Running head: Control of semantic retrieval

The semantic Stroop effect is controlled by endogenous attention

Sachiko Kinoshita

Luke Mills

Macquarie University

and

Dennis Norris

MRC Cognition and Brain Sciences Unit

Word count: 10,119 words (text only), 4 tables

As at December 8 2017

Contact Information:
Sachiko Kinoshita
Department of Psychology and
ARC Centre of Excellence in Cognition and its Disorders (CCD)
Macquarie University
Sydney, NSW 2109
Australia
Email: sachiko.kinoshita@mq.edu.au
Fax: 61 2 9850 6059

Abstract

Using the oral and manual Stroop tasks we tested the claim that retrieval of meaning from a written word is automatic, in the sense that it cannot be controlled. The semantic interference effect (greater interference caused by color-related words than color-neutral words) was used as the index of semantic activation. To manipulate the level of attentional control over the task of reading, the proportion of non-readable, neutral trials (a row of #s) was varied (75% vs. 25%). In all four experiments a high neutral proportion magnified the interference caused by word distractors. With the color-associated words presented in incongruent color (e.g., LEMON in blue), the semantic Stroop effect was weak and did not interact with neutral proportion (Experiment 1 and 2). Experiment 3 and 4 used color names (e.g., GREEN) not in the response set, and here the semantic interference effect was more robust, and the effect was magnified in the high neutral proportion condition. We take these results to argue that semantic retrieval is controlled by endogenous attention in the Stroop task. (173 words)

Keywords: Stroop effect; automaticity; semantic activation; task conflict; informational conflict;

THE SEMANTIC STROOP EFFECT IS CONTROLLED BY ENDOGENOUS ATTENTION

> *Reading a sentence and understanding it were the same thing; as with the crooking of a finger, nothing lay between them. There was no gap during which the symbols were unravelled. You saw the word* castle, *and it was there, seen from some distance, with woods in high summer spread before it, the air bluish and soft with smoke rising from the blacksmith's forge, …* (McEwan, 2001, *Atonement*, p.37)

The idea that semantic retrieval is automatic ("direct and simple" in the words of the protagonist, Briony, in *Atonement*) is widely held not only amongst budding writers but also experimental psychologists. The Stroop interference effect (Stroop, 1935; see MacLeod, 1991, for a review) is often cited as evidence: When asked to name the color in which a word is written, the reader finds it difficult when the meaning of the word is incongruent with the response. More specifically, the fact that the interference is observed even though the participant is instructed to ignore the word, and when reading is harmful to the performance, is taken to indicate that reading is automatic in the sense that it is "involuntary and cannot be controlled" (Neely & Kahan, 2001, p.69; Moors & De Houwer, 2006; Tzelgov, 1997, but see Besner, 2001; Besner & Stolz, 1999 for a contrary view, to be reviewed shortly). The present study tests the idea that retrieval of meaning is automatic in this sense; that it cannot be controlled by endogenous (voluntary/intentional) attention. Using both oral and manual Stroop tasks, our experiments examined the interference produced by color-related words which are not the names of response colors (the "semantic Stroop effect"), presented intact, and tested whether the effect is modulated by the proportion of non-readable, neutral trials. Below, we review the research that provided the rationale for this design.

The semantic Stroop effect. The first major challenge to the idea that semantic retrieval is automatic was mounted by Besner and colleagues (e.g., Besner & Stolz, & Boutiler, 1997; Besner & Stolz, 1999; Stolz & Besner, 1999). They initiated a line of research showing that the Stroop effect is reduced substantially when only a single letter in the word distractor is colored (e.g. when only E in RED is presented in color). Using a manual Stroop task with four response colors, Besner et al. (1997) showed that the Stroop congruence effect - the difference between the incongruent (e.g., the word RED presented in blue) and congruent (the word RED presented in red) trials – was reduced from 103 ms in the standard, all-letters-colored condition to 72 ms in the single letter colored condition, and in the case of the comparison between the incongruent distractors vs. the neutral distractors (pronounceable pseudowords starting with the same two letters as the color names, e.g. ret, grend, blat), the interference effect (34 ms in the all-letters-colored condition) was completely eliminated in the single-letter colored condition. Besner (2001) took these results to argue that the idea that skilled readers cannot prevent themselves from reading the irrelevant word and retrieving meaning from it is a “myth”.

In a review titled “Is semantic activation automatic?” (which concluded in the affirmative), Neely and Kahan (2001) made two points in response to the evidence presented by Besner and colleagues. One was that in the single-letter coloring condition, misdirected spatial attention can impair visual feature integration to other letters in the word such that the word is not ‘seen’ (perceptually encoded). Neely and Kahan stated that “even the staunchest supporter of semantic activation automaticity would not argue that semantic activation should occur under those conditions” (p.88). Indeed, evidence based on paradigms other than the Stroop effect (e.g., masked priming – Lachter, Forster & Rutheruff, 2004; Lien, Ruthruff, Kouchi, & Lachter, 2010) converges on the conclusion that spatial attention to the word is a prerequisite for lexical access. Hence, in the remainder of this paper, we will consider only those studies that presented words intact, as in normal reading. Neely and Kahan's second point, which is a starting point of our study, was that the standard Stroop interference effect observed with word distractors that are color names in the response set (e.g., RED presented in green when red is a response color) may not reflect the word's semantic representations, but instead "nonsemantic, task-relevant response competition" (p.71).

As an alternative to color names in the response set, Neely and Kahan (2001) recommended using color-associated words, such as LEMON and SKY. It has been known since Klein (1964) that words that have strong association to a color produce greater interference when presented in an incongruent color (e.g., LEMON in blue, SKY in red) than noncolor words with no association to a particular color such as TABLE and PUT. Neely and Kahan (2001) endorsed the comparison between the color-associated words and color-unassociated words (which they referred to as the SKY-PUT Stroop design) as a measure of semantic activation independent of response competition.

Augustinova and Ferrand (2014) reviewed the research using the SKY-PUT Stroop design, and agreed with Neely and Kahan (2001) that semantic retrieval is automatic. Their main argument was that while various manipulations have been found to reduce, or even completely eliminate, the classic Stroop effects observed with the names of response colors, they had little impact on the semantic interference effect indexed by the SKY-PUT design. The studies reviewed used the single-letter coloring manipulation [1] described earlier (Augustinova,

[1] Findings using this manipulation are in fact highly mixed. For example, subsequent to Augustinova and Ferrand's (2014) review, Labuschagne and Besner (2015) reported contrary data (see Footnote 3).

Flaudius, & Ferrand, 2010, Manwell, Roberts, & Besner, 2004) as well as “word blindness” suggestions, such as a post-hypnotic suggestion technique that caused highly suggestible individuals to consider the characters they saw as gibberish (e.g., Augustinova & Ferrand, 2012; Raz, Kirsch, Pollard, & Nitkin-Kaner, 2006) or “social priming of dyslexia” manipulation (making literate individuals think about what the everyday life of a dyslexic person might be like, e.g., Augustinova & Ferrand, 2014; Goldfarb, Aisenberg, & Henik, 2011).

We believe however there is a limitation with the body of evidence presented by Augustinova and Ferrand (2014) for “automatic semantic activation”. All of the manipulations – single-letter coloring, hypnotic or social priming suggestion of word blindness – were designed to reduce an effect that is already small. Compared to the classic Stroop interference effect, the semantic interference effect indexed by the SKY-PUT design is substantially smaller. For example, in Augustinova and Ferrand’s (2012) Experiment 1 (see also Augustinova & Ferrand, 2014, for review of this experiment) using an oral Stroop task, relative to the color-unrelated word distractors (e.g., DEAL), the classic Stroop interference effect with response-incongruent distractors was 70 ms, and the semantic interference effect produced by color-associated words like SKY and LEMON was only18 ms; the word blindness suggestion reduced the standard Stroop effect (to 31 ms), but not the semantic Stroop effect (17 ms). When the evidence rests on the insensitivity of a small effect to manipulations that are designed to reduce the effect, naturally a concern arises as to whether it was a floor effect. Augustinova and Ferrand (2014) themselves wrote that “it is … important to demonstrate that any such elimination is specifically due to the elimination of the semantic conflict that was significant before any intervention designed to reduce it was applied” (p. 346). We would add that it is not sufficient to show simply that the effect was significant; the question is whether there was enough power to detect an interaction. Studies reviewed by Augustinova and Ferrand (2014) reporting null modulation of semantic interference effects have not presented such evidence.[2] Moreover, as we explain shortly, there is an emerging line of evidence that suggests that the semantic interference effects observed in previous studies were *a priori* limited in size.

