Running head: Nature of Stroop Reading

Depth of Processing in the Stroop Task: Evidence from a Novel Forced-Reading Condition

Ami Eidels, Kathryn Ryan, Paul Williams

University of Newcastle, Australia

and

Daniel Algom

Tel Aviv University

Address for correspondence:
Ami Eidels
School of Psychology
University of Newcastle
Callaghan NSW 2308
Australia
E mail: ami.eidels@newcastle.edu.au

Abstract

The presence of the Stroop effect betrays the fact that the carrier words were read in the face of instructions to ignore them and to respond to the target ink colours. In this study, we probed the nature of this involuntary reading by comparing colour performance with that in a new forced-reading Stroop task in which responding is strictly contingent on reading each and every word. We found larger Stroop effects in the forced-reading task than in the classic Stroop task and concluded that words are processed to a shallower level in the Stroop task than they are in routine voluntary reading. The results show that the two modes of word processing differ in systematic ways and are conductive to qualitatively different representations. These results can pose a challenge to the strongly automatic view of word reading in the Stroop task.

Key words:

Stroop Effect, Selective Attention, Forced Reading, Depth of Processing

The most compelling demonstration of the mandatory nature of word reading in the laboratory is the Stroop effect (Stroop, 1935; see MacLeod, 1991, and Melara & Algom, 2003, for reviews). Presented with colour words printed in colour, it takes people longer to name the colour of an incongruent stimulus (e.g., the word GREEN printed in red) than that of a congruent stimulus (RED in red). The presence of the Stroop effect -- the difference in mean response time (RT) between incongruent and congruent combinations -- documents the fact that people succumbed to the overwhelming tendency to read the words in the face of instructions to ignore them and concentrate instead on the target ink-colours. Moreover, this reading involves (a modicum at the least of) semantic processing, otherwise congruent and incongruent stimuli lack psychological reality. Because the Stroop effect proved omnipresent, the view that reading and semantic processing is activated upon exposure to a word for any purpose has gained currency (e.g., Bargh, 1992; Bargh & Chartrand, 1999; Brown, Gore, & Carr, 2002; but see Bugg & Hutchison, 2013). It is this view that comes under scrutiny in the present study.

Recognition of the multifaceted nature of reading formed our point of departure. A word can be read via different routes, entailing orthographic, phonological, or semantic analyses. Completion of all processes is not strictly needed for performing various reading tasks including word recognition, pronouncing, or word identification (e.g., Coltheart, Rastle, Perry, Ziegler, & Langdon, 2001; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Evans, Lambton-Ralph, & Woollams, 2012). A widely agreed corollary is that a word can be read to differing levels of semantic analysis. Processing can be deep, entailing full semantic analysis, but it can also be shallower (Craik & Lockhart, 1972; Roediger, Gallo, & Geraci, 2002), yet yielding acceptable performance nonetheless in the pertinent tasks of reading and memory. Our question was this. What level of word reading is accomplished in the standard Stroop task? The task-irrelevant words are clearly read, otherwise there would not be a Stroop effect,

but the depth of this reading process is moot. Is the nature of word reading performed in the Stroop task (when the person actually attempts to *ignore* the word) comparable to that accomplished in more routine contexts in which reading is voluntary?

Past attempts to provide an answer focused on various means of reducing or eliminating the Stroop effect. Numerous studies have shown that the magnitude of the Stroop effect is affected by experimental manipulations such as the spatial separation of components (Kahneman & Chajzcyk, 1983), cueing (Logan & Zbrodoff, 1982), practice (MacLeod & Dunbar, 1988), colour-word discriminability (Algom, Dekel, & Pansky, 1996; Melara & Algom, 2003), the ratio of congruent and incongruent trials (Dishon-Berkovits & Algom, 2000; Lindsay & Jacoby, 1994; Schmidt & Besner, 2008) or single-letter colouring (Besner, Stolz & Boutilier, 1997; Manwell, Roberts & Besner, 2004). Besner et al. (1997) reasoned that, with the colour of a single letter in the word to be named, reading does not run to completion, so that lexical but not semantic analysis was occurring (hence the elimination of the Stroop effect). Because automatic action is ballistic and must run to completion, Besner and his colleagues concluded that the pertinent reading was not automatic. Based on other evidence, Algom and his colleagues (Algom et al., 1996; Pansky & Algom, 1999, 2002) have similarly concluded that word processing in the Stroop task was not strongly automatic. However, other investigators (e.g., Bargh, 1992; Henik & Tzelgov, 1982; Logan, 1988; Tzelgov, 1997), referring to various definitions of the concept of automaticity, consider the Stroop effect the primary example of automatic action.

Clearly, the automaticity debate vis-à-vis the Stroop effect is loaded and not fully settled -- it is moot if it can be resolved solely by experimental means, especially as the concept is so theory dependent. We decided to eschew discussion of the issue in these terms (but see the Tzelgov approach in the General Discussion) and to concentrate instead on the substantive reading processes under test. Experimentally, too, unlike much existing research,

we did not attempt to affect the presence or magnitude of the standard Stroop effect, but kept it intact. We approached the question of word processing by creating a benchmark task for comparison and diagnosis.

