Investigating trial types putatively evidencing semantic conflict in the Stroop task

Nabil Hasshim

A thesis submitted in partial fulfilment of the requirements of Bournemouth University for the degree of Doctor of Philosophy

May 2016

Abstract

Interference in the Stroop task is thought to arise from various stages of processing, including the semantic and response stages. Different experimental methods have been used in an attempt to dissociate the cognitive processes involved in these stages. The work presented in this thesis evaluates two such methods that have been popular, namely the use of a two-to-one response mapping variant of the task and using colour-word distractors that are not valid response options (*non-response set trials*). The results from a series of experiments which utilised behavioural and eye-tracking measures, provided (Bayesian) evidence that two-to-one mapping trials do not involve additional interference compared to non-word neutral trials. Studies that have utilised this method are likely to have been measuring facilitation instead of the intended semantic-based interference, which has obvious ramifications to the conclusions of those studies. The experiments that evaluated non-response set trials indicated them as a better alternative, although during the course of the investigation, it was found that the make-up of Stroop interference is affected by experimental design. This is problematic to extant models of selective attention, as they cannot account for such findings. This led to further investigations of the cognitive mechanisms involved in processing relevant and irrelevant information during the Stroop task. The findings revealed that bottom-up implicit learning processes have a greater role in the allocation of attention and establishing task relevant stimuli, than previously thought. These concepts have generally not been given much consideration in theoretical accounts and the results from these experiments highlight their importance. The methodological and theoretical implications of the findings in this thesis are discussed in the context of theories of selective attention in the Stroop task and automaticity.

A note on the structure of this thesis

This thesis conforms to an 'article format' in which the middle chapters (Chapters 2 – 5) consist of discrete articles written in a style that is appropriate for publication in peer-reviewed journals in the field. The first and final chapters present synthetic overviews and discussions of the field and the research undertaken while a preface is presented at the beginning of each chapter to clarify the contribution of each manuscript to the overall aims and hypotheses of the thesis.

These manuscripts are at various stages of the publication/review process, and the status of each paper is summarised below. The main text in each chapter is presented as exact replications of the submitted manuscript and inevitably, there is some repetition as a consequence. The tables and figures are numbered within each chapter, while American English spelling is used in chapters three and four to conform to the requirements of the respective journals they were published in. There is also variation in the terminology used, depending on the experimental context of the individual chapters. The following are terms used in the thesis that refer to the same concepts:

- *Conflict*, *competition* and *interference* are used interchangeably
- *Semantic category conflict* and *stimulus-stimulus conflict* refer to *semantic conflict*
- *Different-response* trials and *response-set* trials refer to (standard) incongruent trials where the relevant and irrelevant dimensions of the stimulus elicit responses to different response options. The former term is included in parenthesis on every instance of the latter for continuity of reading of the thesis.

Status of manuscripts from this thesis

Experiments1 and 2 of Chapter 3 have been published as:

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.

Chapter 4 has been published as:

Hasshim, N., & Parris, B.A. (2015). Assessing stimulus-stimulus (semantic) conflict in the Stroop task using saccadic two-to-one color response mapping and preresponse pupillary measures. *Attention, Perception & Psychophysics*. *77*: 2601-2610.

Experiments 1 and 2 of Chapter 5 are currently under review.

Table of Contents

List of figures

Chapter 1

Chapter 3

List of tables

Acknowledgements

There are a number of people without whom this work may not have been completed, and to whom I am greatly indebted.

Ben, my supervisor. For taking a chance on me, and for your patience, guidance, and enthusiasm throughout.

All the friends in P104. For sharing the ups and downs of the PhD experience, and teaching me the importance of life outside it.

Sharon and Daphne, my social support network. For being there from before all this began. It made all this time being so far from home that much easier.

Martin and Jamie, our technicians at the department. For all the technical and logistical help, and always being a source for amusing discussions.

Melvin and Kerry, my academic mentors. For the training and experience gained from working with you; which was the best preparation for a PhD.

Finally, I would like to dedicate this work to my parents.

Author’s Declaration

I hereby declare that the work presented in this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature:

Chapter 1: Introduction

Selective attention and inhibitory control

Selective attention refers to the cognitive capacity to select a relevant and important part of the perceptual landscape while ignoring irrelevant parts. An important part of selective attention is inhibitory control, which involves ignoring, or overriding strong irrelevant mental processes or predispositions in order to successfully and efficiently regulate behaviour (MacLeod, 2007). It is a necessary process that makes it possible for us to overcome behaviours that are innate, or have become automatic through continued practice or learnt habit and instead, perform behaviours that are appropriate to a specific situation (Diamond, 2013).

An understanding of how, when and at what level of the cognitive system selective attention operates and how failure of specific inhibitory components differently affect performance is important. Not only will it permit a better understanding of the different types of impairments in selective attention observed in disorders such as attention deficit-hyperactivity disorder, schizophrenia, conduct disorder, posttraumatic stress disorder, depression, and obsessive compulsive disorder (Berggren & Derakshan, 2014), it also allows for a better understanding of regular aspects of life such as mental health, and social and cognitive development (see Diamond, 2013 for a review of inhibition as a core component of Executive Functions and its links to normal behaviour).

Experimentally, selective attention is typically measured using executive control tasks, which elicit cognitive conflict by presenting multiple sources of information that can be relevant or irrelevant to the performance of the task. The process of ignoring such irrelevant information calls on selective attention, which requires additional cognitive resources. This is exemplified by the classic example of the Stroop effect (Stroop, 1935; Klein, 1965). This effect shows that naming the

colour that a word is printed in takes longer when the word spells out a different colour (e.g. the word ‘red’ displayed in blue ink; an incongruent trial) compared to when the word spells out the same colour (e.g. the word ‘red’ displayed in red ink; a congruent trial) or when the word spells out a neutral word (one that is not associated with any colour, e.g. ‘table’). The Stroop task has been a popular paradigm, with the original paper (Stroop, 1935) being one of the most cited in experimental psychology (MacLeod, 1991). It has been described as the gold-standard measure of selective attention (MacLeod, 1992) and has been utilised in influential models of executive functions (e.g. Cohen et al., 1990; Dyer, 1973; Friedman et al., 2006; Glaser & Glaser, 1982; Miyake, 2000; Roelofs, 2003). Variants of the paradigm are also widely used in clinical settings as an aid to assess disorders related to frontal lobe and executive attention impairments (e.g. in attention deficit hyperactivity disorder, Barkley, 1997; schizophrenia, Henik & Salo, 2004; conduct disorder, Bauer & Hesselbrock, 1999; and anxiety, Matthews & MacLeod, 1985; see MacLeod, 1991; 2005 for comprehensive reviews of the Stroop task).

Early accounts of the Stroop effect describe it as exemplifying the difficulty in overcoming the more practiced behaviour of reading a word, which is irrelevant to the task, compared to the relevant, but less practiced behaviour of naming the colour (MacLeod & Dunbar, 1988). However, research has shown that selective attention involves a complex system composed of several different mechanisms such as conflict detection (Botvinick, Braver, Barch, Carter, & Cohen, 2001), biasing attentional resources to a task-relevant stimulus (Cohen, Dunbar, & McClelland, 2001; Desimone & Duncan, 1995; Kastner & Ungerleider, 2000) and goal-maintenance (Engle & Kane, 2004; Kane & Engle, 2003; Unsworth, Spillers, Brewer, & McMillan, 2011).

Subcomponents of the Stroop task

The performance of the Stroop task requires multiple processes and a widely used method to show that the Stroop task involves smaller subcomponents is the classic subtraction method (Donders, 1868/1969; Sternberg, 1969). This is done by comparing the performance between conditions that are thought to differ in the specific component under investigation (e.g. lexicality, semantic relatedness). For example, the influence of word frequency on interference can be measured by comparing a condition containing words that are highly frequent in the English language to another containing only low frequency words. The difference in performance is attributed to the theoretical difference between the two conditions. Giving another example in the context of the Stroop task, the finding that colour naming on non-word trials (e.g. xxxxx) is generally faster than when the letters make up a (colour-neutral) word (e.g. table) demonstrates that lexicality adds to interference. This means that that the mere presence of a word, regardless of its semantic content, produces interference and slows down the overall response to the task (often referred to as task set conflict; see Figure 1 for an illustration of how conflict is dissociated) (e.g. see Klein, 1964; and MacLeod, 1991).

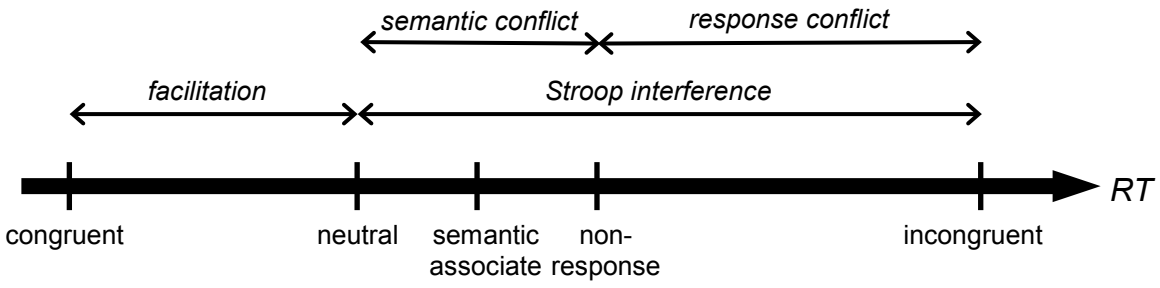

Figure 1 – Illustration of how the subtraction method is used to measure components of the Stroop task

An enduring question in selective attention research surrounds whether we can successfully ignore an irrelevant stimulus and at what point in the stream of

processing are we able to select the appropriate source of information. In the Stroop task, evidence shows conflict can occur independently at different stages of processing such as early stimulus encoding and lexico-semantic processing stages (e.g. Hock & Egeth, 1970; Luo, 1999; Parris, 2014) and at a later response output stage (e.g. Goldfarb & Henick, 2006; Roelofs, 2003; Posner & Snyder, 1975; van Veen & Carter, 2005). Although conflict can stem from either of these stages, the goal of research in selective attention is to assess whether the inhibitory system is able to resolve them within each of these stages, or whether interference builds up and is fully resolved at the later, response stage. However, to assess this effectively one needs robust measures of semantic- and response-based conflict.

Teasing apart semantic and response conflict

In its most common format, conflicts at the semantic and response levels are intertwined in the Stroop task. It is an inescapable fact that the performance of the task requires both processing of the stimulus (semantically) and selecting an appropriate output (response), and these two processes cannot be done independently of each other in the task since they are both expressed via response times. The question this thesis intends to address is how much semantic conflict contributes to Stroop interference and whether one can observe semantic conflict without response conflict. If conflict can be experienced at any particular level there might be a mechanism in place to resolve interference at that level without conflict being experienced at another level (conflict at the semantic level would still result in a difference in response times due to the delay caused at the semantic level, but no conflict at the level of initiating the response effector would be experienced). Conversely, the conflict that stemmed from the semantic level might only be resolved at the response level since that is where the response (and

the response effector e.g. which finger) has to be selected (i.e. conflict at response selection vs. conflict at concept selection). The standard version of the Stroop task leaves open the possibility of both scenarios occurring, and it is not possible to isolate these processes without modifying the task. Attempts to do this will be described below. However, before describing these experimental methods some of the extant models of the Stroop task and their accounts of Stroop conflict will be described.

Response and semantic conflict resolution in extant models of the Stroop task

The Dimensional Overlap taxonomy (Kornblum, 1992; Kornblum, Hasbroucq & Osman, 1990; Kornblum & Lee, 1995), along with other models of the Stroop task (e.g. De Houwer, 2003; Klopfer, 1996) assumes a multi-stage selection mechanism where information can convergence on common sets of intermediate components at each level and conflict that occurs at these components is resolved before going on to the following stage of processing. However, other prominent models such as the parallel distributed processing (PDP) model of Cohen et al. (1990, updated in Cohen & Huston, 1994) and the WEAVER++ model of Roelofs (2003), are single-locus, *response-competition based* models. This means that information from the different dimensions only converges at the response selection stage, and this is where all conflict, including conflict that arises from earlier semantic information, is resolved. Detailed descriptions of these models are given next.

Dimensional Overlap taxonomy

The DO taxonomy (Kornblum, 1992; Kornblum, Hasbroucq & Osman, 1990; Kornblum & Lee, 1995) provides a general framework for classifying executive control tasks according to the involvement of separate stimulus- and response-

based processes. The irrelevant information in a task can occur at the stimulus/semantic level, when the relevant and irrelevant dimensions of the stimulus give conflicting information about which is the target stimulus (stimulus-stimulus, or S-S conflict), or at the response level, when the response dimension conflicts with either stimulus dimension (i.e. relevant or irrelevant). This leads to the activation of an incorrect response in addition to the required response (stimulus-response, or S-R conflict), or a combination of both (S-S and S-R conflict; see Kornblum & Lee, 1999 for an in-depth review of the DO taxonomy).

In the verbal-response Stroop task, overlap at both S-S and S-R causes conflict at stimulus and response dimensions (Kornblum et al, 1990; Kornblum & Lee, 1995; Zhang, Zhang & Kornblum, 1999). It should be noted that although a manual response Stroop task does not have S-R overlap according to the taxonomy, the S-S overlap elicits different eligible responses, and thus produces response conflict that is stimulus-based (Egner, Delano & Hirsch, 2007; Kornblum et al., 1990; Kornblum & Lee, 1995; Zhang, Zhang & Kornblum, 1999). While some results have suggested that the manual response Stroop task does not involve stimulus-based, semantic conflict (Sharma & McKenna, 1998), there is good evidence of consistent semantic interference in the studies mentioned below using similar approaches (and also in a reanalysis of Sharma & McKenna's own data (Brown & Besner, 2001)).

Cohen, et al. (1990) PDP model

The PDP model by Cohen et al. (1990; updated in Cohen & Huston, 1994) describes the processing of stimuli as occurring via activation of a series of modules along two processing pathways, with the possibility of each module being activated by each pathway simultaneously. When different pathways (processes) activate a common module, it results in facilitation (better performance) or

interference depending on whether pathways activated by both the word and colour dimensions are similar or different. For a congruent trial, facilitation results since both word and colour activate the module by providing evidence for the same response, while on an incongruent trial the two pathways provide evidence for different responses, resulting in interference. An important component of the model is the momentary balance of evidence for each response is defined by the strength of evidence in favour of one minus the strength of evidence in favour of the other. When the mean difference between the two pieces of evidence crosses a threshold, selection occurs. Therefore, although the biasing of attention towards a certain pathway (attentional selectivity) begins early in processing, its effect is to reduce competition at the response module to allow for more efficient response (action) selection; thus as in early models of the Stroop task, conflict and its resolution occurs at the response level and not before.

An update of the Cohen et al. model within what is known as the GRAIN (graded, random, activation-based, interactive, and non-linear) framework by Cohen & Huston (1994) modified the connections between the modules to be bi-directionally excitatory such that the stimulus could also affect top-down attention. Importantly for present purposes, this model modified the response selection mechanism (actually removed it) due to the fact that their new model included bidirectional inhibitory connections between units in a module, which naturally causes them to compete. Thus, in principle the model could be modified to include an earlier semantic conflict resolution mechanism should a semantic module be added. Nevertheless, in its current form the model does not contain a semantic module meaning that there is no semantic conflict resolution mechanism.

Roelofs (2003) WEAVER ++ model

The WEAVER ++ model (Roelofs, 2003) is based on a network model of word production, and describes spreading activation of concepts in a network. For example, perceiving the colour red activates the concept of ‘red’, which also activates the superordinate concept of ‘colours’ and other colours (e.g. blue and green) to a lesser extent. The concept ‘red’ is connected to the word ‘red’ at a syntactic level, such that the activation of the concept leads to the activation of syntactic level processes and evidentially to the verbal production of the word. The relative level of activation in comparison to the other nodes determines subsequent action (e.g. since only one word can be the output, only the most active syntactic node will lead to the activation of its word form to be uttered), and this will only occur when the activated conceptual node has been flagged with a goal concept and is a possible response option. It should also be noted that in the model’s processing levels, colour naming and word reading do not interact until the lemma retrieval level, which is where response selection occurs. This indicates that the interference in the Stroop task happens after semantic processing. In other words, this model, like the PDP model, assumes that interference in the Stroop task is resolved only at the response stage.