Task conflict and the neutral proportion manipulation. In the literature reviewed above, the implicit assumption is that the interference in a Stroop task reflects the competition between the information activated automatically by the color and the distractor, for example, in the word SKY presented in red, the semantic features of “sky”, which includes the color feature “blue”, compete with that of “red”. This is referred to as the *informational conflict*.[3] There is a growing recognition in the Stroop literature that Stroop interference also reflects another type of conflict, namely, *task conflict* (e.g., Goldfarb & Henik, 2007; Monsell, Taylor & Murphy, 2001; Roelofs, 2012; Steinhauser & Hübner, 2009).

Task conflict refers to the competition between task sets. The notion of task set has its origin in the task switching literature (e.g., Allport, Styles & Hirsh, 1994; Rogers & Monsell, 1995), and refers to the configuration of cognitive processes necessary to perform the task associated with the stimulus. Monsell, et al. (2001) argued that in the Stroop task, the task set of reading competes with the task set of color naming, and that the interference effect observed with color-neutral words (e.g. ABBEY; MERCY) reflect competition at the level of the task set. This claim is based on the finding that the size of interference in the oral Stroop task was not related

[2] In contrast, Labuschagne and Besner (2015), which presented evidence against Augustinova and Ferrand’s (2014) conclusion, did report Bayes factors. This study presented color-associated words with symbols inserted between the letters (e.g., f6r*o#g) and showed that single-letter coloring eliminated the semantic interference effect. Here, the authors reported that the Bayes factor indicated overwhelming support for the semantic interference effect in the all-letters-colored condition, and clear support for the null effect in the single-letter colored condition.

[3] Informational conflict is sometimes referred to as *response conflict* (e.g., Monsell, et al., 2001; Steinhauser & Hubner, 2009). We use the term informational conflict here to avoid the connotation that the conflict arises late, during the preparation of specific motor response.

to the strength of association between the distractor and its pronunciation, but to the presence of word-like constituents. Specifically, Monsell et al. found that word frequency did not affect the size of Stroop interference even though the high-frequency words (e.g., work) were read aloud faster than low-frequency words (e.g., womb); similarly, lexicality did not affect the size of Stroop interference even though words were read aloud faster than pronounceable pseudowords (e.g., woze). The word and pseudoword distractors however interfered more than non-readable consonant strings (e.g., wlwb), which in turn interfered with color naming relative to a row of Xs. Monsell et al. interpreted these findings as suggesting that the interference from noncolor words is due to competition from the task set of reading which is triggered in part by stimulus characteristics, but is also controlled by endogenous (voluntary) attention. Specifically, they argued that "a control bias is endogenously applied …to enable the relevant task set (the mapping from colors to their names) and *suppress the stronger…task set of reading*" (emphasis ours), but that "the detection of wordlike properties of the stimulus evokes exogenously in literate subjects the associated task set of reading" (p.147).

Monsell et al. (2001) are thus explicit in suggesting that the task of reading in the Stroop task is suppressed by voluntary attention. Goldfarb and Henik (2007; see also Tzelgov, Henik, & Berger, 1992) were the first to show experimentally that this task suppression can be relaxed by increasing the proportion of non-readable, neutral trials. Using a manual Stroop task and congruent (e.g., RED in red), incongruent (e.g., RED in blue) and non-readable neutral distractors, Goldfarb and Henik varied the proportion of neutral trials (75% vs. 25%) while holding the relative proportion of congruent and incongruent trials equal (i.e., both 12.5% in the high-neutral proportion condition or both 37.5% in the low-neutral proportion condition). Following Monsell et al. (2001), Goldfarb and Henik assumed that in a word distractor, a conflict arises between the task of color naming and word reading, and that this task conflict exists regardless of the congruence between the meaning of the word and the color of the word ink. In contrast, a non-readable neutral distractor (like a row of Xs) contains no task conflict. On the assumption that control is relaxed when the proportion of conflict trials is low, greater task conflict should arise for readable word distractors and magnify the interference effect due to task conflict. This should increase the interference effect for incongruent distractors; also, a *reverse facilitation effect* (slower responses to congruent trials relative to neutral trials) may emerge. The results were entirely consistent with this prediction: Relative to the low-neutral proportion condition, the interference effect was magnified and a reverse facilitation emerged in the high neutral proportion condition (see also Entel, Tzelgov, Bereby-Meyer, & Shahar, 2015, for an extended replication using the oral Stroop task). This was not due simply to a shift in the neutral condition becoming faster as the proportion of neutral trials is increased; the net difference between the congruent and incongruent trials was also magnified in the high neutral proportion condition.

Recently, Mills (2017) applied the neutral proportion manipulation to the interference produced by non-color associated word distractors (e.g., ADORE, HOBBY). In the oral Stroop task, in the low-neutral proportion condition in which 25% of the trials were non-readable neutral distractors (a row of #s), the interference produced by the word distractors relative to a row of #s was 62 ms; this was magnified to 106 ms in the high-neutral proportion condition in which 75% of the trials contained the non-readable neutral distractor. The same pattern of results was found with the manual Stroop task: Whereas there was no word interference effect in the low neutral proportion condition (-7 ms), a statistically significant word interference effect (30 ms) emerged in the high neutral proportion condition. As with the studies involving the classic congruent and incongruent trials described above, this pattern was not due to the neutral condition becoming faster in the condition containing a high proportion of neutral trials. The absence of word interference effect replicated the result reported in previous studies using the manual Stroop task (Kinoshita, et al., 2017, Experiment 2; Sharma & McKenna, 1998). Note that in these experiments the proportion of non-readable neutral trials was low.

The neutral proportion effects observed by Mills (2017) extended Goldfarb and Henik's (2007) findings in two important ways. First, the results showed that the manipulation works with word distractors as is not limited to the names of response colors, consistent with the view that the neutral proportion manipulation operates on the task of reading. Second, the manipulation works both with the manual and oral Stroop tasks, and thus it has a potential to reveal the effects of stimulus properties of the readable distractors that have not been observed with the former. This is relevant to a point made by Augustinova and Ferrand (2014; also Augustinova et al., 2010): They recommended against the use of manual Stroop task on the ground that interference effects are generally smaller than in the oral task; but they did not provide a theoretical reason for preferring the oral task over the manual task ("there is nothing intrinsically wrong with this variant of the Stroop task", p. 346). That the effects are generally smaller in the manual task should not be of concern in a design that is aimed at magnifying the effect (provided that there was enough power to detect the interaction). Thus, in the present study, we use both the oral and manual tasks, and examine whether the modulation of the semantic Stroop effect is dependent on the task mode.

The present study. Previous Stroop studies investigated the automaticity of semantic retrieval by testing whether the semantic Stroop effect can be reduced or eliminated (see review by Augustinova & Ferrand, 2014). We pointed out that this approach of demonstrating a reduction of an already small effect is methodologically (and statistically) limiting. The alternative approach we take here is motivated theoretically by the recent Stroop literature indicating that endogenous voluntary attention is used to suppress the task of reading (Monsell, et al., 2001), and that this suppression can be relaxed by increasing the proportion of non-readable neutral trials (Entel, et al., 2015; Goldfarb & Henik, 2007; Mills, 2017). Previous studies that examined the modulation of semantic interference effect (those reviewed by Augustinova & Ferrand, 2014) typically contained few (or no) non-readable neutral trials, thus the endogenous control over the task of reading would have been high, and this would have made the effect small. In the present study, we tested whether the size of semantic Stroop effect can be *magnified* by increasing the proportion of neutral trials, which relaxes the attentional control over the task of reading. We tested this in four Stroop experiments, using the oral, and manual Stroop tasks.

The basic design of the four experiments was the same. The semantic interference effect was indexed as the difference between the color-related words that are not the response color names and non-color related control words, and we tested whether its size can be modulated by the proportion of non-readable neutral trials (a row of #s). Experiments 1 and 3 used the oral Stroop task (in which participants named the color), and Experiments 2 and 4 used the manual Stroop task (in which participants were instructed to press a key to indicate the color). In Experiments 1 and 2, we used words which denoted objects associated with a specific color, e.g., LEMON, SKY; in Experiments 3 and 4, we used color names which were not in the response set. In all experiments, the prediction was that if semantic retrieval is controlled by endogenous attention, the size of semantic interference effect should be magnified in the high neutral proportion condition. We will postpone the discussion of Experiment 1 until after Experiment 2 and discuss the oral and manual Stroop experiments together.