In order to gain insight on the nature of the pertinent reading, we introduced a new task, the *forced-reading* Stroop task. In this task, non-colour words as well as colour words are presented, and the participant's task is to indicate the ink colour *only* if the word is a colour word. Otherwise, she or he is to withhold the (colour) response. In the forced-reading task, the participant must engage the meaning of each and every word or she or he is unable to perform the task.

According to the traditional view (Klein, 1964; MacLeod, 1991; 1992), people access the meaning of the words in the typical Stroop task. Because this common assumption is rarely articulated (beyond mere statement of the presence of reading), we dub it as the standard or the default view. On this view, comparable amounts of the Stroop effect are expected in the classic and in the forced-reading tasks. If, on the other hand, reading and semantic processing is accomplished to a deeper level in the forced-reading task than in the standard task, a larger Stroop effect is expected to emerge in the forced-reading task than in the standard task. These predictions are depicted in a graphical form in Figure 1.

Our participants were presented with *precisely* the same stimuli in the standard and in the forced-reading Stroop tasks. The stimuli comprised colour words in colour and non-colour or neutral words in colour interspersed in a random fashion within a single block. Note that the inclusion of non-colour or neutral words in the list does not constitute a departure from accepted practice in the standard Stroop task. Such words are routinely included in order to allow the partition of the Stroop effect into interference and facilitation.[1] The crucial difference between the standard and the forced-reading tasks was this. In the standard task,

[1] We were not interested in the partition in this study, especially as it is precluded in the forced-reading task.

responding was not selective and the participants were asked to classify ink colour on every trial regardless of the meaning of the carrier word. In the forced-reading task, by contrast, responding was selective, contingent on word meaning. The participants were asked to classify ink colour only if the carrier word was a colour word. Does the Stroop effect derived with respect to the same colour-word stimuli differ across the two tasks?

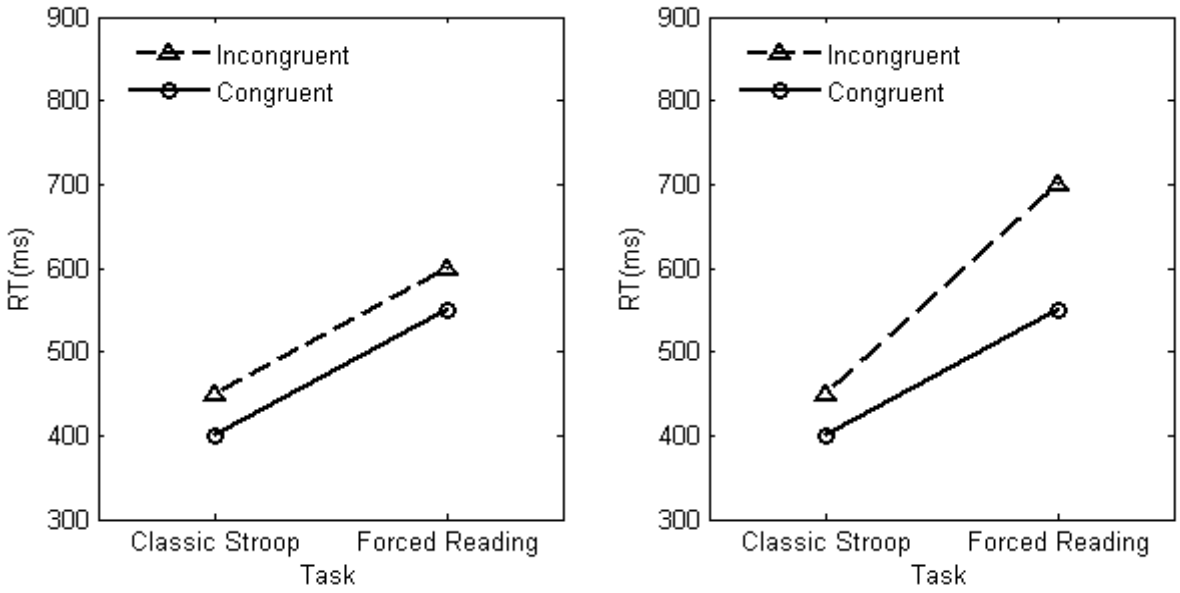

Figure 1: Mean RT patterns predicted by the default, reading-for-meaning hypothesis (left-hand panel) and by the shallow reading hypothesis (right-hand panel) in the classic Stroop task. The vertical separation between the curves gives the Stroop effect. On the former view, although the absolute RTs are longer in the forced-reading task, due to the increased difficulty of this task, the Stroop effects are comparable. On the latter view, responding is again longer under the forced-reading instructions, but now the Stroop effect is larger than that in the standard task. Statistically speaking, one expects an interaction of stimulus type (congruent, incongruent) and task (classic, forced) only under the shallower reading hypothesis.

Experiment 1

Method

Participants. Twenty young students from the University of Newcastle performed in Experiment 1. All were native English speakers with normal or corrected to normal vision.