Relevant to this thesis, although both single and multi-stage models assume that conflict occurs at both the semantic and response stages, the former posits that it is only resolved at the response stage (and thus semantic conflict cannot be controlled or modulated), while the latter model posits that semantic conflict can be reduced (or even resolved) at the semantic stage, before information is parsed to the response stage.

Experimental conditions used to dissociate response and semantic conflict

The typical way of tackling the research question of whether semantic conflict can be resolved independently from response conflict is by investigating whether the

effects indexing semantic conflict can be modulated while not affecting those that index response conflict. To achieve this, a condition that indexes only semantic conflict is first identified, and the performance of the trials of this condition, along with incongruent and baseline trials are compared. The difference in performance between the critical condition and the baseline is taken as a measure of semantic conflict, while the difference between incongruent and the critical condition is a measure of response conflict. In the literature, three conditions have been described that have been used as this critical condition, distinguishing semantic and response conflict in the Stroop task. Each of these conditions, *semantic associates, non-response set trials* and *same-response trials,* along with their limitations are discussed below.

Semantic associates

Semantic associates are words that are semantically or associatively related to a colour (e.g. *frog* – green, *sky* – blue). First used by Klein (1964), they have been used by many studies to isolate semantic conflict (e.g. Glaser & Glaser, 1989; MacKinnon, Geiselman & Woodward, 1985; Risko, Schmidt, & Besner, 2006; Stirling, 1979). Since semantic associates are related to colours only semantically or associatively, any interference has been attributed to non-response based processing. Research using semantic associates has observed interference using semantic associates to be much smaller than that to standard incongruent trials (e.g. Augustinova & Ferrand, 2012, Klein, 1964; Risko, Schmidt, & Besner, 2006; Sharma & McKenna, 1998). Any differences between semantic-associative interference and standard incongruent trials have been taken as evidence for response conflict.

However, semantic-associative interference can be explained with reference to the semantic-associates' non-semantic connection to the response

colours (e.g. Klein, 1964, Roelofs, 2003, also see Roelofs, 2000). That is, semantic-associative interference is the result of the activation of related response set colours and thus the semantic-associative Stroop task does not permit the unambiguous dissociation between semantic and response conflict and resolution. Furthermore, even if no response-based processes are involved when responding to semantic associates, they do not capture all of the semantic processes involved in the Stroop task, as will be noted in the next section.

Non-response set trials

Non-response set trials are trials where the irrelevant colour word is not used in the response set (e.g. the word ‘orange’ when the colour orange is never used and thus not one of the possible responses). Sharma and McKenna (1998) identified non-response set trials as involving an additional level of semantic processing (semantic relevance) when compared to semantic associates. This is an important point in the context of the literature as null effects on the performance of semantic associates have been used as evidence for manipulations not affecting any semantic processes. The fact that semantic associates do not capture semantic processes in their entirety leaves open the possibility of at least some semantic processes being affected.

The difference between the performance of incongruent trials (also known as response set (different-response) trials where the irrelevant word spells out a colour that is a possible response option) and non-response set trials, called *response set (membership) effect*, has been another popular way of isolating response conflict (e.g. Klein,1964; Milham et al., 2001; Risko et al., 2006; Sharma & McKenna, 1998). The response set effect describes interference due to the incongruous irrelevant colour word denoting a colour that is a possible response option.

In his review of the Stroop effect, MacLeod (1991) identified the response set effect as one of 18 well-established findings for which models of the effect need to account and indeed the response set effect is accounted for by extant models (described in a later section), and has been employed/investigated in many studies. While the response set effect is likely to be a good index of conflict that does not involve any semantic component (i.e. a measure of only response conflict), it is not clear whether it captures all the response conflict involved. At least one prominent semantic network model (Roelofs, 2003) accounts for the interference on non-response trials as being due to the representations of the non-response colours (at the response stage) having connections to the representations of the response set colours; in the same way that semantic associates might achieve interfering effects. Thus non-response trials do not elicit unique interference (since in the model, only response relevant colours are flagged) and that their slower performance compared to neutral baselines is due to the non-response colours' semantic link to response relevant colours. The effect on performance is smaller since the activation to these non-response colours is secondary. It should be noted that although the Roelofs (2003) model can be interpreted as attributing all conflict resolution to the response selection stage, teasing apart semantic and response conflict is not one of the main goals of the model. The relationship between non-response and incongruent trials will be explored in Chapter 5 of this thesis.

Same-response trials

The final trial condition to be covered that attempts to dissociate response and semantic conflict is same-response trials. This condition stems from the two-to-one colour-response mapping variant of the Stroop task, first introduced by De Houwer (2003). The paradigm draws from the ideas underpinning the DO model, and has

been gaining popularity in studies distinguishing semantic from response conflict since it claims to be able to remove the influence of response competition. The key attribute to this paradigm is that two colour responses are mapped on to one response button which allows it to dissociate S-S and S-R interference. Typically in studies employing the Stroop task, each response is assigned to a particular key on the keyboard or response box. This ensures that when an incongruent word is presented (e.g. 'red' in blue) the font colour and word will contribute evidence toward *different* response keys (i.e. 'red' will be assigned to the 'z' on the keyboard and 'blue' will be assigned to the 'm' key), ensuring competition at the response output level in addition to that at the semantic level. In the two-to-one paradigm, two colours are assigned the same response button, for example both 'red' and 'green' to the 'z' key (see Figure 2). When the incongruent word *red* is presented in 'green' both dimensions of the Stroop stimulus contribute evidence towards the *same* response key, but still activate different colour concepts (S-S interference).

This enables a distinction between two types of incongruent trials distinguished by whether the relevant and irrelevant stimuli share a common response. That is, the word can spell out a colour that does or does not share the same response as the colour of the stimulus. These *same-response trials* are thought to involve semantic category conflict but not response conflict (since both 'red' and 'green' share a common response) while *different-response trials* (standard incongruent trials) involve both semantic and response conflict.

Studies have shown that RTs progressively increase from congruent, same-response and different-response trials which has been used as evidence for the independent contributions of semantic and response conflict to the Stroop interference effect (Berggren & Derakshan, 2014; Chen, Bailey, Tiernan, & West,

2011; Chen, Tang & Chen, 2013; Schmidt & Cheesman, 2005; Steinhauser & Hubner, 2009; van

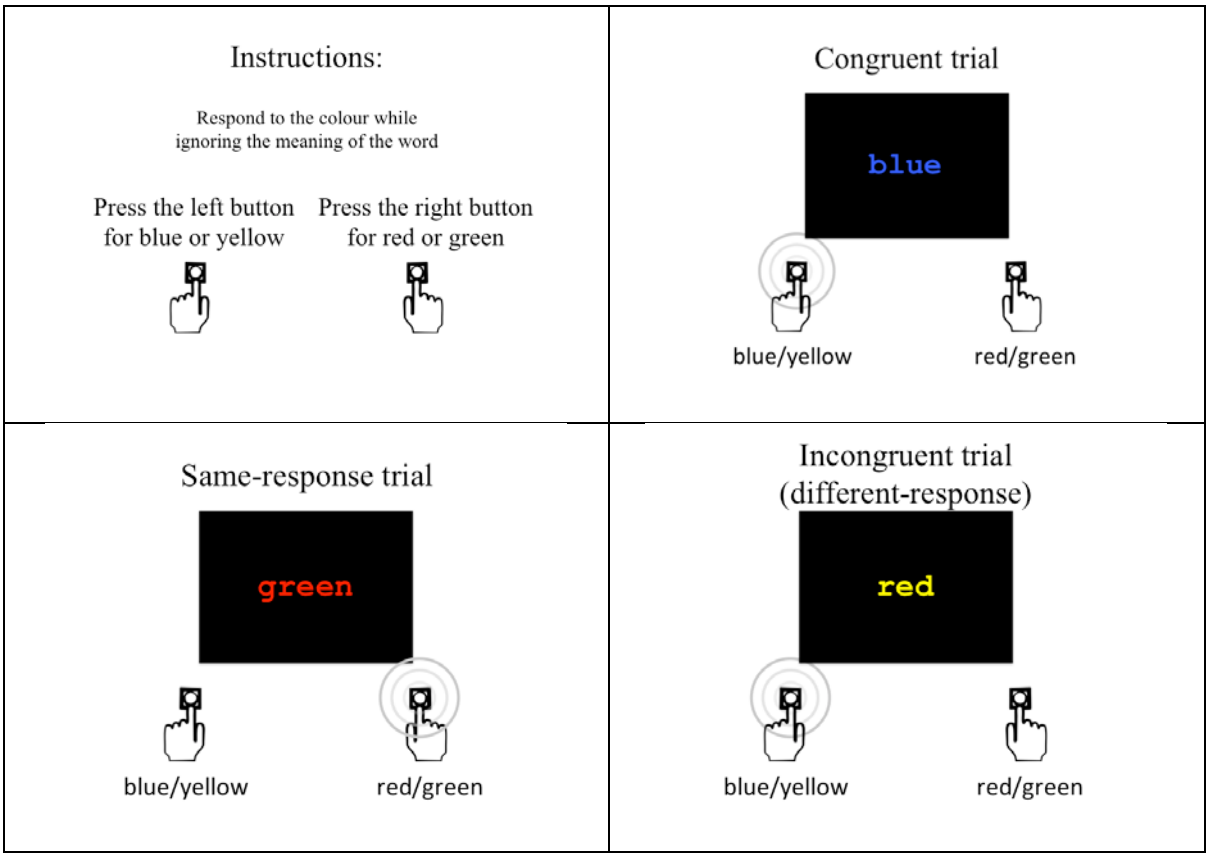

Figure 2 – Instruction and example trials of each condition in the two-to-one paradigm

Veen and Carter, 2005). As neat an idea as this two-to-one colour response mapping is, the studies employing it to dissociate semantic and response conflict have shared one major flaw: In all of the above studies the baseline control condition employed was the congruent trial, which means that the difference between same-response and congruent trials could be facilitation, as proposed by the original paper (De Houwer, 2003), not semantic interference. This possibility will be explored in Chapters 2, 3 and 4 of this thesis.

Importance of the current research

Early accounts of the Stroop effect describe it as exemplifying the difficulty in overcoming the more practiced behaviour of reading a word, which is irrelevant to the task, compared to the relevant, but less practiced behaviour of naming the colour (MacLeod & Dunbar, 1988). This means that the reading of a word is an automatic process where an 'automatic' process is defined as one that does not require attentional resources, happens without intent, and is ballistic (cannot be stopped once started; Brown, Gore & Carr, 2002; Neely & Kahan, 2001; Posner & Snyder, 1975). However, the demonstration that Stroop interference can be reduced (see Figure 3 for an illustration) using manipulations such as the narrowing of spatial attention (e.g. Besner, 2001; Besner et al., 1997; Besner, Risko, & Sklair, 2005; Labuschagne & Besner, 2015, Stolz & McCann, 2000) social priming (Goldfarb et al., 2011) and a post-hypnotic suggestion (e.g. MacLeod & Sheehan, 2003; Parris, Dienes & Hodgson, 2012; Raz & Campbell, 2011; Raz, Moreno- Iñiguez, Martin, & Zhu, 2007; Raz, Kirsch, Pollard, & Nitkin-Kaner, 2006; Raz et al., 2002; 2003) has been taken as evidence against the notion that word reading is automatic.

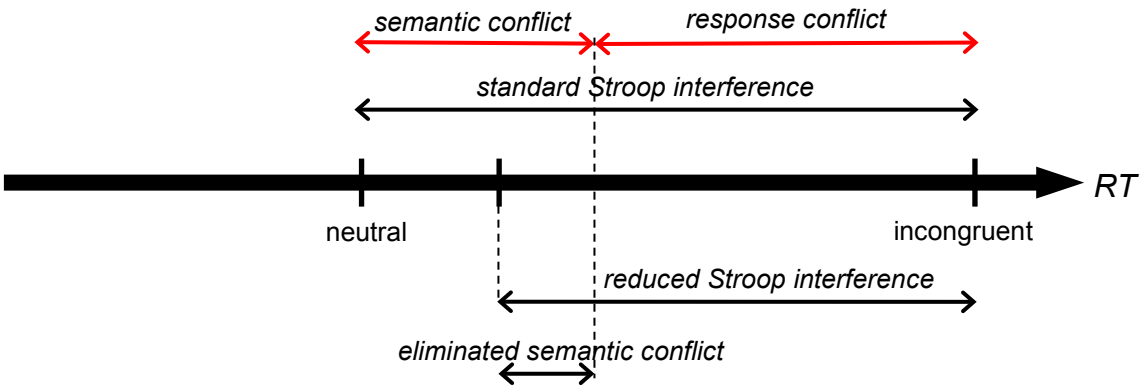

Figure 3 – This figure shows how a reduction in Stroop interference can be taken as evidence for a reduction of semantic conflict. Since Stroop interference is reduced by a large magnitude it is possible that at least some semantic conflict is reduced.

Other researchers (e.g. Augustinova & Ferrand, 2014; and Flaudias & Llorca, 2014) have rightly argued that simply showing a reduction of the Stroop effect is not sufficient to argue against the automaticity of reading. Since Stroop interference is made up of both semantic and response based processes and only the former is assumed to be automatic, Augustinova and Ferrand (2014) argued that such studies need to show that their manipulations affect semantic processes before a claim for control over 'automatic' processes can be made. Thus, to determine whether a process that results in semantic conflict in the Stroop task is preventable, one has to be sure that the measure being used is reliable and accurate. Augustinova and Ferrand also argued that manual button presses, which are popularly employed in the field, are not a good response modality to manipulate semantic conflict since they have been shown to mainly involve response competition (Sharma & McKenna, 1998). Thus they claim that any reduction in RTs would mainly reflect an effect on the predominant response conflict. They suggested the use of vocal responses instead (see Figure 4 for an illustration of this suggested make-up of Stroop interference when using manual responses). However, given the reports of semantic Stroop effects with manual

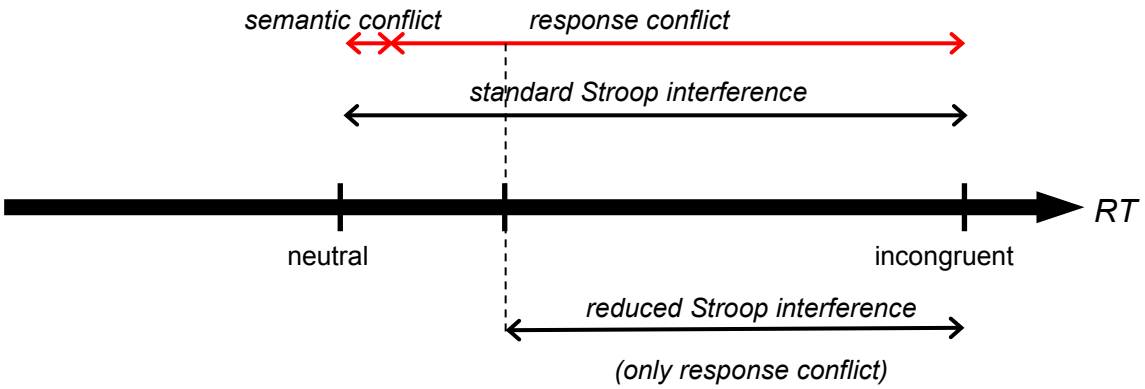

Figure 4 – Argument against a reduction in semantic conflict. Since Stroop interference is mainly made up of response competition in manual responses, reduction in Stroop interference is likely to affect only response conflict.

response in the studies covered in the section above, it is clear that work is needed to identify the best methods to assess response and semantic conflict before claims about reduction/elimination of any type of conflict can be made.

To that end, this thesis concerns the measurement of response and semantic conflict, in the Stroop task. The aim of the thesis is to determine whether putative measures of semantic and response conflict reliably and accurately index semantic and response level processing in the Stroop task. Only then will it be possible to determine the controllability or preventability of semantic processing during reading. The studies in the thesis do this by primarily evaluating the use of two popular measures of semantic competition covered above: same-response trials and non-response trials.

Rationale for initial studies

The first three experimental chapters of this thesis evaluate the utility of same response trials in isolating semantic and response interference. The two-to-one colour-response mapping approach has been employed to distinguish response and semantic conflict in the Stroop task because it allows for a trial type (same-response trials) that theoretically does not involve response conflict, unlike the other methods mentioned earlier.