Experiment 1 (color-associated words, oral response)

Method

Participants. Forty students from Macquarie University participated in the experiment in return for course credit. Twenty were assigned to the high (75%)-neutral proportion condition and the other twenty to the low (25%)-neutral proportion condition, in the order of arrival.

Design. The experiment used the Stroop color naming task, and involved the factor Distractor type (Color-associated word, Control word or Neutral (# signs)) manipulated within subjects, and their relative proportion (75% neutral trials or 25% neutral trials) manipulated between groups. The dependent variables were color naming latency and error rate.

Materials. There were four response colors, red, yellow, green and blue. The color-associated words were selected on the basis that they were associated with one of the response colors, and did not share the initial phoneme with the names of response colors (i.e., they did not start with /r/, /j/, /g/ or /b/). They were FIRE, LEMON, PEA and SKY. The control, color-unassociated words were HOPE, NIP, MERCY and FLY, and they were matched with the color-associated words on length, word frequency (mean 63 for the color-associated words and 61 for the color-unassociated words, range 1.78-175 per million based on Celex frequency, Baayen, Piepenbrock, & Gulikers, 1995; mean 69 for the color-associated words and 108 for the color-unassociated words, range 3.2-320 per million based on Subtitle frequency, Brysbaert & New, 2009)[4], the number of orthographic neighbors (as defined by the N metric, Coltheart, Davelaar,

[4] It may be noted that both here and in Experiments 3 and 4, the mean word frequency values for the two word types vary depending on which frequency counts (Celex vs. SubtitleUS corpus) are used. This is because the frequency of

Jonasson, & Besner, 1977, mean 8.75 for the color associated words and 9.0 for the color-unassociated words, range 1 - 15), and position-dependent bigram frequency (mean 15.6 for both types of words, range 1.5 to 29.6). The neutral distractors were a string of #s matched to the number of letters, i.e., ###, #### or #####. (There were twice as many ### trials as other # string trials, because there were two critical words that were 3-letters long, i.e., PEA and SKY).

The distractors were presented in one of four colors, red (RGB 255, 000, 000), yellow (RGB 255, 255, 000), green (RGB 000, 128, 000) or blue (RGB 000, 000, 255), against a black background. The color-associated words were presented in (one of three) colors other than the associated color (e.g., LEMON was presented in red, green or blue but not yellow), and the matched color-unassociated word and # signs were also presented in the three colors (e.g., MERCY and ##### were presented in red, green and blue but not yellow). Thus there were twelve each of color-word combinations for the color-associated words and color-unassociated words and ten color-#s combinations (a total of 34 stimuli).

Both the High neutral proportion and Low neutral proportion groups received 288 Stroop color naming trials. In the High neutral proportion condition, 216 trials (75%) contained the neutral distractor, and 36 trials contained color-associated words and 36 trials contained color-unassociated words as distractors. In the Low neutral proportion condition, 72 trials (25%) contained the neutral distractor, 108 trials contained the color-associated words and 108 trials contained the color-unassociated words as distractors. Each list was divided into three sublists of 96 trials with each sublist containing the representative proportion of word and #s trials, the three distractor types, and the four response colors occurring equally often. A pseudo-random order of

the most frequent word is quite different in the two frequency counts, which has a disproportionate effect on the mean frequency with only four items for each word type. In any case, word frequency has little effect on the Stroop color naming/response latency (Monsell, et al., 2001).

trials was generated for each sublist such that the same color did not occur in succession. The order of response colors was identical for the High- and Low proportion neutral conditions.

Apparatus and Procedure. Participants were tested individually, seated approximately 60 cm in front of a flat screen monitor, upon which stimuli were presented. Each participant completed 288 color naming trials, presented in three blocks (with each block containing 96 trials) with a self-paced break between the blocks. A practice block of 32 trials containing each of the four colors occurring equally often, and containing the same proportion of word and neutral trials as the test blocks preceded the test blocks.

Participants were instructed at the outset of the experiment that on each trial they would be presented with a word or a row of #s presented in one of four colors, red, yellow, green or blue, and their task was to name the color of the stimulus, as quickly and accurately as possible.

Stimulus presentation and data collection were achieved using the DMDX display system developed by K.I. Forster and J.C. Forster at the University of Arizona (Forster & Forster, 2003). Stimulus display was synchronized to the screen refresh rate (10.01 ms).

Each trial started with the presentation of a fixation signal (a plus sign) for 250 ms, in the center of the screen. It was replaced by a blank screen for 50 ms, then by a word or #s presented in one of four colors (red, yellow, green or blue) for a maximum of 2,000 ms, or until the participant named the color. After the participant's response, the screen went blank for 816 ms after which the next trial started. All stimuli were presented in Arial 10 font. The experimenter sat next to the participant and noted errors in a response sheet. Participants were given no feedback during the experiment.

RESULTS

In this and subsequent experiments, correct RTs and error rates were analyzed according to the following procedure. In the analysis of RTs, we first examined the shape of the RT distribution for correct trials, and excluded those faster than 250 ms as outliers (Most of the fast outliers were voice key trigger errors). In Experiment 1, 145 data points (out of 11,193 trials, 1.3%) were identified as outliers. In Experiment 1 and 2, the fixed factors were Distractor type (Color-associated, Control word, Neutral) and Neutral proportion (High, Low).

We analyzed the RT data in two ways, first using linear mixed effects model with log-transformed data with subjects and stimuli as crossed random factors (Baayen, 2008), and the second, using the untransformed, mean RT from each condition for each subject in a standard ANOVA (F_1 analysis). Linear mixed effects modelling requires RT data to be log- (or inverse/reciprocal-) transformed, to meet the distributional assumptions. These transformations reduce differences at slower RTs, and because of this, concern has been expressed that linear mixed effects modelling may fail to detect an overadditive interaction (Balota, Aschenbrenner, & Yap, 2013). In our analyses, both methods yielded the same pattern in all experiments. Thus, for brevity, and because F_1 analysis is more standard for Stroop experiments in which a small set of stimuli are used repeatedly (and where the within-category item differences are of lesser interest than in psycholinguistic experiments), here we report the F_1 analysis.

RT. The mean correct RT and error rates are shown in Table 1.

--- Insert Table 1 about here ---

RT were analyzed as a 3 (Distractor type: Color-associated word, Control word, Neutral) x 2 (Neutral proportion: High vs. Low) factorial design, and the Semantic interference effect was calculated as the difference between the Color-associated words and the Control word distractors, and the Word interference effect was calculated as the difference between the Control word and Neutral distractors. Averaged over the Hi- and Low neutral proportion conditions, the Semantic interference effect was significant, F(1, 38) = 10.822, η_p^2 = .22, p < .01, as was the Word interference effect, F(1,38) = 86.639, η_p^2= .70, p < .001. The main effect of neutral proportion was non-significant, F(1,38) = 3.695, η_p^2= .09, p = .062. Importantly, Word interference interacted with Neutral proportion, F(1,38) = 10.86, η_p^2= .22, p < .01, with an increased Word interference effect in the High neutral proportion condition, consistent with the view that the neutral proportion manipulation was successful in relaxing attentional control. However, the Semantic interference effect did not interact significantly with Neutral proportion, F(1,38) = 2.332, η_p^2= .06, p = .13.

Error rate. Error rates were analysed with linear mixed effects model with subjects and stimuli as crossed random factors, using the logit function appropriate for categorical variables (Jaeger, 2008). In all experiments, the model tested was: Error rate ~ Neutprop * Distractor type + (1 | subject) + (1 | stimulus). The Neutprop factor was contrast-coded (-.5, .5), and the Distractor type was referenced to the Control word to calculate the Semantic interference effect and the Word interference effect, as for RT. The only significant effect was the semantic interference effect, Z = 2.356, p < .02. Error rates generally mirrored the pattern for RTs.

Experiment 2 (Color-associated words, Manual response)

Experiment 2 was identical to Experiment 1, except for the type of response: Instead of naming the colors orally, participants identified the colors by means of a key press.

Method

Participants. Fifty students from Macquarie University participated in the experiment in return for course credit. Twenty-five were assigned to the high-neutral proportion condition and the other twenty-five to the low-neutral proportion condition, in the order of arrival.

Design. The experimental design was identical to Experiment 1, except that in this experiment participants identified the colors by means of a manual key press.

Materials. The stimulus materials were identical to Experiment 1.