Stimuli, Apparatus, and Design. The stimuli included the colour words RED and GREEN and their corresponding prototypical print colours (with RGB values of 220, 0, 0, and 0, 170, 0, respectively, for red and green). We selected the non-colour words such that each shared all but one letter with the corresponding colour word. For RED, the three non-colour words presented were ROD, BED, and RENT. For GREEN, they were QUEEN, GRAIN, and GREED. These non-colour words were selected to ensure that the participants read the entire word and that they could not respond based on local cues.[2] Words were written in uppercase Arial font, bold, size 30, which, at a viewing distance of 60 cm subtended a maximum width of less than 3 degrees of visual angle. Stimuli were presented on a 17" CRT monitor, using an IBM compatible computer and Presentation software, which also recorded response times (to the 1ms). Each trial entailed the following sequence. A fixation cross appeared in the centre for 500 ms, followed by a blank screen for another 500 ms at the end of which the stimulus (a word in colour) was presented for a maximum of 500 ms. The presentation was response terminated regardless of whether the response occurred within the interval of 500 ms or following a 1000 ms interval during which the screen remained blank. The next trial followed after 400 ms.

For each task, the participant performed in 840 trials. These were partitioned into seven blocks of 120 trials each. We introduced 2 min breaks between successive blocks. The makeup of the block was as follows. Each colour word (RED, GREEN) appeared printed in each ink colour (red, green) 15 times a combination, making for 60 colour word trials in all. Each of the 6 non-colour words appeared printed in each ink colour (red, green) 5 times a combination, making for 60 non-colour word trials in the same block. The order of presentation was random and different for each participant.

[2] We controlled for word frequency, as much as possible, by selecting for non-colour words the orthographic neighbours of the colour words with the closest frequency (based on N-Watch, Davis, 2005). For example, for the colour name RED (log frequency 2.29) we used the nearest-frequency non-colour neighbours ROD, BED, and RENT (log frequencies of 1.47, 2.41, and 1.62, respectively).

Procedure. Each participant performed in both the classic and the forced-reading Stroop tasks, with testing separated by at least 24 hours. The order of tasks was counterbalanced across participants such that a random half performed the classic task first, and the remaining half performed the forced-reading task first. On a trial, a single word in colour appeared at the centre of the screen. The background was grey. In the classic Stroop task, the participants classified, while timed, the colour of *all* the words presented. In the forced-reading task, the participants classified the colour of the colour words but not that of the non-colour words. For the latter they simply withheld the response. Responses were made by pressing the appropriate key on a Cedrus response pad (with responses counterbalanced across colour and side of responding). Both accuracy and response time were recorded.

The participants were tested individually in a dimly lit cubicle. Before performing in each task, the participants were given two short practice sessions of 20 trials each (the first with- and the second without-feedback).

Data Analysis. One participant exceeded the error-rate criterion of 15%, and her results are omitted from further analysis. Error rate for the remaining 19 participants was low (M=3.6%). Errors were fewer on congruent (2.2%) than on incongruent trials (4.9%) [$F(1,18)=18.52$, $p<.001$], and did not differ between the classic and forced tasks (3.8% and 3.3%, respectively) [$F(1,18)=.84$, ns]. These results rule out the possibility of RT-accuracy trade-off (e.g., Luce, 1986, pp. 81-90). Therefore, we discuss RT for correct responses in the following analyses.

Results and Discussion

The results are presented in Figure 2. Salient to visual inspection is the agreement of the pattern of responses with the predictions of the shallower Stroop reading hypothesis. The Stroop effect in the forced-reading condition (93 ms) was larger by an order of magnitude than that in the standard condition (7 ms) for the same set of colour-word stimuli. Statistical

analysis bears out the results of the visual examination. The main effect for stimulus type [$F(1,18)=111.25$, $p<.001$] confirmed the presence of the Stroop effect in the data. This was the case for the relatively small effect in the standard Stroop task [$t(18)=2.11$, $p<.05$] as it was for the much larger effect in the forced-reading task [$t(18)=11.69$, $p<.001$]. Most important, the interaction of stimulus type and task [$F(1,18)=129.36$, $p<.001$] confirmed the presence of a larger Stroop effect in the forced-reading task than in the standard task. This stimulus type x task interaction was also found for 18 out of the 19 participants, confirming that the pattern depicted in Figure 1 characterized the data of virtually all of the individual participants, too.

