# Assessing stimulus–stimulus (semantic) conflict in the Stroop task using saccadic two-to-one color response mapping and preresponse pupillary measures

**Nabil Hasshim**[1] · **Benjamin A. Parris**[1]

Published online: 14 August 2015

**Abstract** Conflict in the Stroop task is thought to come from various stages of processing, including semantics. Two-to-one response mappings, in which two response-set colors share a common response location, have been used to isolate stimulus–stimulus (semantic) from stimulus–response conflict in the Stroop task. However, the use of congruent trials as a baseline means that the measured effects could be exaggerated by facilitation, and recent research using neutral, non-color-word trials as a baseline has supported this notion. In the present study, we sought to provide evidence for stimulus–stimulus conflict using an oculomotor Stroop task and an early, preresponse pupillometric measure of effort. The results provided strong (Bayesian) evidence for no statistical difference between two-to-one response-mapping trials and neutral trials in both saccadic response latencies and preresponse pupillometric measures, supporting the notion that the difference between same-response and congruent trials indexes facilitation in congruent trials, and not stimulus–stimulus conflict, thus providing evidence against the presence of semantic conflict in the Stroop task. We also demonstrated the utility of preresponse pupillometry in measuring Stroop interference, supporting the idea that pupillary effects are not simply a residue of making a response.

**Keywords** Stroop · Semantic conflict · Same response · Pupillometry · Oculomotor

The *Stroop effect* refers to the finding that people are slower to name the color that a word is printed in when the word spells out another color (*incongruent trials*—e.g., the word *red* in blue) than to name the color of a square (Stroop, 1935) or to name a word's color when the word spells out the same color (*congruent trials*—e.g., the word *red* in red; Klein, 1964; see MacLeod, 1991, for a review). The Stroop task has been described as the gold standard for measuring attention (MacLeod, 1992) and has been the focus of influential models of attention (e.g., Cohen, Dunbar, & McClelland 1990; Glaser & Glaser, 1989; Roelofs, 2003).

The Stroop effect has been attributed to having to resolve conflict at the response stage when the color and the meaning of the word each activate different responses (referred to as *response conflict* or *stimulus–response conflict*; Cohen et al., 1990; MacLeod, 1991; Roelofs, 2003). However, some researchers have posited that, in addition to interference/conflict resolution at the response stage, performance in the Stroop task also requires conflict resolution at earlier processing stages (e.g., De Houwer, 2003b; Goldfarb & Henik, 2007; Hock & Egeth, 1970; Klein, 1964; Parris, 2014; Schmidt & Cheesman, 2005; Sharma & McKenna, 1998; H. Zhang & Kornblum, 1998; H. H. Zhang, Zhang, & Kornblum, 1999).

One such stage is semantic processing. This is controversial, however, since key models of the Stroop task account for interference in terms of response-level conflict only (Cohen et al., 1990; Roelofs, 2003). To establish whether semantic conflict is present in the Stroop task, researchers have tended to use semantic–associative Stroop stimuli (e.g., *sky* in red, where *sky* is associated with blue). Numerous studies have shown evidence of a small but consistent semantic–associative Stroop effect (Augustinova & Ferrand, 2012; Klein, 1964; Schmidt & Cheesman, 2005; Sharma & McKenna, 1998). However, the use of such stimuli is problematic, since it is

✉ Nabil Hasshim
nhasshim@bournemouth.ac.uk

[1] Department of Psychology, Faculty of Science and Technology, Bournemouth University, Fern Barrow, Poole BH12 5BB, UK

not clear whether semantic interference is the only effect that slows down semantic–associative trials. For example, in his model of the Stroop task, Roelofs (2003) might account for semantic–associative Stroop effects as resulting from conceptual (semantic) level connections between the semantic–associative stimuli and the response set colors (*sky* is associated with blue, which is a member of the response set). However, the interference would only arise as a result of interactions in the language production (response) architecture. Thus, one might interpret the semantic–associative Stroop effect as being due to response-level, and not to semantic-level, conflict (see also Klein, 1964, for a similar argument). Even if this were an inaccurate representation of Roelofs's model, there is an unavoidable logical conundrum with the use of such stimuli, in that as long as response-level conflict is present, one can never be sure whether the conflict is occurring at the semantic-processing stage or at the response-level stage as a consequence of semantic-level connections to response set colors. Thus, to establish semantic-processing effects, one would need to present a Stroop stimulus that did not involve response conflict.

One such stimulus derives from the dimension overlap (DO) models (see H. H. Zhang et al., 1999, for an in-depth review of the taxonomy of DO models). DO models attribute interference effects in perceptual interference tasks, including the Stroop task, to overlap in the stimulus and response dimensions. This overlap can occur at a semantic level, between the dimensions of the stimulus (known as *stimulus–stimulus* or *S–S* overlap; Kornblum & Lee, 1995), or at a response level, between the stimulus and response (S–R) dimensions (Kornblum, Hasbroucq, & Osman, 1990). S–S overlap refers to similarity (defined as having the same characteristics) between the two stimulus dimensions (in the case of the Stroop task, the two stimulus dimensions, word and color, overlap because they both refer to the category of colors), whereas S–R overlap refers to how relevant a stimulus dimension is to a response dimension. When two dimensions overlap, the resulting effect depends on the compatibility (how much they match) of the stimulus dimensions (De Houwer, 2003a; Kornblum et al., 1990). On a congruent Stroop trial, both the S–S (the word and the color) and S–R (the word and the correct color response patch) dimensions are compatible, whereas on an incongruent trial, both S–S and S–R are incompatible. Congruent trials are typically responded to faster than incongruent trials, which could be due to the effects of compatibility at either or both the S–S and S–R levels.