Chapter 2:The Contribution of Semantic Category Conflict to the Stroop Interference Effect

Abstract

The Stroop interference effect is thought to include a semantic component in addition to a response selection component. Studies evidencing the contribution of semantic category conflict (SCC) have compared the standard incongruent Stroop trial (different-response trial) to trials where two response colours are mapped to the same response key (same-response trials) thereby eliminating response competition but maintaining semantic conflict. In Experiment 1, the semantic category conflict effect was replicated. In Experiment 2, same-response trials were compared to neutral and different-response trials (Experiment 2a); and neutral, different-response and non-response set incongruent trials (Experiment 2b) instead of the typically employed congruent trial baseline. In Experiment 3, Experiment 2b was re-run but performance was compared in mixed vs. pure blocks. The results suggest that measuring semantic conflict utilising congruent trials is mainly indexing congruency facilitation and also show evidence of a specific influence of mixed trial presentation on certain trial types, particularly different-response and non-response trials; a type of trial type homogenisation that does not affect all trial types. Finally, the results support the notion that response competition is the main driving force behind Stroop interference.

The studies in this chapter did not counterbalance the colours-response button mapping and thus there is a possibility that specific colour pairings might be an extraneous factor. The order of performing Experiments 2a and 2b were also not counterbalanced. These issues are addressed in Chapter 3.

The Contribution of Semantic Category Conflict To the Stroop Interference Effect

The classic Stroop task (Stroop, 1935) requires participants to respond as quickly and as accurately as possible to the colour in which a word is printed while ignoring the word's meaning. The *Stroop congruency effect* refers to the slower response times (RT) on incongruent trials (e.g. the word 'red' printed in blue) compared to congruent trials (e.g. the word 'red' printed in red). Alternatively, some researchers use neutral trials instead of congruent trials as the baseline. Neutral trials are trials where non-colour associated words or a series of repeated letters or symbols that do not form words are used (e.g. 'club' 'xxxxx' or '&&&&&', respectively). In this instance, one can compare incongruent and neutral trials, and congruent and neutral trials to give a measure of the Stroop interference effect and the Stroop facilitation effect, respectively. Confusingly, both the Stroop congruency effect and the difference between incongruent and neutral trials are often referred to as the Stroop interference effect in the literature, despite the former also being composed of facilitation effects. While there is a long history of debate about what makes the best baseline (see Jonides & Mack, 1984), the chosen baseline often depends on the purpose of the research (see Brown, 2011).

Regardless of the chosen baseline, the interference effect has been attributed to having to resolve conflict at the response stage when the colour and the meaning of the word each activate different responses (to be referred to as stimulus-response conflict (RC); Cohen, Dunbar, & McClelland, 1990; MacLeod, 1991). However, some researchers have posited that in addition to conflict resolution at the response stage, performance in the Stroop task also requires conflict resolution in earlier processing stages (e.g. De Houwer, 2003; Klein, 1964). Klein (1964) was the first to show that high and low frequency neutral

words that do not have any association to colours still slowed down RT, compared to a series of repeated letters (referred to as lexical conflict). He argued that this slight delay cannot be in the response stage as the words do not have any link to the set of response colours.

Along with response conflict and lexical (or task) conflict, Stroop interference is thought to involve stimulus-stimulus or semantic category conflict (SCC). Semantic category conflict refers to when both dimensions of the stimulus elicit the same semantic category (in this case: colours) and thus produce within-category competition. In an effort to distinguish response conflict and semantic category conflict researchers have modified the relationship between the relevant and irrelevant stimuli such that they either do or do not share a common response (De Houwer, 2003; Schmidt & Cheesman, 2005; Steinhauser & Hubner, 2009; van Veen & Carter, 2005). Typically in studies employing the Stroop task, there will be two or four possible responses and each response will be assigned to a particular key on the keyboard or response box. This ensures that when an incongruent word is presented (e.g. ‘red’ in blue) the font colour and word will contribute evidence toward different responses (i.e. ‘red’ will be assigned to the ‘x’ key on the keyboard and ‘blue’ will be assigned to the ‘n’ key), ensuring competition at the response output level (to be referred to as *different-response trials*). Comparing trials on which two response colours share the same response (*same-response trials;* e.g. ‘red’ and ‘yellow’ are mapped to the ‘x’ key) to different-response trials is thought to yield a purer measure of response conflict (De Houwer, 2003; Schmidt & Cheesman, 2005; Steinhauser & Hubner, 2009; van Veen and Carter, 2005). Since the stimulus is still incongruent on same-response trials, conflict is thought to be present at the level of semantic category. Same-response trials are semantic category-incompatible but response-compatible, while different-response trials are

both semantic category-incompatible and response-incompatible. Semantic category conflict effects have been calculated by subtracting RTs to congruent trials from RTs to same-response trials.

Using these measures of semantic category conflict and response conflict Schmidt and Cheesman (2005) observed a 24ms semantic category conflict effect and a 32ms response conflict effect. When semantic-associative incongruent words were included (e.g. words like fire (red) and sky (blue)), they observed no difference in RT between same-response trials and different-response trials, leading them to conclude that colour associate incongruent trials result in semantic conflict and not response conflict. This result corresponds to the idea that same-response trials involve only semantic and not response conflict.

The paradigm has seen popular use in different settings as a task that dissociates semantic and response conflict. For example, Chen, Bailey, Tiernan, and West (2011) used event-related brain potentials to measure medial frontal negativity (MFN) and conflict slow potential (SP) activity while participants performed the two-to-one Stroop task. They thus concluded that unique brain regions are involved in processing semantic and response conflict, and the results also show that these two types of conflict can be dissociated using neuroscience methods. Besides replicating the behavioural results, they showed that MFN was sensitive to trials involving response interference but not sensitive to those that only involve stimulus interference, while SP was elicited by both stimulus and response incongruent trials. In an fMRI study, van Veen and Carter (2005) used region of interest contrasts, which identifies the amount of overlap in activation between different conditions as defined by the two-to-one paradigm, to investigate whether stimulus and response conflict elicited distinct activation of different brain regions. In addition to the expected activation in the dorsolateral prefrontal and

anterior cingulate cortices, regions of the brain that have been established as critical in executive function processing, the contrasts showed that within these brain regions, there was no overlap in activation of the 2 contrasts, meaning areas with significant activation in one contrast were not significant in the other.

Other evidence of the validity of using same-response trials to measure semantic and response competition using same-response trials have been in showing that they have differing effects in other behavioural measures. For example Steinhauser and Hubner (2005) used ex-Gaussian distributional analysis to show that response conflict affected all the parameters of the RT distribution while semantic based conflict only affected the skewness of the distribution (it should be noted that while they used same-response trials, their definition and measurement of semantic conflict included other concepts such as task switching and condition mixing). Meanwhile, Chen, Tang and Chen (2013) showed different effects of practice on semantic and response conflict, with the former quickly decreasing and eliminated after the first block of trials, but the magnitude of the latter remaining constant throughout.

The use of the two-to-one paradigm to differentiate semantic and response conflict has also been used for applied research in clinical settings. For example, Berggren and Derakshan (2014) used a two-to-one paradigm of a Stroop-like task and showed that trait anxiety affects performance via response competition and did not affect interference at the semantic level.

These examples show the utility and importance of a task that can differentiate semantic and response competition and also the popularity of the two-to-one response paradigm in achieving this.

A potential concern with measuring SCC and response conflict using same-response trials is that same-response trials might also involve facilitation at the

level of response output (what will be termed ‘response facilitation’) since both the colour and the word dimensions elicit the same-response. This means that comparing same-response trials to incongruent trials is likely to overestimate the contribution of response conflict to the Stroop effect. It should also be noted that when congruent trials are used as a baseline under these conditions, there could be some response conflict even on congruent trials, since there are two colours associated with each response key. For example, when the word ‘red’ is presented in red the to-be-ignored dimension would conflict with the other colour that shares the response key. This added interference would increase RTs to congruent trials and thus lead to an underestimation of semantic category conflict. Response conflict would not be present on same-response trials since the irrelevant word matches the other colour, meaning it is less likely that conflict would be experienced when processing all components of the stimulus and the response. Given the presence of multiple processes that work in different directions, it is not clear how much response facilitation uniquely contributes to the RTs on each trial type. In sum, it is unclear how much interfering and facilitating components contribute to RT on both congruent and same-response trials when a response key is associated with two response colours, which means that it is difficult to accurately gauge semantic category conflict and response conflict.

Same-response trials have only ever been shown to produce interference relative to congruent trials raising the question as to how they would differ from non-colour related neutral word trials; another commonly used baseline condition in the Stroop task. Like same-response trials, neutral word trials involve lexical conflict (and hence make a better baseline condition than repeated letters for present purposes), but do not involve semantic category conflict, response conflict or response facilitation (See Table 1 for a summary of the effects present in all of

the conditions of the study). Same-response trials in contrast involve both semantic category conflict and response facilitation. If same-response trials do involve semantic conflict they should take longer to respond to than trials that do not involve semantic conflict, such as congruent and neutral trials. However, if same-response trials involve response facilitation and semantic category conflict one might not expect same-response trials to differ greatly from neutral trials since the semantic category conflict effect could be ameliorated by response facilitation.

Table 1: List of possible effects present in each condition

	Congruent	Neutral	Same-response	Different-response	Non-response
Response facilitation	✓	×	✓	×	×
Semantic competition	×	×	✓	✓	✓
Response competition	×	×	×	✓	×

Another trial type that has been used to differentiate semantic category conflict and response conflict is *non-response set* incongruent trials. The irrelevant dimensions for these trials denote colours that are not in the set of the response colours (e.g. the word ‘orange’ in blue, when the colour orange is not a possible response colour). An example of its use to measure semantic category conflict is by Milham et al. (2001) who tackled the question of whether response and semantic conflict activated different brain regions. In addition, they used neutral trials as the baseline for comparison.

They showed that the anterior cingulate cortex (ACC) was activated more on different-response trials when response conflict was involved and suggested that the ACC involvement in the task is limited to response conflict detection. The behavioural data however showed that while different-response trials were

responded to 44ms slower than the non-response set trials, the two incongruent conditions were non-significantly different; a finding that they attributed to an increase in RTs to neutral trials when different-response trials were presented in the same block. In contrast, Sharma and McKenna (1998) observed a significant difference between these two incongruent trial types (96.7ms) and thus the use of non-response trials to fulfil this role has to be questioned.

To investigate this the present study compared same-response trials to non-response set trials. Since these trials contain incongruent colours there must be semantic category conflict, but because the irrelevant dimension has no associated response there can be no response facilitation or response conflict. The semantic category conflict effect observed in other studies using congruent trials as the baseline was first replicated (Experiment 1). In Experiment 2a the congruent trials employed in Experiment 1 were replaced with non-colour related neutral trials while Experiment 2b was identical to 2a except for the inclusion of non-response set incongruent trials. Finally, Experiment 3 investigated the effect of mixed vs. pure blocks on semantic category conflict, response facilitation and response conflict.

Experiment 1

The aim of this experiment was to replicate the semantic category conflict effect (particularly the effect observed by Schmidt & Cheesman, 2005) relative to a congruent baseline to confirm it was replicable under present test conditions.

Method

Participants

Thirty-six students (12 male) from Bournemouth University participated in the study in exchange for an experiment credit or £5. The average age was 24.7 (*SD* = 6.4).

Design

The three experimental conditions were: 1) congruent trials, where the word spells out the colour of the text; 2) same-response trials, where the word spells out the colour that shares the same response mapping as the colour of the text; and 3) different-response trials, where the word spells out one of the two colours that are mapped to a different response than the colour of the text. Participants went through trials from all conditions, presented in random order.

Apparatus

Stimuli were presented using a standard PC running Experiment Builder software (SR Research Ltd, 2010) and responses were made via a standard Chiclet keyboard with coloured stickers on the corresponding response keys. The colours blue (RGB: 0; 112; 192) and green (RGB: 0; 255; 0) were assigned ‘c’ key while red (RGB:255; 0; 0) and yellow (RGB: 255; 255; 0) the ‘m’ key.

Materials

The stimuli for all three conditions were made up of four colour words (blue, green, yellow, red) presented in each of the four colours. The words were presented in lowercase, bold and in size 20 Courier New font on a black background.

Procedure

On each trial, participants were presented with a white fixation cross in the centre of the screen for 500ms followed by the Stroop stimulus which remained on the screen until a response was made. They were instructed to press the assigned key corresponding to the colour of the text as quickly as possible while ignoring the meaning of the word. An auditory feedback tone was given when an error was

made. Participants went through a practice block of 48 trials followed by four blocks of 72 trials, resulting in 96 experimental trials in each condition in total. Each block contained an equal number of trials from the three conditions presented in random order.

Results

Incorrect responses (5.2% across all conditions) were excluded from the analyses along with responses that were faster than 200ms and slower than 2500ms. This resulted in the total proportion of valid responses to be 94.6%.

A summary of the descriptive statistics is presented in Table 2. A one-way repeated measures Analysis of Variance (ANOVA) to determine whether there were differences in RTs to the congruent, same-response and different-response conditions was conducted. The difference across the three groups was significant ($F(2,70) = 31.32$, $p < .001$, $r = .56$). A priori follow-up tests revealed that RTs for the congruent condition ($M = 571.53$ ms, $SE = 20.68$) were significantly faster than those on same-response trials ($M = 601.56$ms, $SE = 23.69$; $t(35) = 4.42$, $p < .001$, $r = .60$) and different-response trials ($M = 626.54$ms, $SE = 25.34$; $t(35) = 7.15$, $p < .001$, $r = .77$) while the same-response condition was faster than the different-response condition ($t(35) = 3.95$, $p < .001$, $r = .56$).

Discussion

The results of Experiment 1 replicated the finding by Schmidt and Cheesman (2005) closely, where same-response and different-response trials were slower than congruent trials, and same-response trials were faster than different-response trials. This is congruous with the idea that both semantic category conflict and response conflict are involved in the Stroop task and that semantic category conflict contributes to the Stroop congruency effect. It should however be

noted that since the congruent trial baseline condition contains a facilitatory component, and under shared response conditions could even include an interference component since the other response colour associated with the correct response key is incongruent to the correct colour response, it is likely that semantic category conflict is being inaccurately estimated. It is also possible that response conflict is being overestimated since same-response trials could also include response facilitation.

Experiments 2a and 2b

Experiments 2a and 2b were designed to better understand the contributions of semantic category conflict and response facilitation to the pattern of results observed in Experiment 1. The purpose of Experiment 2 was to investigate the presence of semantic category conflict, response facilitation and response conflict effects and how they affect performance in the Stroop task. In Experiment 2a the congruent condition was replaced with a non-colour-related neutral word condition to determine whether interference, facilitation or no difference is observed when comparing same-response trials to a condition that does not involve semantic category conflict or response facilitation. To differentiate the effects of response facilitation from semantic category overlap, Experiment 2b introduced an additional incongruent condition, non-response set trials, that involve a colour word that is not part of the response set as the irrelevant dimension. While all participants completed Experiments 2a and 2b it was important to introduce these manipulations in different experiments (referred to as ‘Experiments’ to avoid confusion with the use of blocks used in the present work and because they were analysed separately) to control for the influence of one on the other. The non-response set trials involve semantic category conflict, but do not involve response

Table 2: Mean RT (SEs) of all conditions in all experiments

	Exp 1	Exp 2a	Exp 2b	Exp 3 (mixed)	Exp 3 (pure)
Congruent	571.53 (20.68)	-	-		
Same-response	601.56 (23.69)	714.07 (18.53)	679.41 (17.23)	623.00 (16.51)	605.05 (17.64)
Different-response	626.54 (25.34)	738.07 (19.08)	692.68 (20.30)	637.35 (16.26)	662.51 (18.58)
Neutral	-	709.03 (17.85)	673.32 (17.71)	610.25 (14.75)	604.55 (17.01)
Non-response set	-	-	698.38 (18.88)	633.35 (17.03)	612.52 (16.38)

trials and same-response trials, as the former involves neither semantic category conflict, nor response facilitation while the latter has both semantic category conflict and response facilitation. Comparing the non-response set trials to neutral trials and same-response trials would give an indication of the effect of semantic category conflict and response facilitation respectively.

If semantic category conflict has a strong influence on same-response trials, associated RTs should be longer than those to neutral trials and similar to those on non-response set trials. Given this outcome, previous studies utilizing same-response trials can be interpreted as accurately gauging semantic category conflict and response conflict. Alternatively, if response facilitation has a strong influence, associated RTs would be equal to or shorter than those to neutral trials and shorter than those to non-response set trials. Given this outcome, previous studies utilizing same-response trials can be interpreted as inaccurately estimating the contribution of semantic category conflict and response conflict to colour naming RTs.