Apparatus and Procedure. The apparatus and the general procedure were identical to Experiment 1, except that participants identified the color by means of a manual key press. They were instructed to press Z for red, X for yellow, N for green and M for blue. Key assignment was explained with reference to a card with four colored circles arranged in a horizontal line (red circle in the leftmost position and blue circle in the rightmost position), and the card remained in view during the key assignment training trials (8 trials containing colored string of #s) and the 32 practice trials. Participants were tested individually, or in pairs.

RESULTS

The general procedure for the analysis of RT was identical to Experiment 1. In Experiment 2, no data point was excluded (out of 13694 trials) as an outlier (faster than 250 ms).

The mean correct RT and error rates are shown in Table 2.

--- Insert Table 2 about here ---

RT. As in Experiment 1, RT were analyzed as a 3 (Distractor type: Color-associated word, Control word, neutral) x 2 (Neutral proportion: High vs. Low) factorial design, and Semantic interference effect was calculated as the difference between the Color-associated words and the Control word distractors, and the Word interference effect was calculated as the difference between the Control word and Neutral distractors. Averaged over the Hi- and Low neutral proportion conditions, the Semantic interference effect was significant, F(1, 48) = 4.775, ŋ$_p^2$= .09, p < .05, as was Word interference effect, F(1,48) = 27.24, ŋ$_p^2$= .36, p < .001. The main effect of neutral proportion was non-significant, F(1,48) < 1.0. As in Experiment 1, the Word interference effect interacted with Neutral proportion, F(1,48) = 14.07, ŋ$_p^2$= .23, p < .001, but, the Semantic interference effect again did not interact with Neutral proportion, F(1,48) < 1.0, p = .456.

For error rate, the same linear mixed effect model as Experiment 1 (Error rate ~ Neutprop * Distractor type + (1 | subject) + (1 | stimulus) using the logit function) was tested. As in Experiment 1, the Neutprop factor was contrast-coded (-.5, .5), and the Distractor type was referenced to the Control word. None of the main or interaction effects reached significance, all p > .41. As can be seen in Table 2, there was little difference in error rate between the experimental conditions.

Combined analysis of Experiment 1 and 2. The RT data from Experiments 1 and 2 were combined, and analysed as a 2 (Task: Oral vs. Manual) x 3 (Distractor type: Color-associated, Control word, Neutral) x 2 (Neutral proportion: High vs. Low) factorial design. Averaged over the two experiments, the Semantic interference effect (Color associated – Control word) was significant, F(1,86) = 15.34, ŋ$_p^2$= .15, p < .001, as was the Word interference effect (Control word – Neutral), F(1,86) = 113.37, ŋ$_p^2$= .57, p < .001. Word interference interacted with Task, indicating that the effect was larger in the Oral task (Experiment 1) than in the Manual task (Experiment 2): F(1,86) = 16.15, ŋ$_p^2$= .16, p < .001. The Semantic interference effect did not interact with Task: F(1,86) = 1.036, ŋ$_p^2$= .01, p = .31. Importantly, the interaction between Semantic interference and Neutral proportion was non-significant, F(1,86) = 2.70, ŋ$_p^2$= .03, p =.10. This was not qualified by the task: Task x Semantic interference x Neutral proportion:

F(1,86) < 1.0. Thus, our attempt to magnify the semantic interference effect was unsuccessful, for both the oral task and the manual task. Word interference interacted with Neutral proportion: F(1,86) = 24.78, $ŋ_p^2$= .22, p < .001, and this interaction was not modulated by Task: Task x Word interference x Neutral proportion: F(1,86) < 1.0. This indicated that for both the oral and manual Stroop tasks, the Neutral proportion manipulation was successful in relaxing attentional control over the task of reading.

In the Introduction, we expressed a concern with previous studies reporting no modulation of the semantic interference effect that they have not provided evidence that the effect had enough power to detect an interaction. To address this concern in our own study, we calculated the Bayes factor using the BayesFactor package (version 0.9.12-2, Morey & Rouder, 2015). Bayes factor is an odds ratio, indicating the relative amount of evidence for two (mutually exclusive) hypotheses, with 1 indicating equal evidence. Jeffreys (1961) recommends that odds greater than 3 be considered “some evidence,” odds greater than 10 be considered “strong evidence,” and odds greater than 30 be considered “very strong evidence” for one hypothesis over another. Bayes factors are particularly useful when the effect in question is non-significant: A large Bayes factor in favor of the null suggests the effect is likely to be absent, whereas a small Bayes factor for the null suggests that the experiment is insensitive (see Dienes, 2014). In this combined analysis, the Bayes factor for the non-significant interaction between semantic interference effect and neutral proportion (averaged over the oral and manual tasks) was 0.3 (or 1.9 in favor of the null), indicating that there was only equivocal evidence for the null. We also calculated the Bayes factor for the semantic interference averaged over the neutral proportion conditions: In the oral task, the Bayes factor was 11, and in the manual task, the

Bayes factor was 1.5. (Note however that the Semantic interference x Task interaction was non-significant, and the Bayes factor for this interaction was 0.29 (or 3 for the null interaction).

Discussion of Experiment 1 and 2

Experiment 1 (oral response) and Experiment 2 (manual response) used color-associated words presented in incongruent colors (e.g., LEMON in blue) and tested whether the semantic Stroop effect (measured relative to color-unassociated control words e.g., MERCY) can be modulated by the proportion of non-readable neutral distractors (a row of #s). The results were negative: Although there was a tendency for the size of semantic interference effect to be magnified in the High- relative to Low- neutral proportion condition (29 ms vs. 10 ms in Experiment 1; 14 ms vs. 8 ms in Experiment 2), the interaction with neutral proportion was not significant. However, the Bayes factor indicated that there was only equivocal evidence for the absence of an interaction, and hence it does not constitute a strong case for the automaticity of semantic retrieval. The failure to modulate the semantic interference effect was not because the neutral proportion manipulation was ineffective in modulating the control over the task of reading. The word interference effect (the interference caused by control word distractors relative to the non-readable row of #s) was magnified significantly in the high-neutral proportion condition relative to the low-neutral proportion condition (72 ms vs. 34 ms respectively in Experiment 1 and 43 ms vs. 5 ms respectively in Experiment 2). Rather, the absence of interaction seems to be due to the fact that the semantic interference effect is not robust. Numerically, the effect was not large, particularly in the manual task. Consistent with this, in the oral task, the Bayes factor was 11,but in the manual task, the Bayes factor was 1.5.[5] We will

[5] Recently, Levin and Tzelgov (2016) examined the semantic interference effect using color-associated words in Hebrew and Russian, using the manual Stroop task. They also reported that the effect was “small and unstable”,

consider possible reasons why the semantic interference effect with color-associated words is not very strong, particularly in the manual task, in the General Discussion. For now, we note that given that the Bayes factor analysis suggested that there was only equivocal evidence for the lack of an interaction, it would seem premature to accept the null interaction between the semantic interference effect and neutral proportion as evidence that semantic retrieval is automatic.

For our next experiments, we sought a different manipulation that would produce a stronger semantically-based interference effect. In a classic study using the oral Stroop task, Klein (1964) reported that color words that are not in the response set (e.g., the word PURPLE when the response colors are red, green, yellow and blue) produced greater interference than the color-associated words like SKY. Sharma and McKenna (1998) replicated this with manual responding (see also Brown & Besner, 2001, for reanalysis of their data). As these words are not the names of response colors, the concerns raised by Neely and Kahan (2001) relating to the interference reflecting “nonsemantic response competition” should not apply. We therefore chose as our semantic interference manipulation color names that were not the response colors.

Because none of the color names are the response colors, they are necessarily all “incongruent”. However, the specific pairing of a color name distractor with a response color can vary in conceptual similarity, such that the color denoted by the distractor can be close to the response color in the color space (e.g., the word YELLOW presented in orange), or distant (e.g., the word YELLOW presented in blue). Which of the (close vs. distant) pairing should be used to index the semantic interference effect?