To dissociate the effects of S–S and S–R compatibility, De Houwer (2003b) introduced a variant of the Stroop paradigm in which each response button maps onto two different colors (e.g., red and blue are assigned one button, whereas green and yellow are assigned another button). This two-to-one response-mapping paradigm allows for a new type of trial (*same-response trials*), in which the stimulus dimensions are of different colors, yet both colors are mapped to the same response (e.g., the word *red* in blue font, and both the "red" and "blue" responses are mapped to the "x" key). This means that, on same-response trials, the S–S relationship is incompatible, whereas the S–R relationship is compatible, allowing for the individual effects of S–S and S–R compatibility to be inferred by comparing the performance on same-response trials to that on congruent and incongruent trials, respectively.

Studies that have isolated S–S effects (De Houwer, 2003b; Schmidt & Cheesman, 2005; Zhang & Kornblum, 1998) have reported that S–S incompatibility independently contributes to the Stroop interference effect. These studies compared same-response trials (S–S incompatible, S–R compatible) to incongruent (S–S incompatible, S–R incompatible) and congruent (S–S compatible, S–R compatible) trials. Faster and slower responses to same-response than to incongruent and congruent trials, respectively, have commonly been observed. The difference between congruent and same-response trials was interpreted as evidence for S–S incompatibility or semantic conflict. The difference between incongruent and same-response trials was interpreted as evidence for a distinction between response and semantic conflict and established the two-to-one mapping approach as key to the argument that semantic-level conflict contributes to Stroop interference (Schmidt & Cheesman, 2005).

Although, at first blush, interpreting the difference between same-response and congruent trials as a form of conflict seems a reasonable interpretation, given the Kornblum et al. (1990) taxonomy, same-response trials might involve response facilitation, since both dimensions of the stimulus provide evidence toward the same response (as was indicated by De Houwer 2003b). A related point is the appropriateness of using congruent trials as a baseline for the measurement of interference, since they involve facilitation effects (T. L. Brown, 2011). This means that any measurement of interference using them as a baseline is potentially exaggerated by facilitation effects, which consequently indicates the need for a more appropriate baseline.

Typical baseline conditions used in Stroop paradigms have been nonword letter strings (e.g., *xxxx*) and neutral (non-color-related) words. T. L. Brown (2011) argued that these two conditions generally show different RTs, with the slower responses to neutral trials being attributed to a lexicality cost. Any baseline against which to compare same-response trial would therefore have to include a lexical component. Laeng, Ørbo, Holmlund, and Miozzo (2011) emphasized the same point in recommending neutral words over nonwords as baselines for pupillometry studies, because measurements that involve comparing them to color word trials would potentially include differences in lexical information in addition to semantic processing.

Despite this, subsequent studies using the two-to-one mapping paradigm have interpreted the difference between same-response and congruent trials as evidence of semantic conflict (e.g., A. Chen, Bailey, Tiernan, & West, 2011; van Veen & Carter, 2005). To investigate whether this measurement of semantic conflict is affected by facilitation to either congruent or same-response trials, Hasshim and Parris (2014) compared performance on same-response and non-color-word neutral trials (e.g., *wall* in blue) in two experiments. If same-response trials produced slower responses than non-color-word neutral trials, it would be evidence of semantic interference; alternatively, if same-response trials produced faster responses than non-color-word neutral trials, it would be evidence of response facilitation. In fact, the difference in the RTs was shown to be statistically nonsignificant in both experiments, and Bayes factors provided evidence for no difference between the two trial types. It was suggested that this finding could be interpreted as either (1) being due to two different processes (semantic interference and response facilitation) working in opposite directions, resulting in a negligible net effect, or (2) evidence for no effect of S–S incompatibility/semantic conflict in the Stroop task. This latter possibility is important to consider, because not only is it contrary to studies that have attributed same-response trial performance to semantic input effects (De Houwer, 2003b; Schmidt & Cheesman, 2005), but the two-to-one response-mapping paradigm has been employed in recent studies putatively evidencing a dissociation between response and semantic conflict (Berggren & Derakshan, 2014; A. Chen et al., 2011; Z. Chen, Lei, Ding, Li, & Chen, 2013; Steinhauser & Hübner, 2009; Wendt, Heldmann, Münte, & Kluwe, 2007). Researchers have utilized congruent trials as a baseline to measure response conflict and have successfully differentiated response- and semantic-based conflict using distribution analysis (A. Chen et al., 2011; Steinhauser & Hübner, 2009). Furthermore, researchers have claimed to show that S–S and S–R forms of incompatibility activate different brain regions using neuroimaging (A. Chen et al., 2011; van Veen & Carter, 2005).

Although Hasshim and Parris (2014) did find evidence for no difference between nonresponse and neutral trials in their first experiment, the Bayes factor for the second experiment was only 0.58, which suggests that the null results in that experiment might have been due to the data being too insensitive to detect the effect (Dienes, 2014). In the present study, we investigated whether S–S incompatibility/semantic interference effects during the Stroop task could be revealed using a new, more sensitive measure of performance and an online measure of effort expenditure.