It was also of interest to compare the non-response set condition to the different-response condition. The only difference between these conditions is that the different-response condition involves response conflict while the non-response set condition does not. Thus, the difference between their RTs would be a purer measure of response conflict i.e. one that is free from any effects of semantic category conflict and response facilitation. For Experiment 2b, it was expected that the RT for the non-response neutral condition to be slower than that of the neutral and same-response conditions but faster than the different-response condition.

Method

Participants

A new group of 36 students (6 male) were recruited from the same participant population and had an average age of 20.7 years (*SD* = 3.3).

Apparatus, Materials and Procedure

The apparatus and procedure were similar to those in Experiment 1. For Experiment 2a the design and materials were the same except the congruent trials was replaced with colour-neutral trials (neutral condition). The words used in the neutral condition were: 'DUE', 'WALL', 'STORY', 'MARVEL'. The colours used as responses were, yellow and red on one key, and orange (RGB:255; 192; 0) and green on the other.

In Experiment 2b, an additional an additional non-response set condition was included. The stimuli on these trials were words presented in the four target colours, as with the trials from the other conditions, but the words spell out colours that are not a valid response (i.e. not mapped on to any response button). These *non-response* colour words were 'PURPLE', 'WHITE', 'BLUE' and 'GREY'. Because of the additional condition, an additional block of 96 trials was added to Experiment 2b.

As with Experiment 1, trials from all conditions appeared within the practice block and all five experimental blocks in random order. The response set used different colours from Experiment 1 to ensure that the lexical properties of the words in all conditions were matched. Words in each condition had been matched for frequency and length using the English Lexicon Project (Balota et al., 2007). Participants performed Experiment 2a first.

Results

Experiment 2a

The proportion of valid responses for all participants amounted to .925 (*SD* = .049). The main effect from the repeated measures ANOVA across the three

conditions was significant ($F(1.63,57.18) = 11.36$, $p < .001$, $r = .41$), showing that there was a difference across the three conditions. Pairwise comparisons between each of the conditions showed that the different-response condition (M = 738.07ms, SE = 19.08) was significantly slower than the neutral (M = 709.03ms, SE = 17.85; $t(35) = 3.95$, $p < .001$, $r = .55$) and same-response (M = 714.07ms, SE = 18.53; $t(35) = 3.37$, $p = .002$, $r = .49$) conditions. No difference was detected between the neutral and same-response conditions ($t(35) = 1.06$, $p = .295$, $r = .17$). This non-significant effect is consistent with either evidence for no difference between the two conditions or simply with the absence of evidence for a difference. To determine if there was evidence for no difference between the two conditions, Bayes Factors (Dienes, 2011) were used, where the theory that there was a difference between the two conditions with the null hypothesis that there was no difference was contrasted. The predictions of the theory of a difference was modelled with a uniform distribution between -15 and 30ms i.e. any effect was as plausible as any other in the full range (30ms is the size of the semantic category conflict effect in Experiment 1, so defines the largest amount by which the two conditions would be expected to differ; -15 is used as the lower bound because the presence of response facilitation on same-response trials could feasibly lead to RTs shorter than those to neutral trials). The Bayes Factor was .46, indicating that there is not strong evidence supporting the null hypothesis (.33 and below being the cut off for strong evidence for the null; a Bayes Factor of 3 or above can be taken as strong evidence for a difference. See Dienes, 2011).

Experiment 2b

The proportion of valid responses for all participants was .931 (SD = .044). The mean latencies and accuracy of each condition are presented in Table 2. The repeated measures ANOVA measuring whether there is a difference across the

neutral, nonresponse colour, same-response and different-response conditions was significant ($F(3, 105) = 5.07$, $p = .003$, $r = .21$). Pairwise comparisons revealed that the neutral condition was faster than the different-response condition ($t(35) = 2.91$, $p = .006$, $r = .44$) and, importantly, the nonresponse colour condition $t(35) = 3.27$, $p = .002$, $r = .48$). As with Experiment 2a, no significant difference was identified when comparing neutral trials to the same-response condition ($t(35) = 0.85$, $p = .404$, $r = .14$). Applying the same upper and lower bounds as above, the comparison returned a Bayes Factor of .57 again indicating weak evidence for no difference between the two conditions. The difference between the same-response and different-response conditions was also non significant ($t(35) = 1.58$, $p = .124$, $r = .26$) in this experiment, contrasting with the results from Experiment 1. However, of direct interest to the research question for Experiment 2b, the non-response set condition was slower than the same-response condition ($t(35) = 2.69$, $p = .011$, $r = .41$), but surprisingly, was not significantly different from the different-response condition ($t(35) = .884$, $p = .383$, $r = .15$) a finding that contrasts with that from previous studies (Milham et al., 2001; Sharma & McKenna, 1998). To understand whether this latter non-significant effect is evidence for the null hypothesis or the absence of evidence for an effect a Bayes Factor using the magnitudes of the differences between the two conditions using a uniform distribution between 0ms-100ms (from Sharma & McKenna, 1998) was computed. The Bayes Factor was 0.19 indicating strong evidence for the null hypothesis of no difference.

Discussion

The motivation for Experiment 2 was to compare same-response trials to neutral as opposed to congruent trials, and to non-response set trials, which involve semantic category conflict, but not response facilitation. The results from this

experiment could be taken as evidence that RTs on same-response trials are more determined by response facilitation than semantic category conflict. Evidence for this comes from three sources; 1) No difference between neutral trials and same-response trials in Experiments 2a and 2b was detected, and in fact some, albeit weak, evidence for the null hypothesis was presented; 2) Same-response trials were responded to more quickly than non-response set trials which involve semantic category conflict, but not response facilitation; 3) An observation of a difference between neutral trials and non-response set trials. Thus, when taken together the results suggest that response facilitation is likely to contribute to RTs on same-response trials, and that the difference between same-response and congruent trials cannot be attributed solely to interference.

Neither semantic category conflict nor response facilitation were strong enough to enable the detection of a difference between same-response trials and neutral trials. It is possible that the similarity in RTs is driven by the concomitant forces of semantic category conflict and response facilitation affecting it in opposing directions. However, the lack of difference cannot be taken as strong evidence for the presence of two opposing effects or indeed as strong evidence for no difference between the two conditions. Nonetheless, the data suggest weak evidence for no difference and by comparing same-response trials to non-response set trials in Experiment 2b, the effects of response facilitation and semantic category conflict were dissociable.

One problem with the above interpretation of response facilitation is that it is somewhat dependent on the veracity of the observed mean of the non-response set trials. An unexpected result from the analysis of Experiment 2b showed that the RTs to the different-response and non-response set conditions were not significantly different. This result mirrors that found by Milham et al. who also

observed no significant difference between these two conditions, although their raw effect size was much larger than ours (6ms vs. 44ms). However, Sharma and McKenna (1998) observed a significant ~100ms difference between these two conditions. Indeed it was predicted since different-response trials involve response conflict whereas non-response set trials do not; the interpretation of the existence of response facilitation is predicated on the difference between the non-response set trials and the same-response trials. One difference between Sharma and McKenna (1998) compared to the studies of the current research, and Milham et al. (2001), is that Sharma and McKenna presented each trial type in a separate block. An experiment comparing mixed vs. pure block presentation of the stimuli employed in Experiment 2b is reported next.

Experiment 3

This experiment was designed to investigate the lack of difference between the different-response and non-response set conditions observed in Experiment 2b. As noted earlier, this was an unexpected result since Sharma and McKenna (1998), observed a difference between these two conditions, and is the basis for the present argument that same-response trials involve response facilitation. However, the current finding is not unique as Milham et al. (2001) also observed no significant difference comparing similar trials.

If a difference between non-response set and different-response conditions is interpreted as revealing response conflict, the result from Experiment 2b could be interpreted as showing that response conflict is not involved, even on different-response trials, under the present conditions. Once semantic category conflict is accounted for, the effects of response conflict become negligible. Such a position contrasts strongly with extant models of Stroop interference that place most interference as resulting from response conflict (Cohen et al. 1990; Melara &

Algom, 2003; Roelofs, 2003). However, the use of the manual response in the present study might make this interpretation more likely. Studies have reported a lack of Stroop interference with a manual response (see MacLeod, 1991), and an influential study has shown that Stroop interference with a manual response is not the result of fundamental inhibitory limitations but to a failure to fully engage goal maintenance mechanisms (De Jong, Berendsen & Cools, 1999). An alternative explanation, however, is that non-response set trials actually involve response conflict through their association with response set (different-response) trials (Roelofs, 2003).

One key difference between the present experiment and that of Sharma and McKenna's was that the different trial types in this experiment were presented in a random order in mixed blocks while they employed pure blocks. Milham et al. (2001) separated the same and different-response trials into different blocks but included neutral trials in both blocks and hence used mixed trial blocks. To investigate whether the use of mixed vs. pure blocks is the cause of the differing results, Experiment 2b was re-run on a different group of participants; once in a mixed block paradigm (direct replication) and once in pure blocks in counterbalanced order. It was predicted that the results of the mixed blocks will replicate Experiment 2b, where neutral and same-response trials are non-significantly different from each other but significantly faster than the other two conditions, while the different-response and non-response set trials will be non-significantly different as well. In the pure blocks conditions, it is expected that the different-response trials will have the slowest RTs while the other three types of trials will not show marked difference, as in Sharma and McKenna (1998).

Method

Participants

A new group of 36 students (18 male) were recruited from the same participant population and had an average age of 20.5 (*SD* = 3.18).

Design

A 4(condition: neutral, same-response, different-response, & non-response) x 2 (presentation format: pure blocks & mixed blocks) repeated measures design was used.

Apparatus, Materials and Procedure

In Experiment 3, participants essentially performed the trials in Experiment 2b twice. Once when the conditions were presented in mixed blocks (direct replication of Experiment 2b) and another time where the four conditions were presented in pure blocks the order of which was counterbalanced across participants.

The apparatus and procedure were similar to those in Experiment 2b. The colours used as responses were, yellow and red on one key, and purple (RGB: 204; 0; 255) and green on the other. The words used in the non-response set condition were orange, white, blue and brown. The response set used different colours from Experiment 2b as some participants had indicated an initial difficulty in differentiating some of the colours. Participants went through a practice block of 72 trials and trials from all conditions appeared within the practice block in random order. On the experimental trials, participants went through all the pure blocks either before or after all of the mixed blocks. The order of the pure blocks presented was counterbalanced, as was whether they performed the pure or mixed blocks first. The lexical properties of the words in all conditions were matched.

Results

The proportion of valid responses was for all participants were .932 (*SD* = .054) for the mixed blocks and .936 (*SD* = .051) for the pure blocks. The mean latencies of each condition are reflected in Table 2.

Omnibus ANOVA showed a significant presentation format (pure or mixed) by condition interaction, $F(3,105) = 5.91$, $p = .001$, $r = .23$. The follow-up repeated measures ANOVA measuring whether there is a difference across the neutral, nonresponse colour, same-response and different-response conditions was significant for both mixed and pure blocks ($F(3 ,105) = 6.86$, $p < .001$, $r = .25$ and $F(3 ,105) = 14.89$, $p < .001$, $r = .35$ respectively).

To further study the impact of presentation format on non-response set and different-response trials, two 3 x 2 ANOVAs were conducted, which were similar to the omnibus test except that in each, either the non-response set or different response conditions was removed. This was done to determine whether the interaction would still be significant in each case and if not, it would suggest that the omitted condition was the main source of the interaction. Results showed that in the analysis without the different-response trials, the interaction was non-significant ($F(2,70) = 1.15$, $p = .324$, $r = .13$) but significant when the non-response set trials were omitted ($F(2,70) = 6.18$, $p = .003$, $r = .28$).

Pairwise comparisons showed that the difference between the different-response and non-response set conditions was non-significant in mixed blocks ($t(35) = 0.54$, $p = .593$, $r = .09$), but significant when administered in pure blocks ($t(35) = 4.43$, $p < .001$, $r = .60$). The Bayes Factor for the former non-significant comparison was 0.15 indicating strong evidence for the null hypothesis of no difference. This showed that the different-response condition was significantly slower than the non-response set condition only when administered in pure blocks. To further explore the nature of this difference the comparison between the non-

response set RTs in the pure and mixed blocks which revealed $t(35) = 2.07$, $p = .046$, $r = .33$, showing that RTs increased in the mixed block. The same comparison in the different-response condition approached significance ($t(35) = 1.99$, $p = .055$, $r = .31$), which was the result of RTs increasing in the pure blocks.

Discussion

The main goal of Experiment 3 was to better understand the impact of mixed vs. pure block presentation on RTs to non-response set trials to enable a clearer interpretation of effects observed in Experiment 2b. Furthermore, it was aimed to reconcile the differences between the results observed in Experiment 2b and those of Sharma and McKenna (1998). Sharma and McKenna (1998) found that different-response incongruent trials were slower than non-response set trials, which data from Experiment 2b did not corroborate. In the present mixed block, RTs were similar to the findings of Experiment 2b, in which the difference between the different-response and non-response set conditions was non-significant. The follow up analyses suggests that condition mixing mainly affects the different-response trials, lowering their RTs compared to pure blocks. The data provide some evidence for an effect of mixing on non-response set trials, but this did not reach significance in the 3 x 2 interaction with different-response trials omitted.

In the present study's pure block, results were similar to Sharma and McKenna's finding that the different-response trials were significantly slower than the non-response set trials. Indeed, the other conditions were also consistent with their results since no difference between trial types that did not elicit response conflict was detected. The consequences of these findings will be discussed in the general discussion.

General Discussion

The aim of this Chapter was to examine the nature of same-response trials utilised in previous studies to separate semantic category conflict and response conflict. To do this, same-response trials were compared to neutral trials that involve lexical conflict, but no semantic category conflict or response facilitation, and non-response set trials that involve lexical conflict and semantic category conflict, but no response facilitation. Across two experiments no difference between neutral and same-response trials was detected and some evidence for the null hypothesis of no difference was presented. This suggests that there is little, if any, semantically based conflict effects involved in the standard colour Stroop task and that the effects found in previous studies were likely to be mainly due to congruency facilitation in the congruent trial baseline. To be clear, it is not being suggested that the two trial types are identical; even if there were strong evidence for no difference. What is being noted however is that the findings suggest that semantic category conflict has been over-estimated in previous experiments employing the congruent baseline since same-response trials are equal to neutral trials, and therefore any difference between congruent and same-response trials is largely due to the facilitation associated with congruent trials. Same-response trials might still be employed to dissociate semantic category conflict from response conflict, but any difference is unlikely to be more informative than when using neutral trials as the baseline. Following on from this, the findings also support the notion that the major contributor to Stroop interference is response conflict (Cohen et al. 1990; Melara & Algom, 2003; Roelofs, 2003). The effect is even more pronounced in pure blocks where different-response trials, which were the only ones with response conflict, had the slowest RT, while trials from the other conditions did not differ in their latencies.

Same-response trials to non-response set trials were also compared, which putatively involve semantic category conflict but not response facilitation or response conflict. In Experiment 2b, a difference between same-response trials and non-response set trials was observed, which is interpretable as showing that same-response trials involve response facilitation, suggesting that semantic category conflict is inaccurately measured using same-response trials. However, in Experiment 2b, there was strong evidence for no difference between non-response set and different-response set trials, which complicated the interpretation of the difference between same-response and non-response set trials. The lack of difference between non-response set and different-response trials indicating that, contrary to predictions, either the former involves response conflict, or the latter does not. The pure block condition of Experiment 3 showed that different-response trials do differ from non-response set trials in certain contexts. The results from Experiment 3 also showed that the effect of mixing was mainly driven by faster RTs to different-response trials in pure blocks compared to mixed blocks. The opposite effect was observed for non-response set trials. Indeed, the lack of a significant interaction effect when only neutral, same-response and non-response set trials were included in the analysis permits the conclusion that any observed differences between same-response and non-response set trials can be taken as evidence of the involvement of response facilitation on same-response trials. The finding of a significant effect at the level of pair-wise comparisons between non-response set trials in the mixed and pure blocks reduces the strength of this claim however. Hence, any interpretation that is based on comparisons involving non-response set trials in mixed blocks has to be made with caution.