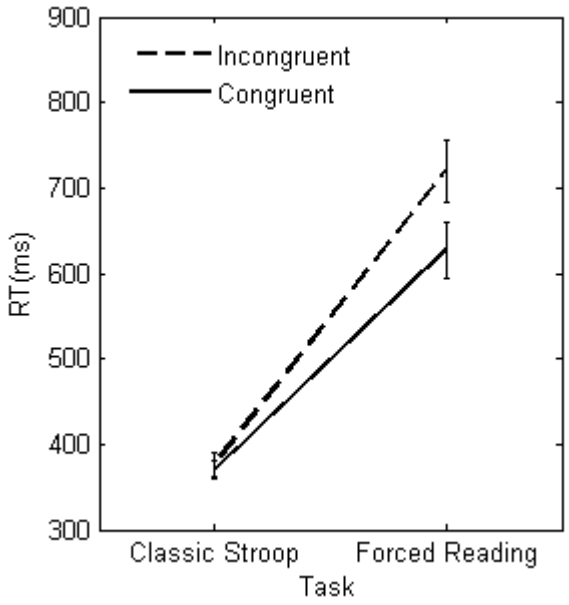

Figure 2. Results of Experiment 1: Mean RTs for correct identification of the ink colour of colour words as a function of stimulus type (congruent, incongruent) and task (classic- or forced-reading Stroop). Error bars represent one standard error around the mean.

One should be a bit circumspect though before drawing too strong a conclusion in favour of the shallower Stroop reading view. As we surmised, the forced-reading task proved more difficult than the standard task [means of 673 and 373 ms, respectively; $F(1,18)=90.31$, $p<.001$]. Given (a) the general tendency for longer RTs to produce larger effects (Faust, Balota, Speiler, and Ferraro, 1999), and (b) the specific tendency for the Stroop effect to

increase with absolute RT (Melara & Algom, 2003; Shalev & Algom, 2000), the inflated Stroop effect observed in our new task could have derived from the greater task difficulty. Cognizant of this possibility, we repeated the statistical analyses on (a) z-score transformations of the mean response times, and (b) logarithmic transformation of the individual response times.

For z-scores, we transformed the data by subtracting each participant's overall mean for a given Stroop task from the mean of each condition within this task, and divided the result by the standard deviation of that condition. This standardization effectively removes the influence of the main factor of task. Rather than eliminating the over-additive interaction, the transformed data actually provided even stronger evidence in favour of the critical interaction between stimulus type and task [$F(1,18)=151.91$, $p<.001$].

Following the logarithmic transformations, the average latencies in the two tasks were nearly equal (the mean RT in the standard task was 93% of the mean in the forced-reading task), yet the critical interaction [$F(1,18)=113.86$, $p<.001$] remained intact, again documenting the large asymmetry in the Stroop effects favouring the forced-reading task. Collectively, the two analyses suggest that the size of the Stroop effect in the latter task was larger beyond what might be expected on the basis of generic slowdown alone.

Nevertheless, given (a) the theoretical weight of our results along with (b) the rather small Stroop effects in the standard task (but see Eidels, Townsend, & Algom, 2010, and Eidels, 2012), we deemed a replication and extension warranted. In Experiment 2, we used the procedures of Experiment 1 with a single notable exception. We added a third colour word and its corresponding ink colour to the stimulus ensemble. We expected to find larger Stroop effects with this preparation.[3]

[3] With three colour words printed in three colours, there are twice as many incongruent trials as there are congruent trials. This creates a negative correlation between the colour and word dimensions, and the expected outcome is a larger Stroop effect.

Experiment 2

Method

Participants. A fresh group of ten participants from the same pool of University of Newcastle students performed in the classic Stroop task and the forced-reading Stroop task.

Stimuli, Design, and Procedure. The word stimuli included the words RED, GREEN and BLUE as well as their non-colour neighbours (ROD, BED, and RENT; QUEEN, GRAIN, and GREED; BASE, GLUE, and BLUR). These words appeared printed in red, green, or blue (with RGB values of 0, 0, and 240 for the latter; RGB values for red and green were the same as in Experiment 1).

Given the larger stimulus set, the makeup of the blocks changed. Each colour word (RED, GREEN, BLUE) appeared printed in each ink colour (red, green, blue) 6 times, making for 54 colour word trials. Each of the 9 non-colour words appeared in each print colour twice, making for 54 non-colour word trials. Mixing these combinations in a random fashion made for a block of 108 trials in all. An experimental session comprised the presentation of 8 such blocks of 108 trials each. For each task (classic, forced), there were 2 sessions, so that each participant performed in 4 separate sessions overall. In all other respects the procedures of Experiment 2 were the same as those of Experiment 1.

Data Analysis. Error rate for all participants was low (M=3.7%). As in Experiment 1, analysis confirmed the lack of an RT-accuracy trade-off: There was not a significant difference between the error rates for the classic (4.3%) and the forced (3.1%) tasks [$F(1,9)=1.75$, ns), or between congruent (2.9%) and incongruent (4.5%) trials [$F(1,9)=5.01$, ns]. Below, we report RT analyses for correct responses.

Results and Discussion

The results are presented in Figure 3. Again, the pattern of data is consistent with the predictions of the shallower reading hypothesis. The main effect for stimulus type

[F(1,9)=80.36, p<.001] confirmed the presence of the Stroop effect. There was an appreciable effect in the standard task [37ms, t(9)=3.86], which nonetheless quadrupled in the forced-reading task [122 ms, t(9)=9.65, p<.001]. Notably, the stimulus type x task interaction [F(1,9)=37.71, p<.001] once again documented the escalation of the Stroop effect under forced-reading. At the individual level, this critical interaction appeared in the data of 9 out of the 10 participants.