## Oculomotor measures of performance

As Logan and Irwin (2000) noted, eye movements are controlled by anatomical pathways that are separate from those that control hand movements, which might suggest that eye movement responses can reveal effects that are not present with manual responses. Moreover, they have noted that eye movements often precede hand movements, suggesting that mechanisms in operation early in processing might dissipate before hand movements are made. Sullivan and Edelman (2009) have noted that the link between attention and saccade programming is greater than the link between attention and manual motor programming.

Saccadic responses have recently been employed as an alternative to manual or vocal responses as a means to reliably measure Stroop interference. Hodgson, Parris, Gregory, and Jarvis (2009) utilized a saccadic Stroop task, in which participants responded to stimuli by moving their gaze to a different location on a screen instead of by pressing a button. They found that the latencies of the saccades showed Stroop effects, with the saccades for incongruent trials being initiated more slowly than those for congruent trials. Taken together, this work suggests that the oculomotor Stroop task might provide an alternative measure of potential differences between the conditions. Moreover, the use of eyetracking also permits the measurement of pupil dilation.

### Pupillometry as a measure of effort

Eyetracking not only permits the measurement of response latencies, but also provides a measure of changes in pupil size. Pupillometry, the measurement of change in the size of the pupil, has been used as a measure of effort in psychology (Laeng, Sirois, & Gredebäck, 2012; see Loewenfeld, 1993, for a review), with the pupil becoming larger as more cognitive effort is exerted. Evidence for this has been shown in larger pupil sizes being measured when the experimental stimuli presented were more intense (Stelmack & Siddle, 1982) and with increased memory load (Beatty, 1982; Granholm, Asarnow, Sarkin, & Dykes, 1996; Kahneman, 1973; Kahneman & Beatty, 1966). In the context of the Stroop task, it has been shown that the diameter of the pupil is largest during incongruent trials, relative to both neutral (Laeng et al., 2011) and non-word neutral (G. G. Brown et al., 1999) trials, which in turn elicit larger pupil diameters than congruent trials (Siegle, Steinhauer, & Thase, 2004). This means that change in pupil diameter is a robust measure of Stroop effects and can be used in conjunction with other measures, such as saccadic latencies, to differentiate between trials in different conditions (Laeng et al., 2012). Moreover, pupil measurement imposes no additional task requirements on the process being studied, since changes in pupil dilation are involuntary.

Importantly for the present purposes, research has shown that pupil dilation and response times (RTs) do not necessarily track each other. Porter, Troscianko, and Gilchrist (2007) showed that effort registered using pupil dilation can index difficulty during a visual search task when RTs do not.

Similarly, Chiew and Braver (2013) showed that transient pupillary effects indexing reward incentives are present even when RT performance is matched. Conversely, van der Meer et al. (2010) used pupillometry to show that individuals with higher fluid intelligence respond faster during low-level cognitive tasks while expending amounts of effort equal to those of individuals with lower fluid intelligence. Taken together, this research shows that it is possible that the factors that affect RTs may not be the same as those that influence pupil dilation, and as such, pupil dilation might reveal influences on performance that RTs do not. Here we investigated whether pupillometry can dissociate between same-response trials and neutral trials, on the assumption that same-response trials involve either opposing influences of semantic conflict and response facilitation, or just semantic conflict. One would assume that resolving opposing influences or S–S incompatibility would require effort, and that pupillometry might provide a method sensitive enough to detect this.

### Preresponse measures of pupil size

Typically, pupillometric measures are taken by averaging pupil size within an entire block of trials (e.g., G. G. Brown et al., 1999), which means that each block can only contain one experimental condition. Laeng et al. (2011) and Siegle et al. (2004) addressed this when they investigated the time course of pupillometric change within each trial by measuring the size of the pupil every 20 ms in each trial, up to 2,000 ms after stimulus onset. Their results showed that generally, the size of the pupil increases after the presentation of the stimulus, initially peaking about 400 ms after onset before decreasing again back to baseline levels. This is followed by a larger dilation that peaks about 1,400 ms after response. The second peak is where the biggest difference in pupil sizes across the different condition occurs, with the largest pupil diameters occurring after the presentation of incongruent trials. Laeng et al. (2011) indicated that an issue with using a post-behavioral-response measure is the possibility that it may simply indicate residual change due to the response that was made (Simpson, 1969). Although Laeng et al. argued that the differing patterns induced by the different conditions suggested that the second peak was not simply a reflection of the behavioral response, they highlighted the need for further research into pupillometry as a measure of cognitive processes, especially since it is a delayed measure, with the dilation occurring after a behavioral response has been made. This is of primary importance in the present study, since it is important that methods be adopted that increase the likelihood of the pupillometric measure not simply being a residual change due to the response that was made.

Pre-behavioral-response measures of changes in pupil diameter have generally not been used, because the initial peak that occurs within this timeframe is not significantly different across the different conditions (e.g., Laeng et al., 2011). However it should be noted that the time-course measurement of pupil size across the trials does show differences in the dip just before a behavioral response is given. There are differences in the minimum sizes of the pupil and in when the minimum sizes occur when different conditions are presented. Hence, it would be a worthwhile endeavor to investigate whether pupillometric data taken before a response can be used as a measure of Stroop interference. If Stroop interference can be reliably measured with preresponse pupillary data, this can be considered a simpler alternative to postresponse pupil size, and this is also useful when the task design does not allow for the long response–stimulus interval that the measurement of the postresponse peak requires.