One such interpretation is that performing the task in pure or mixed blocks somehow involves different mechanisms; particularly that response competition

does not have an effect in mixed blocks. The line of reasoning for this interpretation is as follows: In the mixed blocks, RTs of the non-response set trials when compared to same-response and neutral trials. This was taken as evidence for the presence of response facilitation in addition to semantic category conflict on same-response trials since the two effects work in opposing directions. Thus, non-response trials are a purer measure of semantic category conflict because non-response trials do not involve response facilitation. These two effects were thought to be of similar magnitude and acting in opposing directions, occluding each other when neutral and same-response trials are compared. Since non-response trials do not involve response conflict and the non-response trials and different-response trials did not differ (in both Experiments 2b and 3 mixed) it suggests that different-response trials are similar to non-response trials and do not involve response conflict in mixed blocks. Although this account fits the data well, the notion that response conflict competition is not involved in the Stroop effect (when presented in mixed blocks) is unlikely given the evidence for it in the literature.

There is a non-Stroop literature looking at the effect of mixing different trial types (e.g. Los, 1996; Lupker, Kinoshita, Coltheart and Taylor, 2003) and cases where the RT of easy trials becoming slower while those of difficult ones becoming faster is not uncommon (see Lupker et al., 2003). The theories that have been postulated to explain this phenomenon, which they termed a “homogenization” pattern, include shifting the time-based response threshold and having to adjust to the different number of strategies in each type of block. What is interesting in the data is how this homogenization pattern is mainly found in the different-response and non-response trials but not the other two conditions. It is possible that response conflict effects could be a factor for such a pattern to occur but further research is required to better understand the mechanisms at play.

Previous studies employing same-response and congruent trials to measure semantic category conflict and response conflict might need to be reinterpreted in light of the present results given the possible presence of response facilitation on same-response trials and the lack of a convincing difference between same-response and neutral trials. The difference in RT between same-response and congruent trials is likely to be due to congruency facilitation and not evidence for a semantic interference effect as proposed by De Houwer (2003). Schmidt and Cheesman's (2005) finding that semantic associates influence processing at a semantic level, might also be complicated by the presence of congruency facilitation effects. The results also have implications for neuroimaging studies such as van Veen and Carter (2005) that use the Stroop task to measure brain activity for the neural substrates of response and semantic conflict. They had used the two-to-one paradigm to measure semantic category conflict and response conflict and identified that the two activate non-overlapping areas of the brain. Even though unique brain regions were involved, it is likely that congruency facilitation was measured instead of semantic category conflict.

In sum, the results highlight the importance of using neutral trials as a baseline for measuring interference effects in the Stroop task. Measuring semantic conflict by comparing same-response trials and congruent trials is most probably mainly indexing congruency facilitation. The current results support the notion that response conflict is the main driving force behind Stroop interference. The results also provide evidence of a specific influence of trial type mixing on certain trial types, particularly different-response (standard incongruent trials) and non-response trials; a type of trial type homogenisation that does not seem to effect neutral and same-response trials.

Limitations of the current study

There are two potential methodological confounds in the present experiments. The first is the possible effect of contingency. While ensuring that there are an equal number of trials in each condition is common practice, it is possible that this may result in response contingency effects. Response contingency refers to the situation when a word stimulus is more strongly associated with one particular colour than another. Having two colours mapping onto the same button causes the proportion of responses to a stimulus to be different from chance (67% instead of 50%), meaning that words might be predictive of the response key to press (see Schmidt & Besner, 2008; Schmidt & De Houwer, 2012; for an in depth discussion of contingency and congruency effects). The second possible issue is that the colours in the response set and non-response set were not counterbalanced across participants. Although the lexical properties of the words such as frequency and length were controlled for, it is possible that certain colours might inherently be easier to inhibit. These issues are addressed in the following chapter.

Chapter 3:Two-to-one color-response mapping and the presence of semantic conflict in the Stroop task

This chapter aims to address possible weaknesses in the design of the Experiments in Chapter 2. The first is the possible effect of contingency. While it is common practice to ensure that an equal number of trials are presented in each condition, it is possible that this may result in contingency effects. This occurs when the irrelevant words are predictive of the response key to press (see Schmidt & Besner, 2008; Schmidt & De Houwer, 2012, for an in depth discussion of contingency and congruency effects) so that, for example, the word ‘red’ is presented more often in red than in any other colour. The colours in the response set and non-response set were also not counterbalanced across participants which might be an issue as it is possible that certain colours might be easier to inhibit.

This chapter reports new, methodologically improved versions of Experiments 2a and 2b. Experiment 1 from the previous chapter, which did not suffer from the issue of counterbalancing, is also presented in this chapter because the chapter presents a complete article, published in the journal *Frontiers in Psychology: Cognition.* A new, methodologically improved version of Experiment 3, which is not part of the published manuscript, is also included at the end of the chapter.

Abstract

A series of recent studies have utilized the two-to-one mapping paradigm in the Stroop task. In this paradigm, the word red might be presented in blue when both red and blue share the same-response key (same-response trials). This manipulation has been used to show the separate contributions of (within) semantic category conflict and response conflict to Stroop interference. Such results evidencing semantic category conflict are incompatible with models of the Stroop task that are based on response conflict only. However, the nature of same-response trials is unclear since they are also likely to involve response facilitation given that both dimensions of the stimulus provide evidence toward the same-response. In this study we explored this possibility by comparing them with three other trial types. We report strong (Bayesian) evidence for no statistical difference between same-response and non-color word neutral trials, faster responses to same-response trials than to non-response set incongruent trials, and no differences between same-response vs. congruent trials when contingency is controlled. Our results suggest that same-response trials are not different from neutral trials indicating that they cannot be used reliably to determine the presence or absence of semantic category conflict. In light of these results, the interpretation of a series of recent studies might have to be reassessed.

Two-to-one color-response mapping and the presence of semantic conflict in the Stroop task

The classic Stroop task (Stroop, 1935) requires participants to respond as quickly and as accurately as possible to the color in which a word is printed while ignoring the word's meaning. The Stroop congruency effect refers to the slower response times (RTs) on incongruent trials (e.g., the word "red" printed in blue) compared to congruent trials (e.g., the word "red" printed in red). This 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 in earlier processing stages (e.g., Klein, 1964; Sharma & McKenna, 1998; Zhang & Kornblum, 1998; Zhang et al., 1999; De Houwer, 2003; Schmidt & Cheesman, 2005). For example, semantic category conflict (an example of stimulus-stimulus conflict, or conflict that arises during stimulus processing independently of response processes) refers to when both dimensions of the stimulus elicit two different items from the same semantic category and thus produce within-category competition. In the case of a typical Stroop task, both the word and color dimensions activate color concepts, which results in competition at the semantic category level of "colors". It should be noted that studies in the literature typically use the general term "semantic conflict" while the current research defines semantic category conflict as its main source.

In an effort to distinguish response conflict and semantic category conflict researchers (De Houwer, 2003; Schmidt & Cheesman, 2005; van Veen & Carter, 2005; Steinhauser & Hubner, 2009) have used a variation of the Stroop task first

introduced in De Houwer (2003) that maps two color responses to one response button. Typically in studies employing the Stroop task, each response is assigned to a particular key on the keyboard or response box. This ensures that when an incongruent word is presented (e.g., “red” in blue) the font color and word will contribute evidence toward different -response keys (i.e., “red” will be assigned to the “z” on the keyboard and “blue” will be assigned to the “m” key), ensuring competition at the response output level. It is possible, however, to assign both “red” and “blue” to the “z” key. When the incongruent word red is presented in blue both dimensions of the Stroop stimulus contribute evidence toward the same-response keys, but still activate different color concepts. This two-to-one paradigm enables a distinction between two types of incongruent trials determined by whether the relevant and irrelevant stimuli share a common response. We will refer to these incongruent trials as different-response and same-response trials, respectively. Same-response trials are thought to involve semantic category conflict but not response conflict (since both “red” and “blue” share a common response) while different-response trials involve both semantic and response conflict.

This paradigm has been used to differentiate semantic and response based conflict. Comparing different-response trials to same-response trials is thought to yield a pure measure of response conflict, while comparing same-response trials to congruent trials is thought to measure semantic category (or sometimes called stimulus-stimulus) conflict (De Houwer, 2003; Schmidt & Cheesman, 2005; van Veen & Carter, 2005; Steinhauser & Hubner, 2009). Since congruent trials are also trials on which both dimensions of the stimulus contribute evidence toward the same-response, but also contribute evidence toward the same semantic item, it is assumed that the difference between the two conditions is semantic category

conflict. In short, same-response trials are semantic category-incompatible but response-compatible, different-response trials are both semantic category-incompatible and response-incompatible, and congruent trials are both semantic category compatible and response compatible.

Schmidt and Cheesman (2005) observed a 24 ms semantic category conflict effect and a 32 ms response conflict effect. In an fMRI study, van Veen and Carter (2005) compared brain activity associated with response and semantic conflict and showed that each activated unique brain areas. They found that the contrast between same-response and congruent trials, reflecting semantic category conflict, did not overlap with the contrast between different response and same-response trials. This was taken as evidence for the two types of conflict being detected and resolved by distinct regions of the brain. Using ex-Gaussian distribution analysis, Steinhauser and Hubner (2009) used same-response trials to get a purer measure of response conflict and observed response conflict in the Gaussian component of the distribution while task conflict (a form of semantic based conflict) was observed in the exponential component. Highlighting its utility, other recent studies have also employed the paradigm or similar two-to-one mapping paradigms (Wendt et al., 2007; Chen et al., 2011, 2013; Berggren & Derakshan, 2014).

In sum, in the present literature there is a debate as to whether semantic processes contribute to Stroop effects. Same-response trials have been used to provide evidence for the influences of semantic processes in the Stroop task, particularly semantic category conflict. According to some models such conflict should not exist since according to these models all interference in Stroop-like tasks is attributable to response conflict (Cohen et al., 1990; Roelofs, 2003). In light of the uptake of this paradigm, and the theoretical ramifications of the

presence of semantic category conflict, the present study sought to assess whether one can measure the contribution of semantic category conflict to Stroop effects using same-response trials. In Experiment 1 we aimed to replicate the semantic category conflict effect observed in previous studies. In Experiment 2, participants completed two counterbalanced blocks of the Stroop task. In one block, consistent with previous studies and Experiment 1, participants were exposed to congruent, same-response and different-response trials. In this block, non-color word neutral trials (e.g., “stage” in blue) were also included. In the other block, the congruent stimuli were replaced with non-response set incongruent stimuli (i.e., stimuli in which the word dimension is a color word that is not one of the possible response colors, e.g., “purple” in red). Furthermore, in both blocks we controlled for response contingency (Schmidt et al., 2007; Schmidt & Besner, 2008). We explain the motivation for each of these modifications below.

Inclusion of non-color word neutral trials

There is a potential issue with calculating semantic category conflict by comparing same-response trials to congruent trials as all previous studies have done. This is because, while congruent and same-response trials could involve response facilitation because the color concepts from both dimensions in each case provide evidence toward the same-response, congruent trials likely involve a unique semantic facilitation effect (Brown, 2011) which would result in faster RTs. Thus, this might not make them a suitable baseline to isolate semantic conflict since any difference in RT between the two trial types could be due in part to the presence of semantic facilitation. In order to remove the influence of semantic facilitation, Experiment 2A included non-color word neutral trials which do not involve semantic or response facilitation or semantic or response conflict. Slower RTs on same-response trials compared to neutral trials would be supportive evidence of

semantic category conflict, as is predicted by multiple-stage accounts (Klein, 1964; Zhang & Kornblum, 1998; De Houwer, 2003; Schmidt & Cheesman, 2005; Zhang et al., 1999). Should same-response trials be faster than neutral trials it would be evidence for an effect of response facilitation on same-response trials, not solely semantic conflict as has previously been assumed. Moreover, it would mean that studies comparing same-response and different-response trials for a purer measure of response conflict would also have to be reassessed. Importantly, even evidence for no difference between the trial types would be meaningful since it would indicate that same-response trials should not be used to infer the presence or absence of semantic category (or stimulus–stimulus) conflict.

Inclusion of non-response set incongruent trials

Non-response set incongruent trials (e.g., “purple” printed in blue, when the color purple is not used on any trial) involve semantic category competition but no semantic facilitation, since both dimensions of the Stroop stimulus activate different color concepts, but little or no response competition (Klein, 1964; Sugg & McDonald, 1994) and response facilitation because the word dimension is not a possible response. If responses to same-response trials are faster than those to non-response set trials it would provide support for the existence of response facilitation on the former. Moreover, since non-response set trials do not include response facilitation, the comparison between these trials and neutral trials might give a better measure of semantic category conflict than same-response trials. Finally, the comparison between non-response set trials and different-response trials might provide a purer measure of response competition.

Controlling for response contingency

Recent work has shown effects of contingency on congruent trial RTs (Schmidt et al., 2007; Schmidt & Besner, 2008). The contingency effect shows that the associations between word and response are implicitly learnt throughout an experiment and used to predict specific responses to each word, which facilitates RTs to trials where the correct response is highly correlated to the word. This is the case with congruent trials since they often make up half the trials. For example, with a four-response Stroop task there are only four possible word-color combinations to create the congruent stimuli whereas there are a possible 12 word-color combinations when creating incongruent stimuli. This means that the words are more often associated with their congruent color counterparts. When contingency is absent, RTs to congruent trials increase (see Schmidt et al., 2007; Schmidt & Besner, 2008). Although not explicit, contingency has been controlled in some studies employing same-response trials (De Houwer, 2003; Schmidt & Cheesman, 2005), while it was not controlled in others (van Veen & Carter, 2005; Steinhauser & Hubner, 2009). Importantly for present purposes, contingency is also likely to affect same-response trials. Since Experiment 2A involved congruent trials, we controlled for contingency by having twice as many different-response trials than congruent and same-response trials, which ensures that for each color word, the probability of any of the responses being the correct response is be equal. Thus, any difference remaining between same-response/congruent trials and other trials types would therefore represent influences attributable to other factors.

Summary

Thus the main goal of the current research was to determine whether same-response trials truly index semantic category conflict by addressing possible influences of semantic and response facilitation while controlling for contingency.

The critical comparisons in the experiment were as follows: (1) Same-response trials vs. neutral trials: the difference between these trials would be a more accurate measure of semantic category competition since neutral trials involve neither response facilitation nor semantic category conflict; (2) Same-response trials vs. non-response set trials: the comparison of these trials would also inform us whether there is facilitation involved when processing the former as an inhibition only based account of same-response trials predicts no difference between the two, while one that includes a response facilitation component would predict faster responses to same-response trials; (3) Same-response trials vs. congruent trials when contingency is controlled: If contingency does have an effect, we would expect the difference between the two conditions to be smaller when it has been controlled for; (4) Same-response trials vs. different-response trials when contingency is controlled: If contingency is affecting RTs to same-response trials the difference observed between these two trial types in some previous studies is likely to overestimate response competition.

Before reporting the key experiment of the paper (Experiment 2), we first report a replication (Experiment 1) of the two-to-one mapping paradigm as it has been most commonly employed: Including different-response, same-response and congruent trials but without neutral and non-response set trials and without controlling for contingency. To foreshadow the findings of this paper, using Bayesian statistics we provide evidence for no difference between neutral and same-response trials suggesting that studies utilizing same-response trials to measure semantic category conflict or response conflict will have to be reassessed.

Experiment 1 is reported to establish the magnitude of the effects under present conditions and for later use in the calculation of Bayes Factors where we

test whether any null effects observed are evidence for the absence of an effect or the absence of evidence for an effect and was not run as a within-subjects manipulation with Experiment 2 to avoid learned contingencies carrying over. Experiment 2 consisted of two counterbalanced blocks of trials in which contingency was controlled. In one block, only neutral, same-response, congruent and different-response trials were included (Experiment 2A). The other block was the same except that the congruent trials were replaced by non-response set trials (Experiment 2B).

Method

Participants

Two different groups of 36 students (12 male in Experiment 1, 6 in Experiment 2) participated in each of the experiments in exchange for course credit or £5. The average age was 24.7 (SD = 6.4) for Experiment 1 and 21.0 (SD = 5.0) for Experiment 2.

Apparatus and Materials

Stimuli were presented using standard PC running Experiment Builder software (SR Research Ltd, 2010) and responses were made via a standard chiclet keyboard with colored stickers on the corresponding response keys. In Experiment 1, the colors blue (RGB: 0; 112; 192) and green (RGB: 0; 255; 0) were assigned “c” key while red (RGB: 255; 0; 0) and yellow (RGB: 255; 255; 0) the “m” key.

