect: one due to true emotional effects on attentional processes, the other due to lexical differences in the word lists. Studies that are designed in a way that aggregates these two components, which is the majority of studies in our list of 32, will necessarily overestimate the size of the true effect. True emotion effects may be there. Indeed, Algom et al. (2004, Experiment 5) found lagged effects of emotional words. That is, correct lexical decisions for nonwords were slower if they appeared in a block of negative words compared with those that appeared in a block of neutral words. The size of this carryover slowdown (26 ms) was, however, much smaller than the size of traditionally computed interference effect (a slowdown of 48 ms for the emotional words compared with neutral words in lexical decision time).

Another aspect of the Algom et al. (2004) study was that it was done in the Hebrew language. The authors provided English translations for the Hebrew words used in their study, and the frequencies reported above were those obtained in English-speaking samples. In fact, out of the 32 studies used in the present analyses, 7 were conducted in languages other than English (e.g., Japanese, Swedish, Dutch, German). Hence, one may be concerned about the frequency estimates derived from the English equivalents. However, Bates' et al. (2003) have recently reported quite remarkable stability in conceptual word use frequency across different languages. For example, in correlating the frequency measures of a large sample of words across seven languages, they found that all possible cross-correlations were positive, significant, and robust, ranging from a low of .37 between Bulgarian and Chinese to a high of .66 between German and Hungarian. Consequently, whereas the length and orthographic differences between English and non-English words would be obviously different in different languages, it is likely that the frequency-of-use estimates would be similar between languages and that our frequency findings should hold for the 7 studies in our database that were conducted in non-English languages. We found that these 7 studies accounted for 295 words in our database. However, because of redundancy in our original list of 1,401 words, when we deleted the non-English words, our list of unique words shrank from 1,033 to 965 words. We reran all analyses on this English-only set of words, and the results for the lexical characteristics remained essentially unchanged.

Could it be that word negativity *does* capture some degree of cognitive processes, above and beyond the variability explained by lexical features of the words? To estimate the true magnitude of any generic slowdown to negative words, we examined the effects of word category on lexical decision and naming speed in our set of 1,033 words after controlling for the important lexical differences among the word categories. When we did this, we found that the speed differences between the negative words and the neutral control words disappeared. That is, the negative words no longer were associated with a slowdown in processing above and beyond that accounted for by lexical differences. However, we did have a category of words we coded as disorder specific that remained significantly but modestly slower than the control words after controlling for lexical characteristics.

The 78 disorder-specific words we analyzed were primarily drawn from studies on clinical populations, such as depressed persons, rape victims, persons with snake phobia, and persons with public speaking anxiety. As such, many of the words in this category are specifically

[5] Algom et al. (2004) did not use an objective indicator of frequency of use in selecting their words. Instead, they had 14 judges rate the words on subjective familiarity on a scale of 0 (*not familiar*) to 5 (*very, very familiar*). They did not explicitly match the words on these ratings but instead selected words from the average range (e.g., 3 to 4 on this scale). Moreover, although subjective ratings of familiarity are highly correlated with objective measures of frequency of use, these measures can account for independent variance in word recognition performance (Balota et al., 1994).

relevant to these disorders, such as, respectively, *hopelessly, intercourse, constrictor,* and *stutter*. However, many of the other words from this category may be generically threatening to most people, and so might make good candidates for future Stroop studies. Words such as *ache, bite, bleeds, bruises, cramp, defeat, disfigure, dishonesty, germ, hideous, illness, incest, infected, lying, rejection, repulsive, ridicule, tumor,* and *vomit.* These words were from the category that remained associated with generic slowing even after controlling for lexical features. Nevertheless, future research using such words should still pay careful attention to the lexical features of the control words. Because these negative words are less frequently used than average, the researcher should also compile a list of control words that are equivalent on an objective measure of frequency. In addition, word lists should be equivalent on length and orthographic neighborhood as well, because these characteristics can also be important to word recognition.

The issue of lexical balance in word lists can be addressed in several ways. One strategy would be to use unbalanced word lists but then control for lexical features in the analysis by using them as covariates. This strategy is well demonstrated by Wentura et al. (2000), who used a regression approach to control for lexical features between emotional and control word categories. However, none of the other 31 studies in our database used such an approach to control for word frequency. Clearly, an approach that controls for lexical features before determining the degree of emotional interference would be an effective solution to the problem we point to. Wentura et al. successfully used a repeated-measures multiple regression approach as a way of correcting for frequency differences between word lists before examining emotion differences.