There are not many studies that have examined the effect of color-space distance, and the extant literature paints a mixed picture. For example, using an oral Stroop task, Klopfer (1996)

with the Bayes factor generally indicating “anecdotal evidence” in conditions comparable to the present Experiment 2.

examined the effect of color space distance between each of five response colors (yellow, green, orange, blue, and purple) and their corresponding names as distractors and reported finding an *inhibitory* effect (e.g., the color orange was named more slowly when it was presented in the word YELLOW than in the word BLUE). In contrast, Flowers and Blair (1976) required participants to sort incongruent color-word stimuli into two by color categories (i.e., a manual Stroop task) and found that responses were facilitated by high intra-category similarity (red-orange-yellow and green-blue-purple) relative to low (red-yellow-blue and orange-green-purple). From this, it may be expected that the effect of color space distance would be facilitatory in the oral Stroop task and inhibitory in the manual Stroop task. However, these results may be specific to color name distractors that are the response colors. Notwithstanding this empirical uncertainty, the "distant" condition is conceptually more like the standard "incongruent" condition. Accordingly, in the present experiments we compared the control word distractors to the ”distant” pairing to index semantic interference.

To recap, in Experiment 3 (oral Stroop task) and Experiment 4 (manual Stroop task) we used words which were color names that were not the response colors to index semantic interference. The manipulation of attentional control was the same as in Experiments 1 and 2, namely, the proportion of non-readable neutral trials.

Experiment 3 (Color-names, Oral response)

Method

Participants. Forty students from Macquarie University participated in the experiment in return for course credit. Twenty-one were assigned to the High-neutral proportion condition and 19 to the Low-neutral proportion condition, in the order of arrival.

Design. The experimental design was the same as Experiment 1, using the Stroop color naming task, and involving the factor Distractor type (Semantic, Word, Neutral), manipulated within-subjects, and their relative proportion (75% neutral trials or 25% neutral trials), manipulated between groups. The only difference was that instead of color-associated words like LEMON and SKY, color names that were not the response colors were used as the semantic distractor. The dependent variables were color naming latency and error rate.

Materials. In this experiment (and Experiment 4), the response colors were red (RGB 255 000 000), blue (RGB 000 000 255), orange (RGB 255 165 000) and white (RGB 255 255 255), presented against a black background, and the color-name distractors were GREEN, YELLOW, GREY and PINK. As in Experiment 1 and 2, none of the word distractors shared the initial phoneme with the names of response colors. The control noncolor-name words were TWICE, WINNER, GRIP and TANK, and they were matched with the color name words on length, and as close as possible on syllable structure, word frequency (mean 88.6 for the color names and 29.1 for the noncolor-name words, range 6.7-154 per million based on Celex frequency, Baayen, et al., 1995; mean 36 for the color-names and 32.3 for the noncolor-name words, range 3.2-320 per million based on Subtitle frequency, Brysbaert & New, 2009), the number of orthographic neighbors (as defined by the N metric, Coltheart, et al., 1977, mean 6.3 for the color name words and 5.8 for the noncolor-name words, range 1 - 14), and position-dependent bigram frequency (mean 38.4 and 69.4, range 12.7-200.4). The neutral distractors were a string of #s matched to the number of letters, i.e., ####, ##### or ######.

In Experiment 3 (and 4), each color name word distractor was paired with two response colors. One was closer perceptually to the color depicted by the color name and the other was perceptually more distant, for example, the word YELLOW was presented in orange (close) and blue (distant). The *Close* pairings were the word GREEN in blue, YELLOW in orange, GREY in white and PINK in red; the *Distant* pairings were GREEN in red, YELLOW in blue, GREY in orange and PINK in white. The matched control noncolor-name word distractors were similarly paired with two response colors each.

Both the High neutral proportion and Low neutral proportion groups received 384 Stroop color naming trials. In the High neutral proportion condition, 288 trials (75%) contained the neutral distractor, and 48 trials contained color-names (with half containing the Close color-word pairings and the other half containing the Distant color-word pairings) and 48 trials contained control words as distractors. In the Low neutral proportion condition, 96 trials (25%) contained the neutral distractor, 144 trials contained color names (with half containing the Close color-word pairings and the other half containing the Distant color-word pairings) and 144 trials contained the control words as distractors. Each list of 384 trials was divided into six sublists of 64 trials with each sublist containing the representative proportion of word and #s trials, the three distractor types, and the four response colors occurring equally often. A pseudo-random order of trials was generated for each sublist such that the same color did not occur in succession. The order of response colors was identical for the High- and Low proportion neutral conditions.

Apparatus and Procedure. The apparatus and the general procedure were identical to Experiment 1. Each participant completed 384 color naming trials, presented in six blocks (with each block containing 64 trials) with a self-paced break between the blocks. A practice block of

48 trials containing each of the four colors occurring equally often, and containing the same proportion of word and neutral trials as the test blocks preceded the test blocks.

Participants were instructed at the outset of the experiment that on each trial they would be presented with a word or a row of #s presented in one of four colors, red, orange, blue or white, and their task was to name the color of the stimulus, as fast and accurately as possible.

RESULTS

The general procedure for the analysis of RT and error rates was identical to Experiment 1.

RT. In the analysis of RTs, the preliminary treatment of RT data excluded 158 data points (out of 14873 trials) faster than 250 ms as outliers.

The mean correct RT and error rates are shown in Table 3.

--- Insert Table 3 about here ---

RT were analyzed as a 3 (Distractor type: Color name-distant, Control word, Neutral) x 2 (Neutral proportion: High vs. Low) factorial design. The Semantic interference effect was indexed by the difference between the Color name-distant and Control word distractors, and the Word interference effect was indexed by the difference between the Control word and Neutral distractors. Averaged over the Hi- and Low neutral proportion conditions, the Semantic interference effect was highly significant, F(1, 38) = 62.098, ŋ$_p^2$= .62, p < .001, as was Word interference effect, F(1,38) = 121.72, ŋ$_p^2$= .76, p < .001. The main effect of neutral proportion was non-significant, F(1,38) < 1.0. As in previous experiments, the Word interference effect interacted with Neutral proportion, F(1,38) = 26.40, ŋ$_p^2$= .41, p < .01, indicating that the neutral proportion manipulation was successful in relaxing attentional control and increasing the interference caused by the word distractors. However, the Semantic interference effect did not interact with Neutral proportion, F(1,38) = 1.88, $ŋ_p^2$ = .05, p = .18.

The effect of color space distance was indexed by the difference between the Colorname-distant and Colorname-close conditions. The main effect of the color space distance was non-significant, F(1,38) = 1.53, $ŋ_p^2$= .04, p = .22, and it did not interact with neutral proportion, F(1,38) < 1.0.

Error rate. For error rate, the same linear mixed effect model as Experiment 1 (Error rate ~ Neutprop * Distractor type + (1 | subject) + (1 | stimulus) using the logit function) was tested. As for the RT, the Neutprop factor was contrast-coded (-.5, .5), and the Distractor type was referenced to the Control word. Averaged over the neutral proportion conditions, both the semantically-close and semantically distant words produced more errors than the control word: Z = 2.167, p < .009, and Z = 2.586, p < .01, respectively. No other main or interaction effects reached significance, all p > .09.

Experiment 4 (Color-names, Manual response)

Method

Participants. Forty-one students (36 females, 5 males, mean age 21.0 years) from Macquarie University participated in the experiment in return for course credit. Twenty were assigned to the high-neutral proportion condition and 21 to the low-neutral proportion condition, in the order of arrival.

Design. The experimental design and stimuli were identical to Experiment 3, except that in this experiment participants identified the colors by means of a manual key press.

Materials. The stimulus materials were identical to Experiment 3.

Apparatus and Procedure. The apparatus and the general procedure were identical to Experiment 3, except that participants identified the color by means of a manual key press. They were instructed to press Z for red, X for orange, N for white and M for blue. As in Experiment 2, key assignment was explained with reference to a card with four colored circles arranged in a horizontal line (red circle in the leftmost position and blue circle in the rightmost position), and the card remained in view during the key assignment training trials (8 trials containing colored string of #s) and the 48 practice trials. Participants were tested individually, or in pairs.

RESULTS

The general procedure for the analysis of RT and error rates was identical to Experiment 1.

RT. In the analysis of RTs, the preliminary treatment of RT data excluded no data points (out of 14941 trials) faster than 250 ms as outliers.

The mean correct RT and error rates are shown in Table 4.