In sum, in the present study we investigated whether S–S incompatibility effects during the Stroop task, as measured by the difference between same-response trials and non-color-related neutral word trials, would be revealed using an oculomotor version of the Stroop task—a new, more sensitive measure of performance—as well as via pupillometry—a well-established measure of effort expenditure in cognitive tasks. With the latter index, we employed a preresponse measure of pupil size to reduce the influence of the response on pupil size.

## Method

### Participants

Thirty-three students (25 female, eight male) from Bournemouth University participated in the study in exchange for course credit or £5. The average age was 22.15 years (*SD* = 4.61). The data from five other participants were excluded from the analyses because accurate calibration could not be maintained during the session and they were unable to complete all of the experimental trials.

### Apparatus and materials

Stimuli were presented using a standard PC running Experiment Builder software (SR Research Ltd) and displayed on a color monitor displaying at 120 Hz. The movement of only one eye was recorded using an EyeLink 1000 (SR Research Ltd.), recording both pupil and corneal reflection and sampling at 500 Hz (every 2 ms). Participants went through a nine-point calibration and validation before the start of each block. Eye movement and pupillometric parameters were extracted offline using Data Viewer (SR Research Ltd.).

During the task, participants placed their head and chin on a headrest positioned 60 cm from the screen. The stimuli were presented in the center of the screen in one of four colors: blue (RGB: 0, 125, 255), green (RGB: 0, 255, 0), red (RGB: 255, 0, 0), and yellow (RGB: 255, 255, 0). Two white squares 200 ×

200 pixels in size appeared in the top left and right corners of the screen, and participants made saccadic responses to one of the squares. Each square corresponded to a pair of colors (e.g., "if the color of the word is either blue or red, look at the square on the left; if it is either green or yellow, look at the square on the right"). There were four trial conditions: congruent, neutral, same-response, and different-response trials. On *congruent* trials, the word spelled out the corresponding color that it was presented in, whereas on neutral trials, the word was a non-color-related word. On same-response trials, the word spelled out an incongruent color, which shared the same response location as the relevant color dimension, whereas in different-response trials, the incongruent color word always referred to a color whose response location was on the opposite side from that of the correct response. The neutral words *wall*, *due*, *story*, and *marvel* were used in the neutral trials and were matched for frequency and length to the color words. The words were presented in lowercase, bold, and in size-20 Courier New font on a black background.

### Procedure

At the beginning of each trial a fixation cross appeared in the center of the screen, and as soon as it was fixated on, it was replaced by the Stroop stimulus, and the two response squares appeared at the top corners of the screen. Participants were asked to move their gaze toward the square that corresponded to the correct response of the stimulus, and to do so as quickly and accurately as possible. Once a fixation of 100 ms had been made in the area of the correct square (up to 100 pixels around the square), the stimulus and squares were replaced with the fixation cross for the next trial.

At the start of each session, participants went through a practice block of 48 trials made up of strings of hash symbols (#) from three to six characters in length. Color patches corresponding to the colors assigned to the response squares were placed above the white squares, to aid participants in remembering the response locations, and subsequently were removed during the experimental trials. This was followed by 240 experimental trials, consisting of 48 trials each of the congruent, neutral, and same-response conditions and 96 trials of different-response trials, broken down into three blocks of 80 trials each. The number of different-response trials was double that in the other conditions in order to control for contingency effects (see Schmidt & Besner, 2008, and Schmidt, Crump, Cheesman, & Besner, 2007, for reviews).

### Analyses

Pupil size (area) was calculated by the eyetracking software and recorded in pixels. After each participant completed the task, a single measurement of a 4-mm dot was recorded from the same camera location (the placement of the camera was adjusted for each participant), and this was used as a reference point to convert all measurements from pixels to millimeters. Pupillary information from the onset of the stimuli to the when an initial saccade of >5 deg was made were used in the analyses. Trials in which the initial saccade was not within 45 deg of the direction toward the correct square were classified as invalid and were not included in the analyses, along with trials in which the time taken to make the initial large saccade was <200 ms or >2,500 ms. Incorrect trials were defined as those in which the initial saccade was made within 45 deg of the direction toward the incorrect square, and these were omitted from the main analyses, as well. Using these criteria, 88.43 % of the total responses were included in the analyses.

## Results

### Analysis of errors

The proportions of error trials were 4.5 %, 4.6 %, 3.6 %, and 5.5 %, respectively, for the congruent, neutral, same-response, and incongruent trials. A one-way analysis of variance (ANOVA) was conducted on these proportions and was found to be statistically significant [$F(3, 96) = 3.29$, $p = .024$, $r = .18$], and pairwise comparisons revealed that incongruent trials had more incorrect trials than same-response trials [$t(32) = 3.11$, $p = .004$, $r = .48$]. The other pairwise comparisons were statistically nonsignificant [congruent vs. neutral, $t(32) = 0.223$, $p = .825$, $r = .04$; congruent vs. same-response, $t(32) = –1.39$, $p = .173$, $r = .24$; congruent vs. incongruent, $t(32) = 1.73$, $p = .093$, $r = .29$; neutral vs. same-response, $t(32) = –1.51$, $p = .140$, $r = .26$; neutral vs. incongruent, $t(32) = 1.59$, $p = .123$, $r = .27$].

### Saccadic latencies

The mean RTs of valid saccades for congruent, neutral, same-response, and incongruent trials were 437.55, 460.53, 462.10, and 478.79 ms. A one-way repeated measures ANOVA was conducted and was found to be statistically significant [$F(3, 96) = 14.37$, $p < .001$, $r = .36$].