Another way to approach the issue of lexical imbalance is simply to ensure that words in the emotion list are completely balanced to those in the control list for important lexical features. Along these lines, the Web-based lexical database developed by Balota et al. (2002), which we used in this project, is reverse searchable. That is, researchers can enter the lexical parameters they desire and the database will generate a list of words that fit those parameters. This allows researchers to easily generate word lists that are matched on important lexical features. We have used this feature to construct lists of emotional words (30 negative, 30 positive) and control words (60 total) that are completely balanced on length, frequency of use, and orthographic neighborhood size. We will make these lists available to anyone who sends us a request. Using lists so constructed in emotional Stroop experiments would overcome the important threat to internal validity that we have identified in this article.

References

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.

Andrews, S. (1989). Frequency and neighborhood size effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 802–814.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 234–254.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review, 4,* 439–461.

Arntz, A., Appels, C., & Sieswerda, S. (2000). Hypervigilance in borderline disorder: A test with the emotional Stroop paradigm. *Journal of Personality Disorders, 14* , 366–373.

Balota, D. A., Cortese, M. J., Hutchison, K. A., Neely, J. H., Nelson, D., Simpson, G. B., Treiman, R. (2002). The English Lexicon Project: A Web-based repository of descriptive and behavioral measures for 40,481 English words and non-words. Available at http://elexicon .wustl.edu

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

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A. et al. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin and Review, 10,* 344–380.

Becker, E. S., Rinck, M., Margraf, J., & Roth, W. T. (2001). The emotional Stroop effect in anxiety disorders: General emotionality or disorder specificity? *Anxiety Disorders, 15,*147–159.

Besner, D., & Stolz, J. A. (1999). What kind of attention modulates the Stroop effect? *Psychonomic Bulletin and Review, 6,* 99–104.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting RT in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers, 30,* 272–277.

Burt, J. S. (1999). Associative priming in color naming: Interference and facilitation. *Memory & Cognition, 27,* 454–464.

Burt, J. S. (2002). Why do non-color words interfere with color naming? *Journal of Experimental Psychology: Human Perception and Performance, 28,* 1019–1038.

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind, 11,* 63–65.

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

Forster, K. I. & Shen, D. (1996). No enemies in the neighborhood: Absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 696–713.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods, Instruments, & Computers, 28,* 203–208.

MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163–203.

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

Stroop, J. R. (1935). Studies of interference in serial–verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach- and avoidance-related 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.

Williams, J. M. G., & Nulty, D. D. (1986). Construct accessibility, depression and the emotional Stroop task: Transient mood or stable structure? *Personality and Individual Differences, 7,* 485–491.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47,* 1–29.

Appendix

Publications Providing Emotion Stroop Words Used in This Research

Arntz, A., Appels, C., & Sieswerda, S. (2000). Hypervigilance in borderline disorder: A test with the emotional Stroop paradigm. *Journal of Personality Disorders, 14*, 366–373.

Becker, E. S., Rinck, M., Margraf, J., & Roth, W. T. (2001). The emotional Stroop effect in anxiety disorders: General emotionality or disorder specificity? *Anxiety Disorders, 15,* 147–159.

Bentall, R. P. & Kaney, S. (1989). Content specific information processing and persecutory delusions: An investigation using the emotional Stroop test. *British Journal of Medical Psychology, 62,* 355–364.

Buhlman, U., McNally, R. J., Wilhelm, S., & Florin, I. (2002). Selective processing of emotional information in body dysmorphic disorder. *Anxiety Disorders, 16,* 289–298.

Cassiday, K. L., McNally, R. J., & Zeitlin, S. B. (1992). Cognitive processing of trauma cues in rape victims with post-traumatic stress disorder. C*ognitive Therapy and Research, 16,* 283–295.

Compton, R. J., Heller, W., Banich, M. T., Palmieri, P. A., & Miller, G. A. (2000). Responding to threat: Hemispheric asymmetries and interhemispheric division of input. *Neutropsychology*, *14,* 254–264.