--- Insert Table 4 about here ---

RT were analyzed as a 3 (Distractor type: Color name-distant, Control word, Neutral) x 2 (Neutral proportion: High vs. Low) factorial design, and the Semantic interference effect was calculated as the difference between the Color name-distant and Control word distractors, and the Word interference effect was calculated as the difference between the Control word and Neutral distractors. Averaged over the Hi- and Low neutral proportion conditions, the Semantic interference effect was significant, F(1, 39) = 19.65, ŋ$_{p}^{2}$= .33, p < .001. In this experiment, averaged over the neutral proportion conditions, the Word interference effect was non-significant, F(1,39) = 2.044, ŋ$_{p}^{2}$= .05, p =.16. The main effect of neutral proportion was non-significant, F(1,39) < 1.0. As in previous experiments, the Word interference effect interacted with Neutral proportion, F(1,39) = 26.63, ŋ$_p^2$= .41, p < .001. Critically, in this experiment, the interaction between Semantic interference and Neutral proportion was significant, F(1,39) = 4.66, ŋ$_p^2$= .11, p < .04.

The main effect of color space distance was non-significant, F(1,39) < 1.0. The interaction with neutral proportion did not reach significance, F(1,39) = 2.95, ŋ$_p^2$= .07, p = .09.

Error rate. For error rate, the same linear mixed effect model as Experiment 1 (Error rate ~ Neutprop * Distractor type + (1 | subject) + (1 | stimulus) using the logit function) was tested. As for the RT, the Neutprop factor was contrast-coded (-.5, .5), and the Distractor type was referenced to the Control word. The model did not converge; however, it can be seen from Table 4 that the error rate data mirrored the RT data.

Combined analysis of Experiment 3 and 4. The RT data from Experiments 3 and 4 were combined, and analysed as a 2 (Task: Oral vs. Manual) x 4 (Distractor type: Color name-close, Color name-distant, Control word, Neutral) x 2 (Neutral proportion: High vs. Low) factorial design. Averaged over the two experiments, the Semantic interference effect (Color name-distant – Control word) was highly significant, F(1,77) = 78.49, ŋ$_p^2$= .51, p < .001, as was the Word interference effect (Control word – Neutral), F(1,77) = 62.23, ŋ$_p^2$= .45, p < .001. Both effects interacted with Task, indicating that the effects were larger in the Oral task (Experiment 3) than in the Manual task (Experiment 4): Semantic interference x Task: F(1,77) = 8.99, ŋ$_p^2$= .11, p < .01; Word interference x Task: F(1,77) = 31.83, ŋ$_p^2$ = .29, p < .001. Critically, the interaction between Semantic interference and Neutral proportion was significant, F(1,77) = 6.05, ŋ$_p^2$= .07, p < .02, providing support for the claim that semantic retrieval is modulated by attention. The Bayes factor for the interaction was 1.9. This interaction was not qualified by the task: Task x Semantic interference x Neutral proportion: F(1,77) < 1.0. As in previous experiments, Word interference interacted with Neutral proportion: F(1,77) = 51.92, $ŋ_p^2$= .40, p < .001, and this interaction was not qualified by the task: Task x Word interference x Neutral proportion: F(1,77) < 1.0. As in Experiment 1 and 2, Bayes factor was calculated for the semantic interference effect averaged over the neutral proportion conditions: In the oral task, the Bayes factor was 5478257; in the manual task, 133, indicating very strong evidence for the semantic interference effect.

The main effect of color space distance did not reach significance, F(1,77) = 2.42, $ŋ_p^2$ = .03, p = .12. Its interaction with neutral proportion was marginal, F(1,77) = 3.54, $ŋ_p^2$= .04, p = .06. The Color space distance x Task interaction was not significant, F(1,77) < 1.0, nor was the Color space distance x Task x Neutral proportion, F(1,77) < 1.0.

In summary, Experiments 3 and 4 showed that with color names that were not in the response set, the semantic interference effect – the difference between color names and control words – was highly robust, and produced an interaction with the proportion of non-readable, neutral trials. The latter finding indicates that the semantic Stroop effect is under attentional control. The interaction between semantic interference and neutral proportion was not qualified by the task, indicating that the same pattern was observed in the oral and manual Stroop tasks, even though the magnitude of semantic interference effect was larger overall in the oral task than the manual task. The color distance between the response color and the color name tended to increase the size of interference in both the oral task and the manual task, but this effect did not reach significance.

General Discussion

In four Stroop experiments, we investigated the claim that semantic retrieval is automatic, in the sense that it is not controllable by endogenous attention. We tested whether the semantic interference effect – indexed by the interference produced by words that are related to color but that are not the names of the response colors, relative to color-unrelated control words – is modulated by the proportion of non-readable, neutral distractors (a row of #s). The rationale for manipulating the neutral proportion was based on the view that one basis of conflict in the Stroop task is the competition between the task set of reading and color processing which is driven in part by endogenous voluntary attention (Monsell, et al., 2001), and that the high proportion of non-readable, neutral trials relaxes the suppression of the task of reading (Goldfarb & Henik, 2007; Mills, 2017). Consistent with the assumption, and replicating the results reported by Mills (2017), increasing the proportion of neutral trials magnified the word interference effect (greater interference caused by color-unassociated words like MERCY relative to non-readable neutral distractors) in all four experiments, in both oral and manual Stroop tasks. Experiments 1 and 2 used words denoting objects associated with a specific color, presented in incongruent colors, for example, LEMON in blue. Here the semantic interference effect did not interact significantly with neutral proportion, whether the required response to color was oral (Experiment 1) or manual (Experiment 2). The Bayes factor analysis however indicated that the evidence for the null interaction was equivocal, suggesting that it does not provide a compelling evidence that semantic retrieval is automatic. In Experiment 3 (oral Stroop task) and Experiment 4 (manual Stroop task), we used a manipulation known to produce a larger semantic interference effect (Klein, 1964; Sharma & McKenna, 1998) - color names that are not the response colors. Here, the semantic interference effect interacted with neutral proportion[6]; and the interaction was not qualified by the task mode. We take this result as the first evidence that semantic retrieval in the Stroop task is controlled by endogenous attention. Below, we discuss what the specific aspects of the present results reveal about the control of semantic retrieval in the Stroop task.

Task conflict and informational conflict. Our manipulation of the proportion of non-readable neutral trials was based on the demonstration by Goldfarb and Henik (2007) that increasing the proportion of neutral trials magnified the classic Stroop interference effect with the response-incongruent color names. Of relevance to the argument that this manipulation modulated task conflict was the fact that it also produced a “reverse facilitation effect”, i.e., slower response to congruent trials (e.g., the word RED presented in red) relative to the neutral trials.

Following Goldfarb and Henik (2007), others tried to isolate task conflict and informational conflict, and to determine if they are independent (e.g., Entel, et al., 2015; Roelofs, 2012; Steinhauser & Hübner, 2009; see also Levin & Tzelgov, 2014, for a review). These studies used the standard congruent and incongruent color names and neutral stimuli. While there is a general consensus that the informational conflict can be indexed by the difference between the congruent and incongruent conditions, it has been more difficult to find agreement on how to measure task conflict independent of informational conflict. Goldfarb and Henik (2007) originally put forward the reverse facilitation effect as a marker of task conflict, however, Entel et al., (2015) argued that the reverse facilitation is not “an exhaustive marker of task

[6] A reviewer pointed out that in all four experiments, there was a trend towards an increased semantic interference effect in the high neutral proportion condition, even when the interaction was not significant (Experiment 1 and 2), and suggested calculating the evidence for the interaction over all four experiments. The Bayes factor for the semantic interference effect x neutral proportion interaction across all four experiments was 5 (with the associated $F(1,163) = 8.90$, $\eta_p^2 = .0022$, $p < .001$), consistent with the trend observed in each experiment.

conflict" (p.915), that is, its absence does not mean that task conflict was absent. Entel et al. (2015) suggested instead that the contrast between the color words (congruent and incongruent) and neutral serves as a "general marker of task conflict" (p.925). Steinhauser and Hübner (2009) referred to this contrast as the "bivalency cost", following its usage in the task-switching literature. Steinhauser and Hübner however noted that that using the bivalency cost as a measure of task conflict in the Stroop task is problematic, pointing out that averaging the congruent and incongruent conditions to estimate the bivalency cost works "only if it is assumed that facilitation by a congruent stimulus and interference by an incongruent stimulus is rather similar" (p. 1399), a condition rarely met in classic Stroop experiments -in the Stroop task, typically, interference effect is much larger than the facilitation effect.