The more demanding forced-reading task took a toll on the speed of responding in Experiment 2, too. The mean RTs in the standard and the forced-reading tasks were 500 and 767 ms, respectively [F(1,9)=43.01, p<.001]. This difference shrank in the log-transformed data (with mean latency in the standard task amounting to 94% of that in the forced-reading task). Nevertheless, the stimulus type x task interaction remained intact with both the z-transform [F(1,9)=33.58, p<.001] and the log-transform [F(1,9)=18.02, p<.01] analyses. We conclude that the enhanced Stroop effect in the forced-reading task is a genuine one and it is not explained by task difficulty.

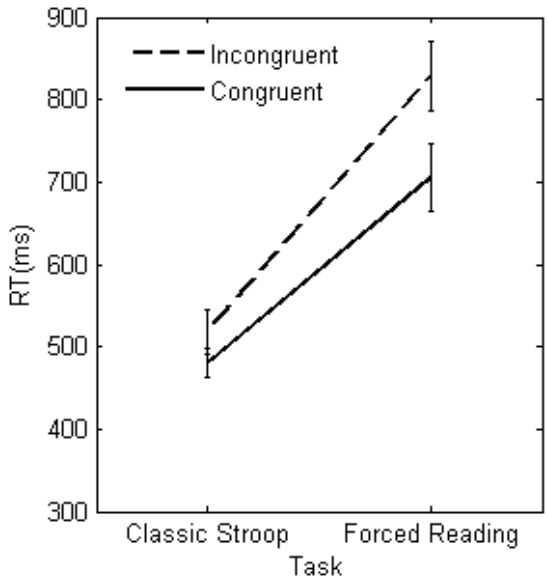

Figure 3. Results of Experiment 2: Mean RTs for correct identification of the ink colour of colour words as a function of stimulus type (congruent, incongruent) and task (classic- or forced-reading Stroop). Error bars represent one standard error around the mean.

The overall results, analyses, and conclusions of Experiment 2 granted, we wished to test a remaining, partly alternative explanation. The longer RTs in the forced reading task possibly allowed the words to be retained longer in working memory, thereby making them available for deeper semantic analysis. In order to rule out this account, in an auxiliary experiment we employed yet another novel task, the *forced typesetting detection* or *forced proofreading for letter-case* task. In this task, half of the presented words entailed one letter in italics. The participant's task was to respond to the ink colour only if the carrier word did not contain a letter in italics; otherwise the participant was to withhold the response. Note that in both the forced *reading* and the forced *proofreading* tasks, the response is contingent on the carrier word. However, the criterion for responding differs: it is word meaning in the former task, but word font in the latter task.

Given the contingent nature of the two forced tasks, the overall RTs were more closely (though not fully) matched, with forced reading taking 80 ms longer than forced proofreading [$F(1,27) = 29.15$, $p<.001$]. Notably, we rank-ordered the participants in terms of matched performance in the two tasks, and selected those exhibiting (nearly) equal performance [$F(1,11) = 2.60$, $p>.05$].[4] For this subset of 12 participants, a larger Stroop effect was still present in the forced reading task in the face of equal overall RTs [$F(1,11) = 6.52$, $p<.05$]. For a powerful analysis on the individual data, we regressed the difference in magnitude of the Stroop effects in the two tasks on the ratio of the respective overall RTs. Both Pearson (.023) and Spearman's (.096) correlations were negligible. The Bayes factor (cf. Rouder & Morey, 2012) was 0.147 meaning that the null model (lack of correlation) was approximately 7 times more likely than the alternative model (the relationship present). Clearly, any

[4] Averaged across these 12 participants, the mean RT in forced-proofreading amounted to 97% of the mean RT in forced-reading.

remaining difference in absolute RT between the tasks is an unlikely candidate to account for the larger Stroop effects observed in the forced reading task.

We conclude that the larger Stroop effect obtained in the forced-reading task in comparison with that obtained in the standard colour-naming Stroop task or in the forced-proofreading task is genuine. When people scan a script for font colour or for typesetting they do not engage in semantic analysis to the same extent that they do when they wilfully read the same words.

The last point granted, a lingering concern nonetheless remains. Considering the overall results of Experiments 1-2, we recorded relatively weak effects in the standard task. Although the Stroop effect of 37 ms obtained in Experiment 2 was larger than the miniscule one found in Experiment 1, it is still fairly small. In order to support further the validity of our results with sufficiently large standard effects, we decided to perform an experiment tailored on Experiment 1 -- with vocal responding. As a rule, vocal responses yield larger Stroop effects than do manual ones (MacLeod, 1991). Vocal responding is also invited in view of the current issue at question: the nature of word processing in the classic Stroop task. Would the magnitude of the Stroop effect be larger in the forced reading condition than in the standard condition with oral responding, too?

Experiment 3

In order to reinforce our results and conclusions, we repeated the procedures of Experiment 1 using vocal rather than manual responses. Participants said aloud the ink colour of all words presented in one session (classic task), but said aloud the ink colour only when the carrier word was a colour word in another session (forced reading task). We expected to find larger overall Stroop effects with the vocal responses. On the background of large effects, we still expected that under forced reading to be the larger one.