For Experiment 2 the neutral words used were DUE, WALL, STORY, and MARVEL. In addition to the colors used in Experiment 1, the colors orange (RGB: 255; 127; 0), pink (RGB: 255; 20; 147), purple (RGB: 0; 125; 255), and white (RGB: 255; 255; 255) were used. For each participant, four of the colors were used as responses while the other four were used as the word dimension in the non-response trials. The colors that were assigned as responses and distractors

were counterbalanced as was which colors were mapped on to the response keys and the order of which participants performed Experiments 2A and B. Words in each condition had been matched for frequency and length using the English Lexicon Project (Balota et al., 2007). Each word was presented in the four response colors equally often. The words were presented in lowercase, bold, and in size 20 Courier New font on a black background.

Procedure

On each trial, participants were presented with a gray fixation cross in the center of the screen for 500 ms followed by the Stroop stimulus which remained on the screen until a response was made. They were instructed to press the assigned key corresponding to the color of the text as quickly as possible while ignoring the meaning of the word. An auditory feedback tone was given when an error was made. Participants went through a practice block of 48 trials. Before the experiment participants were given instructions verbally and written instructions were presented on the screen before each block commenced.

In Experiment 1, participants went through four blocks of 72 trials, resulting in 96 experimental trials in each condition in total. Each block contained an equal number of trials from the three conditions (congruent, same-response, and different-response) presented in random order.

In Experiment 2A, participants went through three blocks of 80 trials, which consisted of 48 trials each of the congruent, same-response, and neutral conditions and 96 trials of the different-response condition. Having twice as many different-response trials is necessary to control for contingency by ensuring that the correct response to each word presented is equal for the two response buttons.

In Experiment 2B, participants went through three blocks of 64 trials which consisted of 48 trials each of the same-response, different-response, neutral, and non-response trials. It was not necessary to have different number of trials of each trial type as congruent trials were not presented.

Results

Experiment 1

Incorrect responses (5.2% across all conditions) were excluded from the analyses along with responses that were faster than 200 ms and slower than 2500 ms. This resulted in the total proportion of valid responses to be 94.6%.

We conducted a one-way repeated measures analysis of variance (ANOVA) to determine whether there were differences in RTs to the congruent, same-response and different-response conditions. The difference across the three groups was significant [$F(2,70) = 31.32$, $p < .001$, $r = 0.56$]. A priori follow-up tests revealed that RTs for the congruent condition ($M = 571.53$ ms, $SE = 20.68$) were significantly faster than those on same-response trials [$M = 601.56$ ms, $SE = 23.69$; $t(35) = 4.42$, $p < .001$, $r = 0.60$] and different-response trials [$M = 626.54$ ms, $SE = 25.34$; $t(35) = 7.15$, $p < .001$, $r = 0.77$] while the same-response condition was faster than the different-response condition [$t(35) = 3.95$, $p < .001$, $r = 0.56$]. Importantly, these results replicate the findings from previous studies showing a semantic category conflict effect (see Figure 1).

The omnibus ANOVA for error rates across the three conditions was statistically significant [$F(2,70) = 12.85$, $p < .001$, $r = 0.39$]. Follow-up pairwise comparisons showed that the error rate in the different-response condition (6.8%) was significantly more than the same-response [4.4%; $t(35) = 3.87$, $p < .001$, $r = 0.54$] and congruent [4.5%; $t(35) = 4.03$, $p < .001$, $r = 0.56$] conditions. The error

rates between same-response and congruent trials were non-significantly different [$t(35) = 0.378$, $p = .708$, $r = 0.06$].

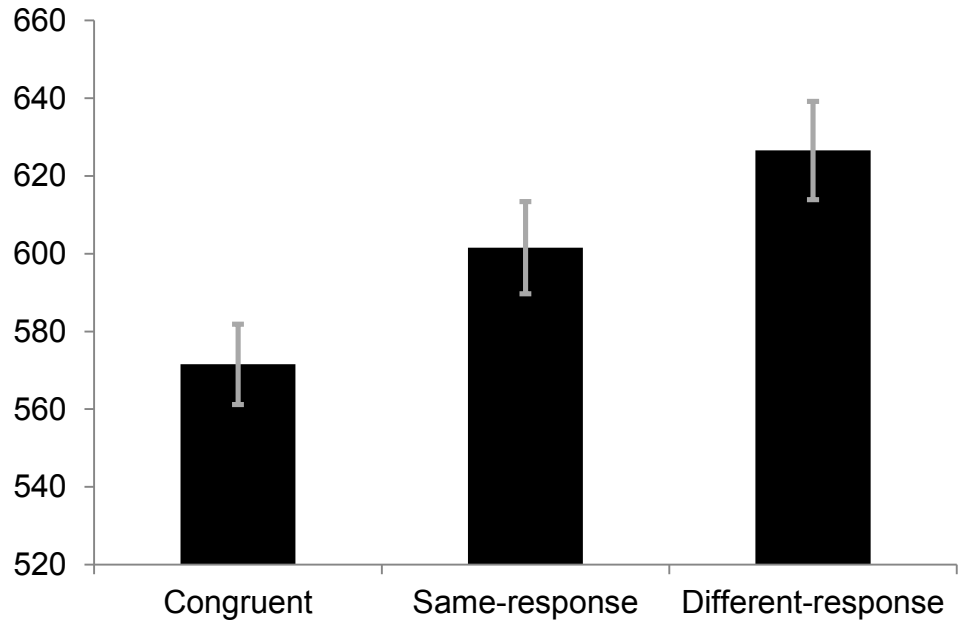

Figure 1: Mean RTs (in ms) for each condition in Experiment 1. Error bars represent standard errors.

Experiment 2A

The same exclusion criteria as Experiment 1 were used which resulted in the proportion of valid responses to be 95.5%. A one-way repeated measures ANOVA was conducted and was found to be statistically significant [$F(3,105) = 8.72$, $p < .001$, $r = 0.23$; see Figure 2]. In the introduction a set of critical comparisons were outlined. Data from this block permit us to test critical comparisons 1, 3, and 4.

No difference was observed between same-response ($M = 602.35$ ms, $SE = 17.40$) and neutral ($M = 601.55$ ms, $SE = 13.75$) trials [$t(35) = 0.089$, $p = .929$, $r = 0.015$]. To determine if there was evidence for no difference between the two conditions, we used a Bayes Factor (Dienes, 2011), where we contrasted the theory that there was a difference between the two conditions with the null hypothesis that there was no difference (0.33 and below being the cut off for strong evidence for the null; a Bayes Factor of 3 or above can be taken as strong

evidence for a difference). To calculate the Bayes Factor we used 6–45 ms as the range and assumed a uniform distribution (i.e., all values within this range were equally likely). This range was chosen based on previous work in our and other labs (De Houwer, 2003; Schmidt & Cheesman, 2005; van Veen & Carter, 2005; Chen et al., 2011, 2013; Parris et al., 2012a,b, 2013; Parris & Dienes, 2013) considering the theory under test (i.e., that semantic category conflict exists/is measurable using same-response trials).[1] For the difference between neutral and same-response trials a Bayes Factors of 0.17 was returned, providing strong evidence for the null hypothesis of no difference relative to the alternative hypothesis. In other words the observed mean difference and SE of the difference between the same-response and neutral trials were sufficiently far from the expected range to be considered evidence for the null. This finding is important and suggests that, at least when using RT as the dependent variable, same-response trials do not index semantic category competition.

For critical comparison 3 we calculated a Bayes Factor for the difference between congruent (M = 601.54 ms, SE = 15.90) and same-response (M = 602.35 ms, SE = 17.40) trials [t (35) = 0.095, p = .925, r = 0.016]. Again we assumed a uniform distribution with all values between 6 and 45 ms being equally likely. This yielded a Bayes Factor of 0.15 providing strong evidence for no difference

[1] To calculate a Bayes factor one must first consider the expected magnitude of the effect under investigation. Schmidt and Cheesman (2005) used experimental methods that most closely resemble the present study and observed a semantic category conflict effect of 24 ms when using congruent trials as the baseline. The size of this effect is comparable, but is at the lower end of the 24–45 ms range observed in other studies using two-to-one mapping in Stroop task (However, the larger value was in a study that presented word primes prior to the Stroop stimuli which may have encouraged greater word processing and thus greater facilitation (Parris et al., 2013). The remaining values range between 15 and 27 ms. If 15 ms, then of the 31.6 ms average raw effect size for the same-response vs. congruent trial comparison we might expect 15 ms to be facilitation (neutral-congruent) and 16.6 ms semantic category interference; in other words the RT for neutral trials falls roughly half-way between congruent and same-response trials. If 27 ms, then we might expect only 3 ms interference. We would certainly not expect the difference between same-response and neutral trials to be greater than the 45 ms maximal difference observed between same-response and congruent trials so we set 45 ms as the upper bound of expected range. To set the lower bound we must consider the smallest raw effect size that would be theoretically interesting. Notably harder to define we selected 6 ms since this is the raw effect size of a recent study using the Stroop task that was theoretically meaningful (Risko et al., 2006).

between the two conditions. This finding contrasts with previous studies showing a semantic category conflict effect when contingency is controlled (De Houwer, 2003; Schmidt & Cheesman, 2005).

For critical comparison 4 we compared same-response (*M* = 602.35 ms, *SE* = 17.40) trials and different-response (*M* = 633.31 ms, *SE* = 15.7) trials when contingency was controlled. As in Experiment 1 here we observed a significant difference between the two conditions [$t(35) = 4.54$, $p < .001$, $r = 0.61$].

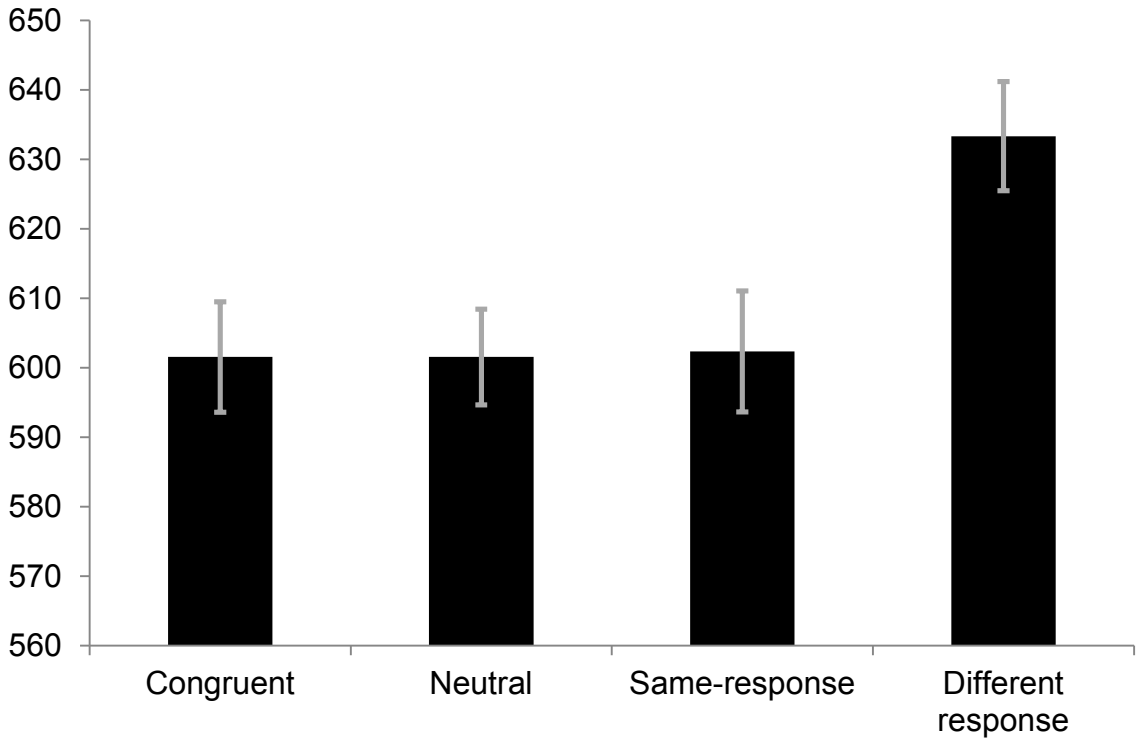

Figure 2: Mean RTs (in ms) for each condition in Experiment 2A. Error bars represent standard errors.

Although not one of the stated critical comparisons, the large apparent effect of contingency on congruent trial RTs was surprising enough to motivate a comparison between the congruent and neutral trials. It was stated that faster RTs on congruent vs. neutral trials would be attributed to facilitation that remains after contingency is controlled, but there was no statistical difference between the congruent and neutral trial RTs ($p > 0.05$) in this study. We modeled the predictions of the theory of a difference with a uniform between 0 and 30 ms, i.e.,

any effect was as plausible as any other in the full range (encompassing the 15–27 ms range suggested by the previous work alluded to above). The difference between the congruent (*M* = 601.54 ms, *SE* = 15.90) and neutral (*M* = 601.55 ms, *SE* = 13.75) conditions showed a Bayes Factor of 0.29. This result suggests that once contingency is controlled there remains no facilitation effect when using a non-color word neutral trial as the baseline. As far as we are aware, this is the first report of this finding, and one that suggests that debates over the mechanisms behind facilitation (MacLeod & MacDonald, 2000; Kane & Engle, 2003; Brown, 2011; Roelofs, 2010) should first consider contingency.

Importantly however, this result also serves another purpose, helping us to interpret the null difference between same-response and congruent trials. This will be discussed later.

The error rates for the congruent, neutral, same-response and different-response were 4.6, 4.3, 3.2, and 4.7% respectively. Analysis of the error rates showed a non-significant difference in the omnibus one-way ANOVA [$F(3,105) = 2.40$, $p = 0.072$, $r = 0.15$].

Experiment 2B

Using the same exclusion criteria as the other two experiments, the proportion of valid responses in this experiment was 94.47%. The repeated measures one-way ANOVA containing all four conditions was statistically significant [$F(3,105) = 7.71$, $p < .001$, $r = 0.26$; see Figure 3]. To test critical comparison 2, a pairwise comparison was made between the RTs of same-response (*M* = 606.21 ms, *SE* = 16.36) and non-response set (*M* = 632.48ms, *SE* = 16.39) trials. The difference was statistically significant [$t(35) = 3.49$, $p = .001$, $r = 0.51$]. This indicated that the non-response set condition had slower RTs than the same-response condition and is supportive of the notion that same-response trials involve response facilitation.

However, the pattern of RTs observed encouraged the comparison of the different-response (*M* = 618.83 ms, *SE* = 14.25) and non-response set trials; a comparison which yielded a non-significant difference [*t*(35) = 1.74, *p* = 0.091, *r* = 0.28]. Slower (but statistically non-significant) RTs to non-response set trials compared to different-response trials was unexpected and makes the difference between same-response and non-response trials difficult to interpret.

Since neutral and same-response trials were used in this block, we compared RTs to these trials to see if the same pattern of results from critical comparison 1 of Experiment 2A would be replicated. Using the same criteria employed to calculate the Bayes Factor in Experiment 2A, the non-significant [*t*(35) = 1.07, *p* = .294, *r* = 0.18] difference between the two conditions returned a Bayes Factor of 0.58 a value that cannot be taken as evidence for nor against the theory under test (Dienes, 2011) and is therefore not considered further.

The error rates for the neutral, same-response, different-response and non-response trials were 5.8, 4.7, 7.2 and 3.9% respectively. Analysis of the error rates showed a significant difference in the omnibus one-way ANOVA [*F*(3,105) = 3.40, *p* = .021, *r* = 0.18]. Post hoc pairwise comparisons between the conditions yielded a significant difference between different-response and non-response trials [*t*(35) = 3.31, *p* = .012, *r* = 0.49] while the other comparisons were non-significant (*p*s > 0.05). The error rate for different-response trials in the present experiment is much higher than in Experiment 2A [*t*(35) = 2.03, *p* = .050, *r* = 0.33], but was only statistically different from the non-response set trials which is largely consistent with the previous block in that errors were no different between different-response, same-response and neutral trials. This is discussed further below.

It is also possible that the introduction of non-response trials influence participants' approach to different-response trials in Experiment 2B since the number of

incongruent trials increases. Pairwise comparisons between the RTs and error rates of different-response trials in the two experiments were run. The results were inconclusive as although the error rates in Experiment 2B were higher the RTs were non-significantly different [$t(35) = 1.56$, $p = .125$, $r = 0.25$].