We agree that in a classic Stroop task it is difficult to isolate the task conflict and information conflict. In the present experiments, we took the interference produced by color-unrelated control words relative to non-readable neutral stimuli as a marker of task conflict, on *a priori* grounds that the former can be read but the latter cannot. Replicating the results reported by Mills (2017), the word interference effect was magnified in the high neutral proportion condition, consistent with the view originally put forward by Goldfarb and Henik (2007) that increasing the proportion of non-readable neutral trials magnifies task conflict. Importantly, in the present study, the high neutral proportion also magnified the semantic interference effect. By definition, the semantic interference effect reflects informational conflict – the color names and the control words are both words hence are equated on task conflict, and they differ only on semantic features. Thus, this result provides the first evidence that manipulating task conflict also impacts informational conflict, i.e., the result indicates not only that the two types of conflict are not independent, but also that informational conflict is dependent on task conflict.

Task mode. In the present four experiments, the word interference effect was always greater in the oral task than the manual task. This is consistent with the result reported by Kinoshita, de Wit and Norris (2017). In a condition comparable to the present low-neutral proportion condition, they examined the interference caused by different types of word-like distractors (words, pronounceable pseudowords, consonant strings as well as incongruent color names in the response set) in an oral and manual Stroop task. Whereas in the oral task, all word-like distractors – in varying degrees - interfered relative to a row of Xs, in the manual task only the incongruent color names interfered.

Why are the word interference effects larger in the oral Stroop task than the manual task? Following Burt (2002) who posited that "a factor likely to be important in the color-naming interference observed in the non-color-word Stroop task is the vocal response requirement" (p. 1033), Kinoshita et al. (2017) suggested that it likely reflects the requirement to produce a speech output in the former, but not the latter. In a model of Stroop color naming couched within his WEAVER++ model of speech production, Roelofs (2003) noted an architectural difference in reading words and naming colors, namely, that "written words in alphabetical systems are intrinsically tied to their sounds, whereas colors are not" (p.6). Thus, word interference - more accurately, interference produced by word-like distractors including pseudowords and nonwords - is expected to be larger in the oral Stroop task than the manual Stroop task.

This provides a principled guideline for when to use the manual vs. oral Stroop task. Recall that Augustinova and Ferrand (2014) suggested that the oral Stroop task is to be preferred ("it (the manual Stroop task) is not suitable for future research aimed at the detailed examination of word reading and its contribution to overall Stroop interference", p. 346), but they did not provide a rationale other than that the interference effects are bigger in the former. Contrary to their recommendation, here we used both tasks and found that the modulation of the semantic interference effect did not depend on the task mode. This is not a surprising outcome when one considers the origin of task effects. It was noted above that it is the process of generating phonology from a printed letter string that interferes with responding to color in the oral, but not manual Stroop task. This is different from retrieving meaning from a word, which is responsible for producing the semantic interference effect. To put it differently, the task of reading is not unitary (Kinoshita, et al., 2017), and the oral and manual Stroop tasks are differentially sensitive to different aspect of reading. The oral Stroop task is to be clearly preferred if the reading process in question concerns the process of generating phonology from a printed letter string (see, e.g., Coltheart, Woollams, Kinoshita & Perry, 1999; Kinoshita & Sulpizio, submitted), but if the process in question concerns semantic processing, the task mode should not matter.

The “semantic gradient”: Color-associated words vs. non-response color names. In the present Experiment 1 and 2 using color-associated words like LEMON, the semantic interference effect was small (particularly in the manual task, with a Bayes factor of less than 2), and it did not produce a significant interaction with neutral proportion. In contrast, Experiment 3 and 4 used color names that are not in the response set, and they produced a larger semantic interference effect, which interacted significantly with neutral proportion. It has been known since Klein (1964, see also Sharma & McKenna, 1998) that color names not in the response set produce greater interference than the color-associated words, but there has been little discussion of the theoretical basis for this “semantic gradient”.

We believe the semantic gradient fits naturally with the view that semantic processing in the Stroop task is goal-directed. The goal of the Stroop color task is to name/identify the color.

Thus, the semantically-based interference reflects the conflict in semantic features between the response color and the meaning of the word, but only those that are diagnostic of color cause interference. Color names used in Experiments 3 and 4 obviously have such semantic features. In contrast, the color-associated words used in Experiment 1 and 2 like LEMON, are names of objects associated with a specific color. Not all of its semantic features (e.g., it is a fruit, it is sour) pertain to color, and we argue this is why they produce less interference than the color names.

The same line of argument (that semantic processing in the Stroop task is goal-directed) was advanced by Kinoshita et al. (2017) to explain the absence of lexicality effect in the Stroop task. In both the oral task and the manual task, (replicating Monsell et al's (2001) original finding with the oral Stroop task) Kinoshita et al. observed no difference in the amount of interference produced by words (e.g., HAT, STORM) and pseudowords (e.g., HIX, STASE). By definition, words but not pseudowords have a meaning; hence words, but not pseudowords are semantically incongruent with the response color. But if so, why do word distractors not interfere more than pseudoword distractors? Kinoshita et al. (2017) argued that it is because the semantic features of the words (e.g., it is inanimate, it is an article of clothing) are not diagnostic of color – they were specifically selected to be color-neutral, and therefore not relevant to the goal of semantic processing at hand, that of identifying the color.

Kinoshita et al. (2017) noted that the absence of a lexicality effect is at odds with the idea that word meanings are activated automatically and produce interference in the Stroop task. Similarly, under this “automatic semantic activation” view, it is hard to explain why color names not in the response set (e.g., GREEN in red when green is not a response color) interfere more than color-associated object names like PEA presented in red: Both types of words are assumed to activate semantic features that are incongruent with the target color. These findings indicate instead that semantic processing in the Stroop task is goal-directed (it specifically evaluates semantic features that are diagnostic of color), and the semantic features that are not diagnostic of color (e.g., peas are small; peas are edible; peas are round, etc) do not interfere with the naming of target color. Thus, both the finding of a semantic gradient and the absence of a lexicality effect on Stroop interference question the assumption that semantic Stroop effects reflect automatic semantic activation (Augustinova & Ferrand, 2014; Neely & Kahan, 2001).

Conclusion. Using color-related words that are not the names of response colors presented intact, the present study showed that the size of the semantic interference effect can be magnified by increasing the proportion of non-readable, neutral trials. On the assumption that the manipulation impacts on the control of the task of reading, we take the results as challenging the widely held view that the retrieval of meaning from words in the Stroop task is automatic, in the sense that it cannot be controlled (Augustinova & Ferrand, 2014; Neely & Kahan, 2001). In contradiction to Briony, the budding writer in *Atonement*, we conclude that reading a word and understanding it are “not the same thing”.

References

Allport, D., Styles, E.A., & Hsieh, S.(1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umlita & M. Moscovitch (Eds.), *Attention and performance XV. Conscious and unconscious information processing* (pp.421-452). Cambridge, MA: MIT Press.

Augustinova, M., & Ferrand, L. (2012). Suggestion does not de-automatize word reading: Evidence from the semantically based Stroop task. *Psychonomic Bulletin & Review, 19*, 521–527.

Augustinova, M., & Ferrand, L. (2014). Automaticity of word reading: Evidence from the semantic Stroop paradigm. *Current Directions in Psychological Science, 23*, 343-348.

Augustinova, M., Flaudias, V., & Ferrand, L. (2010). Single-letter coloring and spatial cuing do not eliminate or reduce a semantic contribution to the Stroop effect. *Psychonomic Bulletin & Review, 17*, 827-833

Baayen, R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (Release 2) CD-ROM. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Balota, D.A., Aschenbrenner, A.J. & Yap, M.J. (2013). Additive effects of word frequency and stimulus quality: The influence of trial history and data transformations. *Journal of Experimental Psychology: Learning, Memory, and Cognition 39*, 1563–1571

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., ..Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278.

Bates, D., M., Maechler, M., & Bolker, B. (2013). *Lme4: Linear mixed-effects models using S4 classes*. R package version 0.999999-2.

Besner, D. (2001). The myth of ballistic processing: Evidence from Stroop's paradigm. *Psychonomic Bulletin & Review, 8*, 324-330.

Besner. D., & Stolz, J.A. (1999). Context dependency in Stroop's paradigm: When are words treated as nonlinguistic objects? *Canadian Journal of Experimental Psychology, 53*, 374-380.

Besner, D., Stolz, J. A., & Boutilier, C. (1997). The Stroop effect and the myth of automaticity. *Psychonomic Bulletin & Review, 4*, 221-225.

Brown, M., & Besner, D. (2001). On a variant of Stroop's paradigm: Which cognitions press your buttons? *Memory & Cognition, 29*, 903-904.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera & Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior and Research Methods, 41*, 977-990.