Pairwise comparisons revealed that congruent trials had the fastest RTs [vs. neutral, $t(32) = 3.48$, $p = .001$, $r = .52$; vs. same-response, $t(32) = 3.92$, $p < .001$, $r = .57$; vs. incongruent, $t(32) = 6.95$, $p < .001$, $r = .78$] and incongruent trials had the slowest RTs [vs. neutral, $t(32) = 2.55$, $p = .016$, $r = .41$; vs. same-response, $t(32) = 2.78$, $p = .009$, $r = .44$].

The difference between the RTs of neutral and same-response trials was nonsignificant [$t(32) = 0.27$, $p = .789$, $r = .048$]. To determine whether there was evidence for no difference between the RTs of the two conditions, a Bayes factor (Dienes, 2011) was calculated using Dienes's online calculator (www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/bayes_factor.swf). A Bayes factor of less than 0.33 indicates support for the null hypothesis, whereas one that is larger than 3.0 indicate support for the alternative hypothesis. Since we were investigating the difference between the same two trial types as Hasshim and Parris (2014), similar parameters were used to calculate the Bayes factor. Using a prior expected range of 6–45 ms for an effect with an assumed uniform distribution (i.e., all values were equally likely), the Bayes factor returned a value of 0.09, indicating strong support for the null hypothesis of no difference between the RTs of the two conditions.

## Pupil size

For each participant, the means of the maximum, average, and minimum pupil sizes during each trial up to the first saccade were obtained and analyzed separately. Table 1 shows the average maximum and minimum pupil diameters, the time after stimulus onset at which they occurred, and the time taken to make a saccade to the correct response. The mean pupil size at the onset of a trial was 4.191 mm (*SE* = 0.055), which indicates that there was a small initial dilation in pupil size, followed by a large constriction.

### Maximum pupil diameter

The mean maximum pupil diameters in the congruent, neutral, same-response, and incongruent trials were 4.204, 4.204, 4.203, and 4.202 mm, respectively. The repeated measures one-way ANOVAs for pupil diameter and the latency at which it occurred were nonsignificant [*F*(3, 96) = 0.017, *p* = .997, *r* = .013 and *F*(3, 96) = 0.646, *p* = .588, *r* = .082, respectively].

### Average pupil diameter

The average pupil diameters in the congruent, neutral, same-response, and incongruent trials were 4.138, 4.132, 4.127, and 4.123 mm, respectively. A repeated measures one-way ANOVA was nonsignificant [*F*(3, 96) = 1.586, *p* = .198, *r* = .127].

### Minimum pupil diameter

The minimum pupil diameters occurred at 316.92, 353.52, 344.38, and 355.41 ms after target onset. A repeated measures one-way ANOVA for the latencies was significant [*F*(3, 96) = 13.69, *p* < .001, *r* = .353], and follow up analyses revealed the latency to congruent trials to be faster than the latencies to neutral [*t*(32) = 6.06, *p* < .001, *r* = .73], same-response [*t*(32) = 3.27, *p* = .003, *r* = .50], and incongruent [*t*(32) = 5.66, *p* < .001, *r* = .71] trials; the other three conditions, on the other hand, were nonsignificantly different from each other [same-response vs. neutral, *t*(32) = –1.23, *p* = .230, *r* = .21; incongruent vs. neutral, *t*(32) = 0.34, *p* = .740, *r* = .06; and same-response vs. incongruent, *t*(32) = 1.85, *p* = .074, *r* = .31].

The mean minimum pupil diameters in the congruent, neutral, same-response, and incongruent trials were 1.879, 1.925, 1.929, and 1.954 mm, respectively, indicating that the pupil constricted to a size smaller than at target onset. The repeated measures one-way ANOVA was significant [*F*(3, 96) = 15.162, *p* < .001, *r* = .37]. Pairwise comparisons showed that congruent trials had the smallest minimum size [vs. neutral, *t*(32) = 3.91, *p* < .001, *r* = .57; vs. same-response, *t*(32) = 3.80, *p* = .001, *r* = .56; vs. incongruent, *t*(32) = 6.68, *p* < .001, *r* = .76] and that incongruent trials had the largest [vs. neutral, *t*(32) = 2.50, *p* = .018, *r* = .40; vs. same-response, *t*(32) = 2.95, *p* = .006, *r* = .46]. The difference between the minimum pupil sizes of neutral and same-response trials was nonsignificant [*t*(32) = 0.36, *p* = .720, *r* = 0.064]. As with RTs, a Bayes factor was calculated to determine whether there was evidence for no difference between the two conditions. Since there were no prior findings on such an effect using minimum pupil sizes, the only reference for the size of the effect was the difference between either neutral and congruent or incongruent and neutral trials. The larger of the differences, 0.045 mm, was used as the upper bound, whereas the lower bound was the proportionate equivalent to the one used in Hasshim and Parris (2014), 0.006 mm. The Bayes factor returned was 0.31, which is evidence for the null hypothesis of no difference between the two conditions.