Method

Participants Twenty young students from the University of Newcastle performed in the classic- and forced-reading Stroop tasks. Each participant performed in two sessions (classic and forced task) separated by at least 24 hours. None of the participants performed in Experiments 1 or 2, and all were native English speakers with normal or corrected to normal vision.

Stimuli, Apparatus, and Design. The stimuli were drawn from Experiment 1. The words RED, GREEN, with their orthographic neighbours (ROD, BED, and RENT; QUEEN, GRAIN, and GREED), appeared printed in either red or green. As in Experiments 1 and 2, a single word in colour was presented on each trial.

Stimulus presentation and recording of the vocal responses were controlled using the ChoiceKey program (Donkin, Brown, & Heathcote, 2009). ChoiceKey is a speech recognition software for Matlab, designed to obtain choice and response times in experiments with a small response set. It requires a short training and testing block at the beginning of each session, which we describe below. The design was otherwise similar to Experiment 1. There were seven experimental blocks of 120 of trials each, making for a total of 840 trials in each task.

Procedure. Each participant performed in both the classic and the forced-reading Stroop tasks with testing separated by at least a day. The order of tasks was counterbalanced across participants. Each session started with a ChoiceKey training and testing block (of 120 trials), followed by a 5-min break, and followed subsequently by the experimental session.

In the ChoiceKey training block, each participant said aloud the words, “red” and “green”, 40 times each. The 80 responses were used in ChoiceKey as a set of exemplars with which the experimental responses of the participant were compared. After the 80 training exemplars were recorded, the identification accuracy of ChoiceKey was tested; The

participant read aloud another set of 40 presentations of the same words, “red” and “green”, with the experimenter recording errors. The training process repeated itself until the accuracy of identification reached 95%. The experimental session started at this point.

In the classic Stroop task of Experiment 3, the participants *said aloud*, while timed, the colour of all the words presented. In the forced-reading task of Experiment 3, the participants *said aloud* the colour of the colour words, but not the colour of the non-colour words.

Data Analysis. One participant exceeded the error-rate criterion of 20%, and his results are omitted from further analysis. Error rate for the remaining 19 participants was 11.7%. This rate likely represents accumulation of errors from incorrect colour reports by the observers as well as mis-identification of the vocal responses by ChoiceKey (especially towards the latter blocks, when participants became fatigued and failed to pronounce the colour names properly). No RT-accuracy trade-off was observed, $F(1,18)=2.35$, ns. The following analyses report RTs for correct responses.

Results and Discussion

There was a main effect of stimulus type, $F(1,18)=78.80$, $p<.01$, confirming the presence of an overall Stroop effect, with responses on congruent trials faster than incongruent trials (696 vs 781ms, respectively). There was also a main effect of task, $F(1,18)=92.68$, $p<.01$, with faster responses in the classic task (mean of 621ms) than in the forced task (mean of 857 ms). Most important, there was a task x stimulus-type interaction, $F(1,18)=8.96$, $p<.01$, confirming the presence of a task-dependent difference in the magnitude of the Stroop effect. The Stroop effect in the classic task amounted to 65 ms [$t(18)=5.27$, $p<.001$], whereas the Stroop effect in the forced task was 105 ms [$t(18)=9.73$, $p<.001$]. Additional testing further confirmed that the effect in the forced reading task was larger than that in the classic task [$t(18)=2.99$, $p<.01$]. Predictably, the vocal responding employed in

Experiment 3 led to large effects in the standard Stroop task (larger than the values obtained by manual responding). Nevertheless, the pattern of larger effects in the forced-reading task was replicated by this mode of responding, too.

Notably, both main effects survived log transformation of the data, but the task x stimulus-type interaction did not. Upon scrutiny, we found that this outcome was driven by a single participant (#8), who exhibited a staggering effect in the classic task (256 ms, three times the value of the second largest effect recorded). Reanalysis of the data after discarding the responses of this individual revealed a significant task x stimulus-type interaction, $F(1,17)=5.33$, $p<.05$. Consequently, we conclude that reading in the forced task was accomplished at another level than that in the classic task. We reached this conclusion earlier, but now it rests on naming aloud, too.

General Discussion

In this study, we introduced a new kind of Stroop task, the forced-reading task. Imposing this task, one can be confident that all of the words presented are read to a fair level of semantic analysis (because the responses are critically dependent on word meaning). If the same words are similarly processed in the standard Stroop task -- a common yet implicit assumption --- the Stroop effects should be comparable. They were not. The effect in the standard task was but a fraction of its value in the forced-reading task. We conclude that the nature of reading in the Stroop task is not comparable to that occurring during voluntary reading or in situations in which one is compelled to read. The level of word reading accomplished in the (standard) Stroop task suffices to generate interference to the target task of colour naming, yet the process is not as deep as it typically is in voluntary, fully intentional reading. When the latter is introduced into the Stroop environment, the resulting interference is appreciably greater than that observed in the standard task.