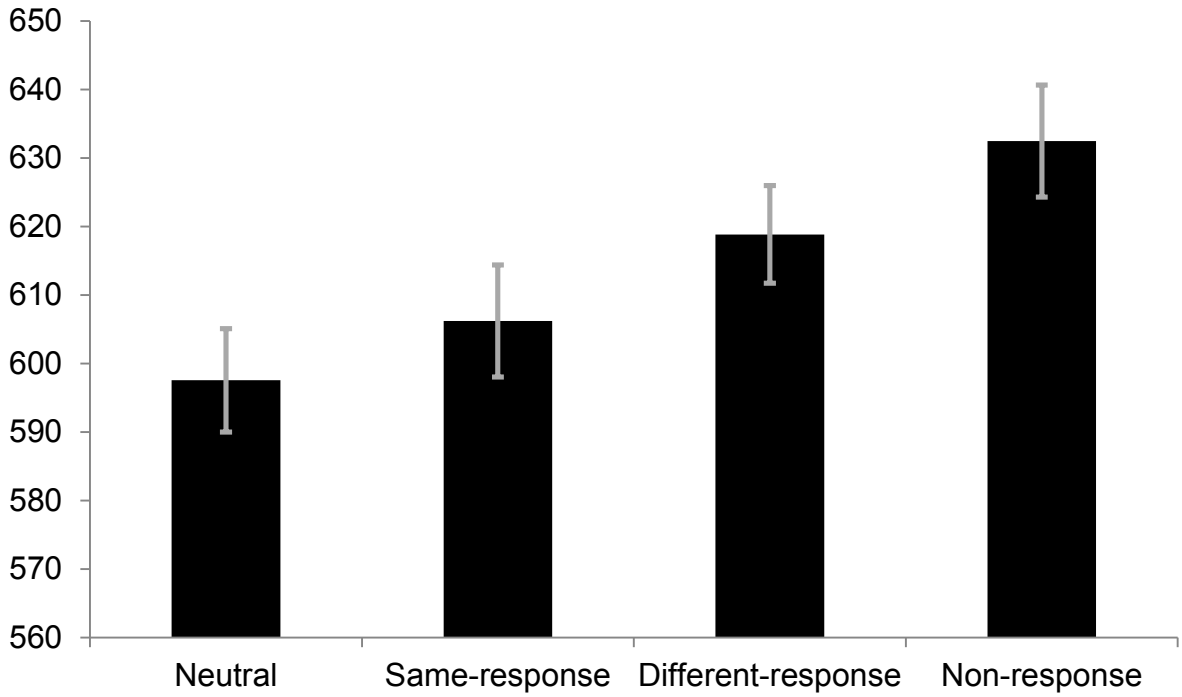

Figure 3: Mean RTs (in ms) for each condition in Experiment 2B. Error bars represent standard errors.

Discussion

The goal of the present study was to assess the utility of the two-to-one mapping manipulation and the nature of same-response incongruent trials in the Stroop task. This was assessed by comparing them to non-color word neutral trials and nonresponse set trials while controlling for response contingency. The key result is the finding of strong (Bayesian) evidence for no statistical difference between same-response and non-color word neutral trials. As stated earlier, two possible scenarios could be the cause of this: either same-response trials involve both response facilitation and semantic category competition, with the two effects canceling each other out, or the more parsimonious explanation that same-

response trials do not involve either effect. Although this result does not allow us to draw conclusions about the mechanisms involved in same-response trials, it shows clearly that same-response trials do not permit a reliable measure of the presence or absence of semantic category conflict and therefore all future studies using the two-to-one mapping paradigm should include a neutral baseline.

Same-response incongruent trials were also compared to nonresponse set trials. Following the assumptions of the two-to-one paradigm, these trials are thought to involve semantic category conflict and not response conflict, just like same-response trials, but in contrast to same-response trials are unlikely to involve response facilitation. We found that non-response set trials were responded to more slowly than same-response trials. This result suggests that RTs to same-response trials are at least partially determined by response facilitation. In light of these results, the significance of a series of recent studies might have to be reassessed (Schmidt & Cheesman, 2005; van Veen & Carter, 2005; Wendt et al., 2007; Steinhauser & Hubner, 2009; Chen et al., 2011, 2013; Berggren & Derakshan, 2014).

However, the longer RT to non-response set trials has to be interpreted with caution since we also observed unexpected results when comparing non-response to different-response trials. RTs to non-response set trials were not different from those to different-response trials, which was not in line with predictions based on previous research. However, recent work in our lab shows that the putative response set effect (different-response trials – nonresponse set trials) is strongly modulated by trial type mixing and is thus not as reliable as previously thought. Hasshim and Parris (submitted) have shown that the response set effect is much larger when different-response and non-response set trials are presented in different, pure blocks. When presented in mixed blocks the response set effect

was substantially reduced; an effect that resulted from a substantial decrease in RT to different-response trials, while no other trial type was affected. Thus, since the present results mirror effects observed in Hasshim and Parris, it is likely that trial type mixing employed here is responsible for the lack of the expected response set effect. Moreover, this means that the RTs observed to the non-response set trials are reliable. Indeed a few studies have reported no difference between non-response and different-response trials under similar mixed conditions (but slightly different presentation formats; e.g., Stirling, 1979; Sugg & McDonald, 1994; Milham et al., 2001). However, the error data from Experiment 2B bear consideration at this point. While the number of errors did not differ from those in the neutral or same-response condition, there were significantly fewer errors in the non-response set condition than in the different-response condition. Assuming that the error trials are the trials on which participants experienced the most difficulty, removing those trials means you are potentially removing the trials that would have increased the overall average RT for the different-response condition, rendering them significantly longer than those to nonresponse trials and hence revealing the expected response set effect. Nevertheless, this would not have altered the RTs to non-response set trials. If anything the RTs to non-response set trials are lower than they would have been had the more difficult trials been included. In sum, the results from Hasshim and Parris permit us to conclude that the finding of shorter RTs to same-response trials than to non-response set trials is best interpreted as supporting the notion that same-response trials involve some form of facilitation.

While the present results are incompatible with multi-stage models of Stroop interference (Klein, 1964; Zhang & Kornblum, 1998; Zhang et al., 1999; De Houwer, 2003; Schmidt & Cheesman, 2005), some such models would predict that

no difference should be expected between same-response and neutral trials when participants respond manually because manual responses (with color patches) do not have access to semantics (Glaser & Glaser, 1989; Sugg & McDonald, 1994; Sharma & McKenna, 1998). Given the use of a manual response with color patches in the present study our data are compatible with such models. However, it is clearly not possible to have same-response trials when using a vocal response, thus we restrict our interpretation to models whose predictions are not modified by response modality.

In the present study we also controlled for response contingency effects to ensure that such effects were not contributing to the RTs on congruent and same-response trials. One surprising effect of controlling for response contingency was the lack of Stroop facilitation effects (neutral-congruent RTs) when we had observed Stroop facilitation when contingency was not controlled in Experiment 1. The mechanism behind Stroop facilitation effects is debated (MacLeod & MacDonald, 2000; Kane & Engle, 2003; Roelofs, 2010; Brown, 2011). Our study was not designed to make this comparison, but we are not aware of any other study that has made a comparison between neutral and congruent trials when contingency is, and is not, controlled. A future study designed explicitly to test for effects of contingency would benefit from a within-subjects comparison to investigate whether, once contingency is controlled, the resulting increase in RTs to congruent trials leaves no facilitation effects to be explained.

A further effect of controlling for contingency is that, in the present data set at least, there was no difference between same-response and congruent trials suggesting that any difference between these two trial types is largely driven by response contingency and not semantic category conflict. More could be made of this result had previous studies not observed a semantic category conflict effect

even after controlling for contingency (De Houwer, 2003; Schmidt & Cheesman, 2005). The present result then could be interpreted as showing no effect of semantic category conflict due to unusually fast responses on same-response trials; that is there is no difference between same-response and congruent trials (and neutral trials) because for whatever reason, semantic category conflict was absent from Experiment 2 of the present study. However, it is not clear why semantic category conflict would be absent in Experiment 2 but not Experiment 1. Furthermore, the RTs to same-response trials in Experiment 1 and 2 are very similar (around 600 ms). Controlling for contingency was predicted to increase RTs to congruent trials and indeed RTs to congruent trials increased by around 30 ms when contingency was controlled. In short, despite contrasting with previous results showing an effect of semantic category conflict when contingency is controlled, the null difference between congruent and same-response trials is most likely an outcome of an increase in RTs to congruent trials brought about by contingency. Notably, congruent trial RTs are also not different from neutral trial RTs which in turn are not different from same-response trial RTs. With the predicted effect of contingency and a neutral word baseline that does not involve semantic or response conflict the results are best interpreted as showing that RTs to same-response trials cannot be used reliably to determine the presence or absence of semantic category conflict. All future studies should include a neutral non-color word baseline when utilizing the two-to-one mapping paradigm.

Since we had removed the effects of response contingency from Experiment 2 we can be confident that the difference observed between the same-response and different-response trials is not overestimated. Indeed, a raw effect size of roughly 30 ms seems to be a common magnitude of difference between these two trial types whether contingency is controlled or not. However, as

mentioned earlier the utility of same-response trials in such a comparison is questioned by the present results given they are not reliably different from neutral trials. In essence, our results suggest that the difference between different-response and same-response trials in terms of RTs is the same as the difference between different-response and neutral trials, meaning that it is a measure of Stroop interference and not a purer measure of response conflict as has previously been assumed (De Houwer, 2003; Schmidt & Cheesman, 2005; van Veen & Carter, 2005; Steinhauser & Hubner, 2009). The analyses on error rates also do not clearly explicate the differences between the different conditions although the trend does suggest a higher error rate for different-response trials generally, which is to be expected. Previous studies using the Stroop task typically do not focus on error rates because the relatively easy task keeps speed-accuracy trade-off to a minimum. Thus the analyses on RTs are the main focus of this paper as well.

The sample size of the present study was selected to match that of Schmidt and Cheesman (2005). However, Schmidt and Cheesman do not report the gender of their participants and so it was not possible to establish whether our participants differed from theirs in that respect. While unlikely it is possible that the differences between our study and theirs (i.e., the effect of contingency on the difference between same-response and congruent trials) were a consequence of the gender differences in the present study. However, we have no reason to assume that gender would influence the present results. Nevertheless, future studies should consider testing equal numbers of male and female participants to eliminate this as a possible account of findings observed.

In conclusion, same-response trials cannot be used to determine the presence or absence of semantic category conflict, at least until the mechanisms

contributing to RTs are better understood. Nor can they be used to index a purer measure of response conflict. Notably, the lack of difference between same-response and neutral trials does not necessarily mean that the two trial types are processed in a similar way. For example, van Veen and Carter (2005) have shown that different brain regions are activated by same-response and different-response trials when both are compared to congruent trials. While our data suggest that any differences observed in previous studies between same-response and congruent trials is likely just greater semantic/response facilitation effects on the latter, it is possible that the competing influences of response facilitation and semantic conflict interact to influence response latency. Sometimes one might win over the other, producing evidence for conflict or facilitation, but until it is known how latency is modulated by each, or even that it actually occurs, RTs to same-response trials must be interpreted with caution. The inability to differentiate neutral and same-response trials is important and reason enough to doubt the latters usefulness in measuring semantic category conflict. Our results show that non-response set trials are potentially a better alternative.

Experiment 3

Experiment 3 of the previous chapter compared the response set effect in mixed vs. pure contexts using the two-to-one colour response mapping paradigm but is subject to the same criticisms of contingency and non-counterbalanced colours raised against the earlier experiments in the same chapter. Here a new, methodologically improved version of that experiment (Experiment 3) is reported.

As with Experiment 3 of Chapter 2, the goal of this experiment was to elucidate the unexpected findings from Experiment 2b in Chapters 2 and 3 where response set effects, a measure of response competition, was found to be non-significant. The aim of this experiment was to replicate the findings of Experiment 2b to rule out the possibility that it was an anomalous result and thus to investigate whether the lack of a response set effect in Expeirment 2b was due to the mixed presentation format as detailed above in the previous chapter. This was done by comparing the magnitude of the response set effects between the mixed and pure block presentation versions of the task.

Replicating the results from Experiment 2b will also permit a further assessment of same-response trials as a measure of semantic conflict to augment those from Chapter 2. As argued in the earlier chapters, same-response trials are potentially problematic as a baseline as they may involve response facilitation, which would exaggerate effects attributed to semantic conflict. This point is further explored in Chapter 4.

In this experiment, four types of trials were presented in pure blocks or in mixed blocks. The four types of trials involved were: 1) neutral word trials; 2) same-response trials, where the incongruous word spells out a colour that is mapped on to the same response button as the correct response; 3) non-response set trials, and; 4) response set (different-response) trials. The goals for this

experiment were as follows: 1) Replicate the mixing effect from Experiment 1; 2) Investigate whether non-response set trials leads to interference compared to a neutral word baseline; 3) Show that the lack of a response set effect in the previous studies in Chapters 2 and the present chapter were due to trial type mixing.

Method

Participants

34 participants (17 male) were recruited from the undergraduate population in exchange for course credit. They had an average age of 20.5 (*SD* = 3.18).

Design

A 4(condition: neutral, same-response, different-response, & non-response) x 2 (block type: pure blocks & mixed blocks) repeated measures design was used.

Apparatus and Materials

In this experiment the two-to-one response mapping version of the Stroop paradigm (De Houwer, 2003) was used. In this version, four colours are possible responses but only two buttons are used as response keys since two colours are mapped onto each button. The four types of trials involved were neutral word trials that were not associated with any colour, same-response trials, where the incongruous word spells out a colour that is mapped on to the same response button as the correct response, non-response set trials and different-response trials.

As with the previous experiments, two versions of the task were administered; counterbalanced across participants. Following the two-to-one mapping paradigm pairs of colours were mapped onto each response key. The colour pairs used were, orange (255; 127; 0) and blue (0; 112; 192); and pink (255; 20; 147) and white (255; 255; 255) for one version, purple (204; 0; 255) and

yellow (255; 255; 0), and green (0; 255; 0) and red (255; 0; 0) on the other. The same neutral words from Experiment 1 were used.

Procedure

The sequence of each trial was the same as that of Experiment 1, as was the way the order of the blocks were counterbalanced. The keys used for responses were the 'z' and '/' keys of the keyboard with the colour pairs assigned to them counterbalanced across participants. Participants went through a practice block of 72 trials, which consisted of trials from all conditions appearing in random order. For the experimental procedure 120 trials of each of the four conditions were presented for each block type. This meant that each participant went through 960 experimental trials in total. Participants went through all the pure blocks either before or after all of the mixed blocks.

Results

Using the same exclusion criteria as before, the proportion of valid responses for all participants were .945 (*SD* = .043) for the mixed blocks and .944 (*SD* = .043) for the pure blocks. The mean latencies of each condition are reflected in Table 1.

An omnibus 4 x 2 repeated measures ANOVA showed a significant trial type (neutral, same-response, non-response set or response set) by presentation format (pure or mixed) interaction, $F(3,99) = 7.872$, $p < .001$, $r = .271$. The follow-up repeated measures one-way ANOVAs measuring whether there is a difference across the neutral, same-response, non-response set, and response set conditions was significant for both mixed and pure blocks ($F(3 ,99) = 8.774$, $p < .001$, $r = .285$ and $F(3 ,99) = 16.624$, $p < .001$, $r = .379$, respectively).

To further study the impact of presentation format on non-response set and response set (different-response) trials, two 3 x 2 ANOVAs were conducted, which

Table 1: Mean RT in ms (SEs) of all conditions in all conditions of Experiment 3

	Mixed	Pure
Neutral	585.98 (14.29)	570.20 (16.26)
Same-response	579.43 (14.12)	577.29 (15.29)
Non-response	606.02 (17.98)	593.41 (18.21)
Response set	603.91 (13.59)	636.75 (19.87)
Non-response set effect (Non-response – Neutral)	20.04	23.21
Response set effect (Response set – Non-response)	-2.11	43.34

was similar to the omnibus test except that in each, either the non-response set or response set conditions was removed. This was done to determine whether the interaction would still be significant in each case and if not, it would suggest that the omitted condition was the main source of the interaction. Results showed that in the analysis without the response set (different-response) trials, the interaction was non-significant ($F(2,66) = 0.981$, $p = .380$, $r = .121$) but significant when the non-response set trials were omitted ($F(2,66) = 9.098$, $p < .001$, $r = .348$). Thus, trial type mixing appears to mainly affect response set (different-response) trials.