## Discussion

Using an oculomotor version of the two-to-one response-mapping manipulation in the Stroop task, the RTs of saccadic responses and minimum pupil sizes were found to be consistent with the findings of the manual-response version used by

**Table 1** Average (with *SE*) maximum and minimum pupil sizes for each condition up to response, along with the average times at which they occurred after stimuli onset

| Condition | Maximum Diameter | | Minimum Diameter | | Saccadic RT |
|---|---|---|---|---|---|
| | Size (mm) | Latency (ms) | Size (mm) | Latency (ms) | |
| Congruent | 4.204 (0.112) | 188.40 (10.05) | 1.879 (0.047) | 316.92 (19.31) | 437.55 (17.80) |
| Neutral | 4.204 (0.110) | 186.46 (8.85) | 1.925 (0.049) | 353.52 (19.98) | 460.53 (18.43) |
| Same response | 4.203 (0.114) | 189.18 (9.85) | 1.929 (0.049) | 344.38 (21.33) | 462.10 (18.43) |
| Incongruent | 4.202 (0.113) | 192.77 (10.71) | 1.954 (0.050) | 355.41 (20.59) | 478.79 (20.96) |

Hasshim and Parris (2014). Saccadic RTs to congruent trials were fastest, followed by those of neutral and same-response trials, and the RTs to incongruent trials were the slowest. The Bayes factor for the difference between neutral and same-response trials indicated evidence for no statistical difference between their RTs. The preresponse pupil size measurements showed that the experimental conditions could not be differentiated by maximum and average pupil sizes. However, the minimum pupil sizes, which occurred after the initial pupil dilations, showed diverging condition effects similar to those of the saccadic RTs. Congruent trials resulted in the smallest minimum pupil size, whereas the minimum pupil size was largest for incongruent trials. The minimum pupil diameters for neutral and same-response trials were larger than in congruent trials, but smaller than in incongruent trials. However, they were nonsignificantly different from each other, with a Bayes factor that suggests evidence for no difference. Since the maximum pupil diameter occurred before a subsequent constriction and was found not to differentiate trial types, it can be inferred that the minimum pupil size was not due to residual effects of the initial dilation.

The latencies at which the maximum pupil diameter occurred were also shown not to differ by condition. In contrast, for the minimum diameter the average latency of congruent trials was different (faster) than those in the other three conditions. The noncorrespondence of these latencies with those of the saccadic RTs indicates that they are not a direct result of one another, and also indicates that the differences in the measurements of minimum diameter are not due to the different preresponse sampling times. The initial pupil dilation is consistent with studies that have looked at the time courses of pupillary measures (e.g., Laeng et al., 2011; van der Meer et al., 2010), and Laeng et al. (2011) suggested that the initial pupil dilation may be due to attentional changes brought about by the appearance of a stimulus. Since the identity of the stimulus cannot be predicted at the start of the trial, the similar level of pupil dilation might be a reflection of the cognitive system being prepared for any condition. As we noted in the introduction, pupil dilation is an indirect index of effort, which suggests that the subsequent constriction could reflect the level of effort required for attentional processing at the start for each of the different trial types. More specifically, since even non-color word neutral trials likely involve some form of conflict, whereas congruent trials involve mainly facilitation in this context, it is possible that the lesser constrictions in the neutral, same-response, and incongruent trials index the extra effort required to deal with the extra conflict.[1]

Researchers have posited that in addition to interference/conflict resolution at the response stage, performance in the Stroop task also requires conflict resolution at earlier processing stages (e.g., De Houwer, 2003b; Goldfarb & Henik, 2007; Klein, 1964; Parris, 2014; Schmidt & Cheesman, 2005; Sharma & McKenna, 1998; H. Zhang & Kornblum, 1998; H. H. Zhang et al., 1999) with the DO model attributing a portion of interference effects to overlap at a semantic level between the dimensions of the stimulus (i.e., S–S overlap; Kornblum & Lee, 1995). Along with the results of Hasshim and Parris (2014), the present results suggest no differences between same-response and non-color-word neutral trials in numerous measures of performance, thereby putting in question the utility of the two-to-one color response-mapping paradigm for measuring semantic or S–S conflict, and equally putting in question the presence of semantic conflict in the Stroop task. However, it should be noted that the previous results were obtained from oculomotor and manual-response paradigms, and thus are not necessarily generalizable to Stroop processing in other response modes. For example, Sharma and McKenna (1998) showed the components of Stroop interference to be different in manual and vocal response modes, with semantic-level components being more prominent in the latter, and they argued that the manual response mode indexed interference at the response level only (however, see M. Brown & Besner, 2001, for a reanalysis of the Sharma & McKenna, 1998, data evidencing semantic conflict with a manual response).

In the context of the DO model, neutral trials have neither S–R nor S–S overlap, which means that the relationship between the stimulus and response dimensions does not affect performance. However, many studies employing the Stroop task have calculated interference by subtracting neutral from incongruent trials and calculated facilitation by subtracting congruent from neutral trials, and have thus shown that interference and facilitation are the products of potentially different mechanisms (Goldfarb & Henik, 2007; Kane & Engle, 2003; Parris, 2014) and should not be directly compared. We have shown that same-response trials do not differ from neutral trials, and thus it seems increasingly unlikely that same-response trials could be used to differentiate the separate contributions of semantic (S–S) and response (S–R) conflict.

One possible explanation for the results from Hasshim and Parris (2014) is that S–S compatibility and S–R incompatibility work in opposing directions and cancel each other out. Since compatibility has a facilitative effect and incompatibility an inhibitory one (De Houwer, 2003b), it would be possible to have a zero net effect if the two were of similar magnitudes. Since pupillometric changes reflect the amount of effort exerted during the task (Laeng et al., 2012; Loewenfeld, 1993) it was assumed that any effort involved in dealing with opposing influences or S–S conflict alone would be measurable via pupillometry. Our data, however, showed no differences between same-response and neutral trials, suggesting no differential effort requirements.

[1] We thank an anonymous reviewer for this suggestion.