Dalgleish, T. (1995). Performance on the emotional Stroop task in groups of anxious, expert, and control subjects: A comparison of computer and card presentation formats. *Cognition & Emotion, 9,* 341–362.

de Houwer, J., & Hermans, D. (1994). Differences in the affective processing of words and pictures. *Cognition & Emotion, 8,* 1–20.

Dijksterhuis, A., & Aarts, H. (2003). On wildebeests and humans: The preferential detection of negative stimuli. *Psychological Science, 14,* 14–18.

Gilboa-Schechtman, E., Revelle, W., & Gotlib, I. H. (2000). Stroop interference following mood induction: Emotionality, mood congruence, and concern relevance. *Cognitive Therapy and Research, 24,* 491–502.

Green, M., Rogers, P. J., & Elliman, N. A. (1995). Change in affective state assessed by impaired color-naming of threat-related words. *Current Psychology: Developmental–Learning–Personality–Social*, *14,* 222–232.

Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition & Emotion, 16,* 29–59.

Lundh, L. G., & Simonsson-Sarnecki, M. (2002). Alexithymia and cognitive bias for emotional information. *Personality and Individual Differences, 32,* 1063–1075.

Mathews, A., & Klug, F. (1993). Emotionality and interference with color-naming in anxiety. *Behavior, Research, and Therapy, 31,* 57–62.

Mathews, A. M., & Sebastian, S. (1993). Suppression of emotional Stroop effects by fear-arousal. *Cognition & Emotion, 7,* 517–530.

McKenna, F. P. (1986). Effects of unattended emotional stimuli on color-naming performance. *Current Psychological Research & Reviews, 5,* 3–9.

McKenna, F. P., & Sharma, D. (1995). Intrusive cognitions: An investigation of the emotional Stroop task. *Journal of Experimental Pschology: Learning, Memory, and Cognition, 21,* 1595–1607.

Miller, M. W., & Patrick, C. J. (2000). Trait differences in affective and attentional responding to threat revealed by emotional Stroop interference and startle reflex modulation. *Behavior Therapy, 31,* 757–776.

Mogg, K., & Marden, B. (1990). Processing of emotional information in anxious subjects. *British Journal of Clinical Psychology, 29,* 227–229.

Myers, L. B., & McKenna, F. P. (1996). The color naming of socially threatening words. *Personality and Individual Differences, 20,* 801–803.

Parker, J., Taylor, G., & Bagby, M. (1993). Alexithymia and the processing of emotional stimuli: An experimental study. *New Trends in Experimental and Clinical Psychiatry, 9,* 9–14.

Paunovic, N., Lundh, L. G., & Ost, L. G. (2002). Attentional and memory bias for emotional information in crime victims with acute posttraumatic stress disorder. *Anxiety Disorders, 16,* 675–692.

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

Price, K. H., Hunton, J. E., Hall, T. W., Coalter, T. M., & Clinton, B. D. (1998). Reactions to majority voting procedures and amelioration of voting member responses. *Personality & Social Psychology Bulletin, 24,* 214–223.

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.

Schwartz, C. E., Snidman, N., & Kagan, J. (1996). Early temperamental predictors of Stroop interference to threatening information at adolescence. *Journal of Anxiety Disorders, 10,* 80–96.

Seddon, K., & Waller, G. (2000). Emotional processing and bulimic psychopathology: Age as a factor among nonclinical women. *International Journal of Eating Disorders, 28,* 364–369.

Segerstrom, S. C. (2001). Optimism and attentional bias for negative and positive stimuli. *Personality and Social Psychology Bulletin, 27,* 1334–1343.

van den Hout, M., Tenney, N., Huygens, K., Merckelbach, H., & Kindt, M. (1995). Responding to subliminal threat cues is related to trait anxiety and emotional vulnerability: A successful replication of Macleod and Hagan. *Behavior Research and Therapy, 33,* 421–454.

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.

White, M. (1996). Automatic affective appraisal of words. *Cognition & Emotion, 10,* 199–211.

Williams, J. M. G., & Nulty, D. D. (1986). Construct accessibility, depression and the emotional Stroop task: Transient mood or stable structure? *Personality and Individual Differences, 7,* 485–491.

Received December 15, 2004
Revision received June 10, 2005
Accepted June 30, 2005 ■