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

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

Coltheart, M., Woollams, A., Kinoshita, S., & Perry, C. (1999). A position-sensitive Stroop effect: Further evidence for a left-to-right component in print-to-speech conversion. *Psychonomic Bulletin & Review*, *6*, 456-463.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(July), 1–17.

Entel, O., Tzelgov, J., Bereby-Meyer, Y., & Shahar, N (2015). Exploring relations between task conflict and informational conflict in the Stroop task. *Psychological Research, 79*, 913–927.

Flowers. J.lI., & Blair. B. (1976). Verbal interference with visual classification: Optimal processing and experimental design. *Bulletin of the Psychonomic Society, 7*. 260-262.

Forster, K.I., & Forster, J.C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods Instruments and Computers, 35*, 116–124.

Goldfarb, L., Aisenberg, D., & Henik, A. (2011). Think the thought, walk the walk: Social priming reduces the Stroop effect. *Cognition, 118*, 193–200.

Goldfarb, L., & Henik, A. (2007). Evidence for task conflict in the Stroop task. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1170-1176.

Jaeger, T.F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press, Clarendon Press.

Kinoshita, S., De Wit, B., & Norris, D. (2017). The magic of words reconsidered: Investigating the automaticity of reading color-neutral words in the Stroop task. *Journal of Experimental Psychology: Learning, Memory and Cognition, 43*, 369-384.

Kinoshita, S., & Sulpizio, S. (2017). The locus of serial effect in reading: Insights from Stroop color naming. Manuscript submitted for publication.

Klein, G.S. (1964). Semantic power measured through the interference of words with color-naming. *American Journal of Psychology, 77*, 576-588.

Klopfer, D.S. (1996). Stroop interference and color-word similarity. *Psychological Science, 7,* 150-157.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2013). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). Version 2.0–11. http://CRAN.R–project.org/package=lmerTest

Lachter, J., Forster, K.I., & Ruthruff, E. (2004). Forty-five years after Broadbent (1958): Still no identification without attention. *Psychological Review, 111*, 880–913.

Labuschagne, E.M., & Besner, D. (2015). Automaticity revisited: When print doesn't activate semantics. *Frontiers in Psychology*, *6,* Article 117, 1-7.

Levin, Y., & Tzelgov, J. (2014). Conflict components of the Stroop effect and their "control". *Frontiers in Psychology,* Article 463. 1-5, doi: 10.3389/fpsyg.2014.00463

Levin, Y., & Tzelgov, J. (2016). What Klein's "semantic gradient" does and does not really show: Decomposing Stroop interference into task and informational conflict components. *Frontiers in Psychology*, *7*, Article 249, 1-16, doi: 10.3389/fpsyg.2016.00249.

Lien, M-C, Ruthruff, E., Kouchi, S, & Lachter, J. (2010). Even frequent and expected words are not identified without spatial attention. *Attention, Perception & Psychophysics*, *72*, 973-988.

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

Manwell, L. A., Roberts, M. A., & Besner, D. (2004). Single letter coloring and spatial cuing eliminates a semantic contribu- tion to the Stroop effect. *Psychonomic Bulletin & Review, 11*, 458–462.

McEwan, I. (2001). *Atonement*. London: Jonathan Cape.

Medler, D.A., & Binder, J.R. (2005). MCWord: An on-line orthographic database of the English language. http://www.neuro.mcw.edu/mcword/

Mills, L. (2017). Does neutral proportion modulate attentional control of task conflict in the Stroop task? Unpublished Masters thesis, Macquarie University.

Monsell, S., Taylor, T.J., & Murphy, K. (2001). Naming the color of a word: Is it response or task sets that compete? *Memory & Cognition, 29*, 137-151.

Moors, A., & De Houwer, J. (2006). Automaticity: A theoreti- cal and conceptual analysis. *Psychological Bulletin, 132*, 297–326.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-2. Retrieved from http://CRAN.R-project.org/package= BayesFactor

Neely, J.H., & Kahan, T.A. (2001). Is semantic activation automatic? In H.L. Roediger, III J.S. Nairne, I. Neath & A.M. Surprenant (Eds). *The nature of remembering: Essays in honor of Robert G. Crowder*. (pp. 69-93). Washington, DC, US: American Psychological Association.

R Core Team (2014). *R: A language and environment for statistical computing*. RFoundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raz, A., Kirsch, I., Pollard, J., & Nitkin-Kaner, Y. (2006). Suggestion reduces the Stroop effect. *Psychological Science, 17*, 91–95.

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review, 110*, 88–125.

Roelofs, A. (2012) Attention, spatial integration, and the tail of response time distributions in Stroop task performance, *Quarterly Journal of Experimental Psychology, 65*, 135-150

Rogers, R.D., & Monsell, S. (1995). The costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*, 207-231.

Sharma, D., & McKenna, F.P. (1998). Differential components of the manual and vocal Stroop tasks. *Memory & Cognition, 26*, 1033-1040.

Steinhauser, M., & Hübner, R. (2009). Distinguishing response conflict and task conflict in the Stroop task: Evidence from ex-Gaussian distribution analysis. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 1398-1412.

Stolz, J., & Besner, D. (1999). On the myth of automatic semantic activation in reading. *Current Directions in Psychological Science, 8*, 61-65.

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

Tzelgov, J. (1997). Specifying the relations between automaticity and consciousness: A theoretical note. *Consciuosness and Cognition, 6,* 441–451. doi: 10.1006/ccog.1997.0303

Table 1.

Mean Color Naming Latencies (RT, in ms) and Percent Error Rates (%E) in Experiment 1 (Oral response)

		Neutral Proportion			
		High		Low	
Distractor type	Example	RT	%E	RT	%E
	(response color = blue)				
Color-associated word	LEMON	667 (24)	4.4	592 (24)	3.8
Control word	MERCY	638 (23)	3.1	582 (23)	2.7
Neutral	#####	567 (21)	2.4	548 (21)	1.9
Semantic interference effect (Color-associated – Control)		29 (8.3)	1.3	10 (8.3)	1.1
Word interference effect (Control word – neutral)		71 (7.7)	.7	34 (7.7)	.8

Standard error of the mean in parentheses

Table 2.

Mean Color Identification Latencies (RT, in ms) and Percent Error Rates (%E) in Experiment 2 (Manual response)

		Neutral Proportion			
		High		Low	
Distractor type	Example	RT	%E	RT	%E
	(response color = blue)				
Color-associated word	LEMON	686 (21)	4.4	691 (21)	4.8
Control word	MERCY	672 (20)	5.1	683 (20)	5.5
Neutral	#####	629 (19)	4.7	678 (19)	4.6
Semantic interference effect (Color-associated – Control)		14 (7.5)	-.7	8 (7.5)	-.7
Word interference effect (Control word – Neutral)		43 (6.9)	.4	5 (6.9)	.9

Standard error of the mean in parentheses

Table 3.

Mean Color Naming Latencies (RT, in ms) and Percent Error Rates (%E) in Experiment 3 (Oral response)

		Neutral Proportion			
		High		Low	
Distractor type	Example	RT	%E	RT	%E
	(Response color = red)				
Color-name - close	PINK	692 (24)	6.4	682 (26)	4.2
Color-name - distant	GREEN	708 (24)	5.0	684 (25)	5.5
Control word	TWICE	639 (21)	3.1	634 (22)	3.1
Neutral	#####	577 (20)	2.2	612 (21)	2.6
Semantic distance effect (Distant – Close)		17 (11.5)	1.4	2 (12.1)	-.7
Semantic interference effect (Color name-distant – Control word)		69 (10.1)	1.9	50 (10.6)	1.6
Word interference effect		62 (6.1)	.9	22 (6.4)	.5

Standard error of the mean in parentheses

Table 4.

Mean Color Identification Latencies (RT, in ms) and Percent Error Rates (%E) in Experiment 4 (Manual response)

Neutral proportion		High		Low	
Distractor type	Example	RT	%E	RT	%E
	Response color = red				
Color-name - close	PINK	646 (25)	4.4	613 (24)	4.6
Color-name - distant	GREEN	669 (25)	7.9	607 (24)	4.8
Control word	TWICE	625 (21)	7.1	591 (21)	4.3
Neutral	#####	594 (21)	5.2	609 (20)	5.1
Semantic distance effect (Distant – Close)		23 (11.8)	3.5	-6 (11.5)	-.2
Semantic interference effect (Color name-distant – Control word)		44 (10.3)	.8	16 (10.1)	.5
Word interference effect (Control word – Neutral)		31 (6.2)	1.9	-18 (6.1)	-1.2

Standard error of the mean in parentheses