MacLeod (1998) suggested that the effect of facilitation could be produced by inadvertent reading, so that some responses were made via the reading of the word, resulting in faster responses to such trials (see also Kane & Engle, 2003). Since such cases would be classified as errors on incongruent trials but not on congruent trials (since the response was still correct), this would result in faster mean RTs to the latter trial type that would be included in later analysis. A similar scenario could occur for same-response trials, since the responses elicited by both dimensions would be correct. However, the analyses of error rates did not support the idea of inadvertent reading, since fewer errors to congruent and same-response trials, as compared to the other trials, would have been predicted. Although incongruent trials showed more errors than same-response trials (which can be attributed to additional response conflict), the error rates for neutral trials were individually nonsignificantly different from those of congruent and same-response trials, which does not reflect an advantage of inadvertent reading in the latter two conditions. More importantly, the inadvertent-reading hypothesis would have trouble accounting for data showing reverse facilitation effects as a result of increased task conflict (Goldfarb & Henik, 2007).

### Preresponse pupil measurement

Previous studies using pupillometry have focused on postresponse information and the average pupil size throughout the whole trial or block. The use of preresponse measures of pupil information is not common in studies of cognitive processes, and to our knowledge this has been the first study to show their usefulness in measuring Stroop interference effects. Typically, studies measuring changes in pupil size have reported the largest pupil dilation for incongruent trials, followed by neutral and congruent trials, with the most rapid dilation occurring after a response was made. However, as we previously described, such a measure has potential theoretical and methodological concerns. Being able to use preresponse pupillary information would support the argument for changes in pupil size being a measure that is independent of making a response decision. Moreover, using this measure would also allow for greater flexibility in the experimental procedure, since there would be no restriction on the trial duration or the response–stimulus interval between trials, which a postresponse measure would require.

Although the preresponse measure of pupil size displayed converging evidence with other measures of Stroop interference, the fact that it did not capture the full range of pupillary change in performing the task made it difficult to establish whether the same processes were responsible for both the pre and post pupillary effects. Richer and Beatty (1985) reported pupil dilation occurring before the onset of a stimulus, which suggests that the different aspects of responding, including preparation, execution, and proprioceptive feedback, are captured. It is likely that pupillary changes in the preresponse time frame would capture only some aspects of the cognitive process, albeit sufficiently to differentiate between standard Stroop effects.

To conclude, although researchers have argued that same-response trials index semantic conflict and have used the two-to-one response-mapping paradigm to isolate semantic conflict from response conflict in the Stroop task, our results with both pupillometry and saccadic RT measures showed evidence for no difference between same-response and neutral trials. These results support the suggestion that the previously measured effect likely indexes, or at the very least is inflated by, facilitation on congruent trials, and is not wholly due to semantic interference, casting doubt on the validity of using same-response trials in such an endeavor. The pupillometry data also showed that the Stroop effect can be measured by variation in pupil sizes *before* a response is made. This shows the utility of such a measure and its usefulness in measuring Stroop interference effects in task designs that do not allow for long response–stimulus intervals, widening the situations in which pupillometry can be used as a measure of Stroop effects.

**Author Note** This work was funded through a Bournemouth University studentship awarded to the first author, who was supervised by the second author. There are no conflicts of interest.

## References

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. doi:10.3758/s13423-012-0217-y

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91,* 276–292. doi:10.1037/0033-2909.91.2.276

Berggren, N., & Derakshan, N. (2014). Inhibitory deficits in trait anxiety: Increased stimulus-based or response-based interference? *Psychonomic Bulletin & Review, 21,* 1339–1345. doi:10.3758/s13423-014-0611-8

Brown, T. L. (2011). The relationship between Stroop interference and facilitation effects: Statistical artifacts, baselines, and a reassessment. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 85–99. doi:10.1037/a0019252

Brown, M., & Besner, D. (2001). On a variant of Stroop's paradigm: Which cognitions press your buttons? *Memory & Cognition, 29,* 903–904. doi:10.3758/BF03196419

Brown, G. G., Kindermann, S. S., Siegle, G. J., Granholm, E., Wong, E. C., & Buxton, R. B. (1999). Brain activation and pupil response during covert performance of the Stroop color word task. *Journal of the International Neuropsychological Society, 5,* 308–319.

Chen, A., Bailey, K., Tiernan, B. N., & West, R. (2011). Neural correlates of stimulus and response interference in a 2–1 mapping Stroop task. *International Journal of Psychophysiology, 80,* 129–138.

Chen, Z., Lei, X., Ding, C., Li, H., & Chen, A. (2013). The neural mechanisms of semantic and response conflicts: An fMRI study of practice-related effects in the Stroop task. *NeuroImage, 66,* 577–584. doi:10.1016/j.neuroimage.2012.10.028

Chiew, K. S., & Braver, T. S. (2013). Temporal dynamics of motivation–cognitive control interactions revealed by high-resolution pupillometry. *Frontiers in Psychology, 4,* 15. doi:10.3389/fpsyg.2013.00015

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332–361. doi:10.1037/0033-295X.97.3.332

De Houwer, J. (2003a). A structural analysis of indirect measures of attitudes. In J. Musch & K. C. Klauer (Eds.), The psychology of evaluation: Affective processes in cognition and emotion (pp. 219–244). Mahwah, NJ: Erlbaum

De Houwer, J. (2003b). On the role of stimulus–response and stimulus–stimulus compatibility in the Stroop effect. *Memory & Cognition, 31,* 353–359. doi:10.3758/BF03194393

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290. doi:10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5,* 781. doi:10.3389/fpsyg.2014.00781

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118,* 13–42. doi:10.1037/0096-3445.118.1.13

Goldfarb, L., & Henik, A. (2007). Evidence for task conflict in the Stroop effect. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1170–1176. doi:10.1037/0096-1523.33.5.1170

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology, 33,* 457–461.