Following Tzelgov (1997; Ganor-Stern, Tzelgov, & Ellenbogen, 2007), one can refer to automatic processes as those that occur without intention, including cases in which the specific (automatic) process is not part of the behavioural goal at hand and that can even hurt the explicit goal for action. Espousing this definition of automaticity, reading in the standard Stroop task is automatic. By virtue of the same definition, reading in the forced task is not automatic because it is done intentionally under conscious control. Various theoretical approaches (e.g., Dienes & Perner, 1999; Dulany, 1991; 1996) converge on the conclusion that automatic processing results in a different, shallower representation. Our results provide empirical support for this distinction. At the same time, the existence of a difference in interference between the two tasks argues against strong automaticity in the standard Stroop task. If this were the case, comparable performance would have been observed. Be this as it may, there is a deep difference in the nature of word processing in and out of the classic Stroop task.

Considering word processing, our results are consistent with the dual route cascaded (DRC) model of word recognition (Coltheart et al., 2001) as well as with Roelof's dedicated Stroop model (2003; notice, in particular, the notion of goal-referenced selection). In the DRC model, for example, words presented for view can be processed along lexical and non-lexical pathways, and, within the former, along semantic or non-semantic routes. Semantic information is not strictly required for word recognition. This level, possibly augmented by a modicum of semantic information, probably sufficies to produce the Stroop effect. Again however, this level is not typical of routine reading.

Commensurate with the present perspective, Coltheart, Woollams, Kinoshita, and Perry (1999) reported a "phonological Stroop effect" (see again, Roelofs, 2003) that can occur without sematic processing of the carrier word. They found that colour naming responses were faster when the printed word shared phonemes with the colour name (e.g., naming the

red colour of the word ‘ROOK’) than when it did not (naming the red colour of the word ‘SOOT’). This occurred in the absence of a semantic relationship between the printed words and the colour names. The findings suggest that phonology can account for at least part of the Stroop effect. Although 35 years old, the study by Regan (1978) is relevant to the present concerns. She found that coloured colour-name initials sufficed to produce the Stroop effect, but that this effect was smaller than that obtained with fully scripted colour names. Clearly, a portion only of normal reading processes is accomplished with a single letter, but even this portion suffices to generate interference (or facilitation) to the target task of colour naming (Regan reported both interference and facilitation; see also Kahneman & Chajzcyk, 1983). Of the processes subsumed under the generic term, reading, a relatively limited subset operates under the conditions of the standard Stroop task in which reading is involuntary (discouraged by the task demands). That level of reading is not negligible, yet it is not on a par with that occurring under normal voluntary reading. This difference is demonstrated in our study by employing the standard preparation used in the vast majority of Stroop studies in the literature (i.e., without modifying the stimulus in any systematic way).

In conclusion, let us consider a new, radical explanation for the present results. It is favoured by applying Occam’s razor, but we offer it at this point as a challenge to accepted modes of thinking vis-à-vis the Stroop phenomenon. Suppose that a given word is read to the *same depth* under the standard and the forced conditions. The difference between the two conditions -- a larger Strop effect under forced-reading -- derives from a difference in the *number* of the items processed in the two conditions. Obviously, all of the items are read in the forced-reading condition. However, it is possible that only a subset is read under the standard condition. Our explanation rests on the fact that virtually all published Stroop studies are based on the *mean* RTs. This means that it is not necessary for each and every word to be

read in order to produce a Stroop effect. It is entirely possible that the observed (and reported) effect is actually based on a subset only of the experimental trials.

The implications for practice are profound. They might undermine the possibility of a unique interpretation for any given Stroop result. Suppose that two people undergo Stroop testing (say, for early diagnosis of Attention Deficit Hyperactivity Disorder, ADHD). Suppose further that they get the same result, so that both exhibit a Stroop effect of 60 ms. Does the common result tap a similar degree of selective attention? Not necessarily! One observer might have read almost all of the words presented, whereas the other only a much smaller portion.

A simple probability-mixture model captures this idea and can account for the present results. Under this model, words are read, or processed to a level that creates interference, on some proportion of the trials. When read, incongruent words may slow down, and congruent words may speed up, the naming of colour. The overall empirical distribution is a binary mixture of two unobserved distributions (Brown, Lehman, & Poboka, 2006; Townsend & Ashby, 1983, p. 263).[5] The observed RT on a given trial is a sample drawn from one of these distributions, either the distribution associated with reading (with probability *p*), or the distribution free of word reading (with probability 1-*p*). By forcing people to read, we have increased the probability of reading to maximum (*p*=1). This should lead to an inflated Stroop effect compared with the standard task, and this is what our data show.

[5] Formally, if $f_R(t)$ is the (unobserved) distribution of response times from trials where the word was *read*, and $f_{NR}(t)$ is the (unobserved) RT distribution from trials where the word was *not read*, with corresponding probabilities of *p* (reading) and 1-*p* (not reading), then the *observed* RT distribution is $g(t) = pf_R(t) + (1-p)f_{NR}(t)$.