Hasshim, N., & Parris, B. A. (2014). Two-to-one color–response mapping and the presence of semantic conflict in the Stroop task. *Frontiers in Psychology, 5,* 1157. doi:10.3389/fpsyg.2014.01157

Hock, H. S., & Egeth, H. (1970). Verbal interference with encoding in a perceptual classification task. *Journal of Experimental Psychology, 83,* 299–303. doi:10.1037/h0028512

Hodgson, T. L., Parris, B. A., Gregory, N. J., & Jarvis, T. (2009). The saccadic Stroop effect: Evidence for involuntary programming of eye movements by linguistic cues. *Vision Research, 49,* 569–574.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154,* 1583–1585.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General, 132,* 47–70. doi:10.1037/0096-3445.132.1.47

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

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: cognitive basis for stimulus–response compatibility—A model and taxonomy. *Psychological Review, 97,* 253–270. doi:10.1037/0033-295X.97.2.253

Kornblum, S., & Lee, J.-W. (1995). Stimulus–response compatibility with relevant and irrelevant stimulus dimensions that do and do not overlap with the response. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 855–875. doi:10.1037/0096-1523.21.4.855

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processing, 12,* 13–21. doi:10.1007/s10339-010-0370-z

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on Psychological Science, 7,* 18–27.

Loewenfeld, I. E. (1993). The pupil. In *Anatomy, physiology, and clinical applications* (vol. 1, pp. 695–707). Ames, IA: Iowa State and Wayne State University Press.

Logan, G. D., & Irwin, D. E. (2000). Don't look! Don't touch! Inhibitory control of eye and hand movements. *Psychonomic Bulletin & Review, 7,* 107–112. doi:10.3758/BF03210728

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163–203. doi:10.1037/0033-2909.109.2.163

MacLeod, C. M. (1992). The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General, 121,* 12–14. doi:10.1037/0096-3445.121.1.12

Macleod, C. M. (1998). Training on integrated versus separated Stroop tasks: The progression of interference and facilitation. *Memory & Cognition, 26,* 201–211.

Parris, B. A. (2014). Task conflict in the Stroop task: When Stroop interference decreases as Stroop facilitation increases in a low task conflict context. *Frontiers in Psychology, 5,* 1182. doi:10.3389/fpsyg.2014.01182

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology, 60,* 211–229.

Richer, F., & Beatty, J. (1985). Pupillary dilations in movement preparation and execution. *Psychophysiology, 22,* 204–207. doi:10.1111/j.1469-8986.1985.tb01587.x

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review, 110,* 88–125. doi:10.1037/0033-295X.110.1.88

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. doi:10.1037/0278-7393.34.3.514

Schmidt, J. R., & Cheesman, J. (2005). Dissociating stimulus–stimulus and response–response effects in the Stroop task. *Canadian Journal of Experimental Psychology, 59,* 132–138. doi:10.1037/h0087468

Schmidt, J. R., Crump, M. J., Cheesman, J., & Besner, D. (2007). Contingency learning without awareness: Evidence for implicit control. *Consciousness and Cognition, 16,* 421–435.

Sharma, D., & McKenna, F. P. (1998). Differential components of the manual and vocal Stroop tasks. *Memory & Cognition, 26,* 1033–1040. doi:10.3758/BF03201181

Siegle, G. J., Steinhauer, S. R., & Thase, M. E. (2004). Pupillary assessment and computational modeling of the Stroop task in depression. *International Journal of Psychophysiology, 52,* 63–76.

Simpson, H. M. (1969). Effects of a task-relevant response on pupil size. *Psychophysiology, 6,* 115–121.

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. doi:10.1037/a0016467

Stelmack, R. M., & Siddle, D. A. (1982). Pupillary dilation as an index of the orienting reflex. *Psychophysiology, 19,* 706–708.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662. doi:10.1037/0096-3445.121.1.15

Sullivan, K., & Edelman, J. (2009). An oculomotor Simon effect [Abstract]. *Journal of Vision, 9*(8), 380. doi:10.1167/9.8.380

van der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., … Wartenburger, I. (2010). Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology*, *47*, 158–169. doi:10.1111/j.1469-8986.2009.00884.x

van Veen, V., & Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: A functional MRI study. *NeuroImage, 27,* 497–504. doi:10.1016/j.neuroimage.2005.04.042

Wendt, M., Heldmann, M., Münte, T. F., & Kluwe, R. H. (2007). Disentangling sequential effects of stimulus- and response-related conflict and stimulus–response repetition using brain potentials. *Journal of Cognitive Neuroscience, 19,* 1104–1112.

Zhang, H., & Kornblum, S. (1998). The effects of stimulus–response mapping and irrelevant stimulus–response and stimulus–stimulus overlap in four-choice Stroop tasks with single-carrier stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 3–19. doi:10.1037/0096-1523.24.1.3

Zhang, H. H., Zhang, J., & Kornblum, S. (1999). A parallel distributed processing model of stimulus–stimulus and stimulus–response compatibility. *Cognitive Psychology, 38,* 386–432.