References

Algom, D., Dekel, A., & Pansky, A. (1996). The perception of number from the separability of the stimulus: The Stroop effect revisited. *Memory & Cognition*, *24*, 557-572.

Bargh, J. A. (1992). The ecology of automaticity: Towards establishing the conditions needed to produce automatic processing effect. *American Journal of Psychology, 105,* 181-199.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462-479.

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

Brown, T. L., Gore C. L., & Carr, T.H. (2002). Visual attention and word recognition in Stroop colour naming: Is word recognition "automatic"? *Journal of Experimental Psychology, 131,* 220-240.

Brown, S. D., Lehmann, C., & Poboka, D. (2006). A critical test of the failure-to-engage theory of task-switching. *Psychonomic Bulletin & Review, 13,* 152-159.

Bugg, J. M., & Hutchison, K. A. (2013). Converging evidence for control of color-word Stroop interference at the item-level. *Journal of Experimental Psychology: Human Perception & Performance*, in press.

Coltheart, M., Rastle, K., Perry, C., Langdon, R. & Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204-256.

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.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal behavior, 11*, 671-684

Davis, C. J. (2005). N-Watch: A program for deriving neighbourhood size and other psycholinguistic statistics. *Behaviour Research Methods, 37,* 65-70.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and non-aphasic speakers. *Psychological Review, 104,* 801-838.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and brain sciences, 22,* 735-808.

Dishon-Berkovits, M., & Algom, D. (2000). The Stroop effect: It is not the robust phenomenon that you have thought it to be. *Memory & Cognition*, *28*, 1437-1449.

Donkin, C.**,** Brown, & Heathcote, A. (2009). ChoiceKey: A real-time speech recognition program for psychology experiments with a small response set. *Behavioral Research Methods, 41,* 154-162.

Dulany, D. (1991). Conscious representation and thought systems. In R. S. Wyer, & T. K. Srull (Eds.), *Advances in social cognition* (pp. 91-120). Hillsdale, NJ: Erlbaum.

Dulany, D. (1996). Consciousness in the explicit (deliberate) and the implicit (evocative). In J. Cohen, & J. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 179-212). Hillsdale, NJ: Erlbaum.

Eidels, A., Townsend, J. T., & Algom, D. (2010). Comparing perception of Stroop stimuli in focused versus divided attention paradigms: Evidence for dramatic processing differences *Cognition, 114*, 129-150.

Eidels, A. (2012). Independent race of colour and word can predict the Stroop effect. *Australian Journal of Psychology*, *64*, 189-198.

Evans, G. A. L., Lambton Ralph, M. A., & Woollams, A. M. (2012) What's in a word? A parametric study of semantic influences on visual word recognition. *Psychonomic Bulletin & Review*, *19*, 325-331.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125*, 777-799.

Ganor-Stern, D., Tzelgov, J., & Ellenbogen, R. (2007). Automaticity and two-digit numbers. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 483-496.

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition*, 10, 389-393.

Kahneman, D., & Chajzcyk, D. (1983). Tests of the automaticity of reading: Dilution of Stroop effects by colour-irrelevant stimuli. *Journal of Experimental Psychology: Human Perception and Perception and Performance, 9*, 497-509.

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

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 219-234.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological review, 95,* 492-527.

Logan, G. D., & Zbrodoff, N. J. (1982). Constraints on strategy construction in a speeded discrimination task. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 502-520.

Luce, R.D. (1986) *Response Times: Their Role in inferring elementary mental organisation.* New York: Oxford University Press.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 102,* 163-203.

MacLeod, C. M. (1992). The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General, 121,* 12-14.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 126-135.

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

Melara, R. D. & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review, 110,* 422-471.

Pansky, A., & Algom, D. (1999). Stroop and Garner effects in comparative judgment of numerals: the role of attention. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 39-58.

Pansky, A., & Algom, D. (2002). Comparative judgment of numerosity and numerical magnitude: attention preempts automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 259-274.

Regan, J. E. (1978). Involuntary automatic processing in color-naming tasks. *Perception & Psychophysics, 24,* 130-136.

Roediger, H. L., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels of processing framework. *Memory, 10,* 319-332

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

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavior Research, 47,* 877-903.

Schmidt, J. R., & Besner, D. (2008). The Stroop effect: why proportion congruent has nothing to do with congruency and everything to do with contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 514-523.

Shalev, L., & Algom, D. (2000). Stroop and Garner effects in and out of Posner's beam: Reconciling two conceptions of selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 997-1017.

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

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* Cambridge, UK: Cambridge University Press.

Tzelgov, J. (1997). Specifying the relations between automaticity and consciousness: A theoretical note. *Consciousness and Cognition, 6,* 441–451.

Author note

We thank Kate Berka, Wendy Devine, and Frank van de Mortel, for their invaluable assistance. We also thank Joseph Tzelgov for helpful comments on an earlier version of this manuscript. The study was supported by an Australian Research Council Discovery ARC-DP 120102907 grant and by the University of Newcastle's New Staff grant to A. E. Please address correspondence to Ami.Eidels@newcastle.edu.au.