Two 2 x 2 ANOVAs were conducted to investigate the effect of presentation type on non-response and response set effects. The analysis on non-response set effects showed the interaction to be non-significant ($F(1,33) = 0.90$, $p = .766$, $r = .163$), while the interaction was statistically significant for response set effects ($F(1,33) = 15.69$, $p < .001$, $r = .568$) indicating that presentation type affected response set effects but not non-response set effects.

Pairwise comparisons showed that the difference between the response set and non-response set conditions was non-significant in mixed blocks ($t(33) = -0.29$, $p = .775$, $r = .050$), but significant when administered in pure blocks ($t(33) = 4.47$, $p < .001$, $r = .614$). In pure block presentation, response set (different-response) trials (636.75 ms) were slower than non-response set trials (593.41ms). To further explore the nature of this difference the non-response set RTs in the pure and mixed blocks were compared which revealed no difference where $t(33) = 1.16$, $p = .253$, $r = 1.198$. The same comparison in the response set condition was significant ($t(33) = 2.60$, $p = .014$, $r = .412$), which was the result of faster RTs in the mixed blocks (603.91ms). To measure non-response set effects in each presentation format, pairwise comparisons between non-response set and neutral trials were conducted. Non-response set trials were slower than neutral trials in both pure ($t(33) = 2.93$, $p = .006$, $r = .454$) and mixed blocks ($t(33) = 2.61$, $p = .014$, $r = .414$).

The comparison between non-response and same-response trials showed the former to be slower in mixed blocks ($t(33) = 3.718$, $p = .001$, $r = .543$) but the difference was non-significant in pure blocks ($t(33) = 1.446$, $p = .158$, $r = .244$).

Error analysis

The 2x4 repeated measures ANOVA on the number of errors committed showed a non-significant interaction $F(3,99) = 1.39$, $p = .251$, $r = .118$. The main effect of condition was statistically significant ($F(3,99) = 10.02$, $p < .001$, $r = .303$) while the main effect of block type was not ($F(1,33) = 0.80$, $p = .378$, $r = .090$).

Discussion

The results of Experiment 3 revealed that trial type mixing modifies the magnitude of the response set effect; an effect that was again driven by faster RTs to response set (different-response) trials in the mixed block condition. RTs to non-

response set and response set (different-response) trials were non-significantly different in mixed blocks indicating that the response set effect was statistically eliminated.

Compared to a neutral baseline non-response set trials produce an interference effect in both pure and mixed blocks despite there being no opportunity for response-level conflict. This confirms that non-response set interference likely results from semantic conflict. The magnitude of non-response set effect did not differ across the two block types in this experiment indicating that the effects of trial type mixing were on response set membership effects (response conflict).

The results from the present study support the notion that the lack of a response set effect in the experiments in Chapter 2 and the present chapter were due to the use of mixed blocks. The results show that non-response set trials are a better index of semantic processing than same-response trials and support the findings from the earlier experiments that show that same-response trials do not differ from a neutral baseline condition. This suggests that same-response trials are either treated as neutral conditions or involve both response facilitation (since both dimensions of the stimulus provide evidence towards the same response) and semantic conflict. Unlike the results from the mixed block presentation, the 16ms advantage of same-response trials on non-response trials in pure blocks was statistically non-significant, which does not support the idea that response facilitation influences the performance of the former. Nevertheless, non-response trials might still be a better alternative to index semantic conflict since their mean RT is dissociable from that of neutral trials, which is a necessary measure of semantic conflict in the two-to-one paradigm.

The effect of presentation format (mixed vs pure blocks) on the response set effect is an important finding and one that deserves greater attention in its own chapter. To that end, the effect of trial type mixing on the response set effect is further explored in Chapter 5 of this thesis where it will be explored using the more common one-to-one colour-response mapping. Chapter 4 describes a further and final attempt to elucidate the mechanisms involved on same-response trials.

Chapter 4: Assessing stimulus-stimulus (semantic) conflict in the Stroop task using saccadic two-to-one color response mapping and preresponse pupillary measures

The experiment in this chapter was designed to further evaluate same-response trials. Evidence that same-response trials are not a reliable measure of semantic interference effects was presented in the studies in Chapters 2 and 3. The aim of this chapter was to use eye-tracking measures as an alternative, potentially more sensitive, index of conflict in the Stroop task and which, importantly, also permitted a pupillometric index of effort. This permitted the further assessment of whether conflicts experienced on same-response and neutral trial types are dissociable. This chapter in its entirety has been published as an article in the journal *Attention, Perception, & Psychophysics*.

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.

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

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 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 non-significant 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.

Pre-response 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 (*SD* = 4.61). Data from 5 other participants were excluded from the analyses as an 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 120Hz. The

movement of only one eye was recorded using an Eyelink 1000 (SR Research Ltd.) recording pupil and corneal reflection, sampling at 500Hz (every 2ms). Participants went through a 9-point calibration and validation before the start of each block. Eye movement and pupillometric parameters were extracted off line using Data Viewer (SR Research Ltd.).

During the task, participants placed their head and chin on a headrest positioned 60cm from the screen. 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 200x200 pixels in size appeared on 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 spelt out the corresponding color it was presented in while on neutral trials, the word was a non-color related word. On same-response trials, the word spelt out an incongruent color, which shared the same response location as the relevant color dimension, while in different-response trials, the incongruent color word always referred to a color whose response location was on the opposite side to 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 on the top corners of the screen. Participants were asked to move their gaze towards the square that corresponds to the correct response of the stimulus and to do so as quickly and accurately as possible. Once a fixation of 100ms 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 hash symbols (#) of 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 were subsequently 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 and broken down in to 3 blocks of 80 trials each. The number of different-response trials was double the other conditions 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 eye-tracking software and recorded in pixels. After each participant completed the task, a single measurement of a 4mm 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 when an initial saccade of >5 degrees was made were used

in the analyses. Trials where the initial saccade was not within 45 degrees towards the correct square were classified as invalid and were not included in the analyses along with trials where the time taken to make the initial large saccade was <200ms or >2500ms. Incorrect trials were defined as those where the initial saccade was made within 45 degrees towards the incorrect square and 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 and was found to be statistically significant ($F(3,96) = 3.29$, $p = .024$, $r = 0.18$) and pairwise comparisons revealed that incongruent trials had more incorrect trials than same-response trials $t(32) = 3.11$, $p = .004$, $r = 0.48$. The other pairwise comparisons were statistically non-significant (congruent vs. neutral: $t(32) = 0.223$, $p = .825$, $r = 0.04$; congruent vs. same-response: $t(32) = -1.39$, $p = .173$, $r = 0.24$; congruent vs. incongruent: $t(32) = 1.73$, $p = .093$, $r = 0.29$; neutral vs. same-response: $t(32) = -1.51$, $p = .140$, $r = 0.26$; neutral vs. incongruent: $t(32) = 1.59$, $p = .123$, $r = 0.27$).

Saccadic latencies

The mean RTs of valid saccades for congruent, neutral, same-response, and incongruent trials were 437.55ms, 460.53ms, 462.10ms, and 478.79ms. A one-way repeated measures ANOVA was conducted and was found to be statistically significant ($F(3,96) = 14.37$, $p < .001$, $r = 0.36$).

Pairwise comparisons revealed that congruent trials had the fastest RTs (vs. neutral: $t(32) = 3.48$, $p = .001$, $r = 0.52$; vs. same-response: $t(32) = 3.92$, $p < .001$, $r = 0.57$; vs. incongruent: $t(32) = 6.95$, $p < .001$, $r = 0.78$) while incongruent trials had the slowest RTs (vs. neutral: $t(32) = 2.55$, $p = .016$, $r = 0.41$; vs. same-response: $t(32) = 2.78$, $p = .009$ $r = 0.44$.

The difference between the RTs of neutral and same-response trials was non-significant ($t(32) = 0.27$, $p = .789$, $r = 0.048$). 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 (http://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 while one that is larger than 3.0 indicate support for the alternate hypothesis. Since we are 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 6-45ms for an effect with an assumed uniform distribution (all values were equally likely), the Bayes factor returned a value of 0.09, indicating strong support for 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 size during each trial up to the first saccade were obtained and separately analyzed. Table 1 shows the average maximum and minimum pupil diameter, the time after stimuli onset 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 4.191mm ($SE = 0.055$), which indicates that there was a small initial dilation in pupil size, followed by a large constriction.

Table 1: Average (SE) maximum and minimum pupil sizes for each condition up to response, along with the average time they occurred after stimuli onset.

	Maximum Diameter		Minimum Diameter		Saccadic RT
Condition	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)

Maximum pupil diameter

The mean maximum pupil diameter in the congruent, neutral, same-response, and incongruent trials were 4.204mm, 4.204mm, 4.203mm, and 4.202mm respectively. The repeated measures one-way ANOVA for pupil diameter and the latency at which it occurred were non-significant ($F(3,96) = 0.017$, $p = .997$, $r = 0.013$ and $F(3,96) = 0.646$, $p = .588$, $r = 0.082$, respectively).

Average pupil diameter

The average pupil diameter in the congruent, neutral, same-response, and incongruent trials were 4.138mm, 4.132mm, 4.127mm, and 4.123mm respectively. The repeated measures one-way ANOVA was non-significant ($F(3,96) = 1.586$, $p = .198$, $r = 0.127$).

Minimum pupil diameter

Minimum pupil diameter occurred at 316.92ms, 353.52ms, 344.38ms, and 355.41ms after target onset. A repeated measures one-way ANOVA for the latencies was significant $F(3,96) = 13.69$, $p < .001$, $r = 0.353$) and follow up analyses revealed the latency to congruent trials to be faster than neutral ($t(32) = 6.06$, $p < .001$, $r = 0.73$), same-response ($t(32) = 3.27$, $p = .003$, $r = 0.50$), and incongruent trials ($t(32) = 5.66$, $p < .001$, $r = 0.71$); while the other three conditions were non-significantly different from each other (same-response vs. neutral: $t(32) = -1.23$, $p = .230$, $r = 0.21$, incongruent vs. neutral: $t(32) = 0.34$, $p = .740$, $r = 0.06$, and same-response vs. incongruent: $t(32) = 1.85$, $p = .074$, $r = 0.31$).

The mean minimum pupil diameter in the congruent, neutral, same-response, and incongruent trials were 1.879mm, 1.925mm, 1.929mm, and 1.954mm respectively, indicating that the pupil constricted to a size smaller than at target onset. The repeated measures one-way ANOVA was significant ($F(3,96) = 15.162$, $p < .001$, $r = 0.37$). Pairwise comparisons showed that congruent trials had the smallest minimum size (vs. neutral: $t(32) = 3.91$, $p < .001$, $r = 0.57$; vs. same-response: $t(32) = 3.80$, $p = .001$, $r = 0.56$; vs. incongruent: $t(32) = 6.68$, $p < .001$, $r = 0.76$) while incongruent trials had the largest (vs. neutral: $t(32) = 2.50$, $p = .018$, $r = 0.40$; vs. same-response: $t(32) = 2.95$, $p = .006$ $r = 0.46$). The difference between the minimum pupil sizes of neutral and same-response trials was non-significant ($t(32) = 0.36$, $p = .720$, $r = 0.064$). As with RTs, a Bayes factor was calculated to determine if there is evidence for no difference between the two conditions. Since there are no prior findings on such an effect using minimum pupil size, the only reference to the size of the effect is either the difference between neutral and congruent or incongruent and neutral trials. The larger of the differences, 0.045mm, was used as the upper bound while the lower bound was the proportionate equivalent to the one used in Hasshim and Parris (2014),

0.006mm. 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 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 non-significantly 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 non-correspondence 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

[1]We thank an anonymous reviewer for this suggestion

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.

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 non-significantly 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.

Chapter 5: The Makeup Of Stroop Interference Depends on Context: Trial Type Mixing Substantially Reduces the Response Set Effect

The three experiments in this chapter were designed to evaluate the response set effect, which like same-response trials has been used to dissociate semantic and response conflict in the Stroop task. Among the findings reported in Chapters 2 and 3, one stood out as being particularly important: Experiment 3 from both chapters showed in the two-to-one response mapping paradigm, the magnitude of the response set effect is substantially affected by whether the trial types are presented in mixed or pure blocks. The motivation for the experiments in this chapter is to investigate whether the effects from the two-to-one paradigm utilised in the previous chapters can be generalised to the more common one-to-one response-mapping task. Thus the experiments reported in this chapter utilises a one-to-one mapping of colour-button.

The first two experiments reported in the present chapter are presented in manuscript form; a manuscript that is currently under review. Experiment 1 reports an investigation of the effect of trial type mixing (i.e. presenting trials in pure vs mixed blocks) on the response set effect using the more standard one-to-one colour-response mapping. The second experiment reports an investigation into how the trial type mixing effect might be operating, giving an insight into how response set effects might be established. A final more recent experiment, which was not part of the submitted manuscript, was also conducted to further test the theory posed in Experiment 2 of this chapter.

Abstract

The Stroop interference effect is thought to mainly comprise of response competition. This is demonstrated by the *response set effect,* which refers to the finding that an irrelevant incongruent colour-word produces greater interference when it is one of the response options, compared to when it is not. Despite being a key effect for models of selective attention, the magnitude of the effect varies considerably across studies. The present study tested the hypothesis that the presentation format (trials from each condition presented in separate blocks vs blocks containing trials from all conditions presented randomly) modulates the magnitude of the response set effect. We show that when each trial type is presented in its own block (pure), the response set effect is substantially larger compared to when blocks contain trial types from all conditions (mixed). A follow-up experiment manipulated the number of colour-words that make up the non-response set of distractors and showed that this modulated the size of the previously demonstrated mixing effect. These results show that 1) contrary to predictions of extant theoretical models, interference from colour-words that are not part of the response set is independent of the interference from colour-words that are; and 2) the magnitude of the response set effect is influenced by the number of active task-irrelevant colour concepts, which suggests that response competition is not the result of strategic selective attentional processes, but rather the result of learning biases brought about by task design. The results are discussed in terms of their implications for research debating the automaticity of reading.

The Make Up Of Stroop Interference Depends on Context: Trial Type Mixing Substantially Reduces the Response Set Effect

Selective attention refers to the process of selecting only relevant and important parts of the perceptual landscape at the cost of less relevant or irrelevant parts. Selective attention makes it possible to overcome behaviours that are innate or have become automatic through continued practice, and instead perform behaviours that are appropriate to a specific situation (Diamond, 2013). Experimentally, selective attention is typically measured using executive control tasks, which elicit cognitive conflict by presenting multiple sources of information that can be relevant or irrelevant to the performance of the task. To facilitate goal-oriented behaviour, mechanisms of selective attention appear to increase activation of goal-salient concepts. This is demonstrated by the *response set effect,* which refers to the well-established finding that items in the response set (task-relevant items) are important and are harder to ignore when in an irrelevant dimension.

The Stroop task requires participants to name the colour of the font in which a word is printed while ignoring the meaning of the word itself. The Stroop effect refers to the finding that naming the colour that a word is printed in takes longer when the word spells out a different colour (e.g. the word ‘red’ displayed in blue ink; an *incongruent* trial) compared to when the word spells out the same colour (e.g. the word ‘red’ displayed in red ink; a *congruent* trial) or when the word spells out a neutral word (one that is not associated with any colour, e.g. ‘table’) (see MacLeod, 1991; 2005 for comprehensive reviews of the Stroop task). In the context of the Stroop task the response set effect refers to the well-established finding that greater interference occurs when the incongruent colour word is a possible response option (part of the response set) compared to when it is not

(referred to as *non-response trials,* e.g. the word ‘orange’ in blue, when the colour orange is not a possible response colour; e.g. Klein, 1964; Milham et al., 2001; Risko, Schmidt & Besner, 2006; Sharma & McKenna, 1998; also see MacLeod, 1991, for a review).

On the other hand, non-response trials have been shown to display interference compared to a neutral non-colour related word or a congruent trial (e.g. Klein, 1964; and Sharma & McKenna, 1998), and this has been attributed to irrelevant non-response word belonging to the same semantic category as the eligible responses (i.e. colours) and is thus interpreted as indexing semantic conflict. This concurs with evidence showing that interference can occur independently at different levels of processing such as earlier stimulus encoding and lexico-semantic processing stages (e.g. Goldfarb & Henik, 2007; Hock & Egeth, 1970; Luo, 1999; Parris, 2014).

In his review of the Stroop effect, MacLeod (1991) identified the response set effect as one of 18 well-established findings for which models of the effect need to account, while two prominent models of the Stroop task (Cohen, Dunbar, & McClelland, 1990 and WEAVER++; Roelofs, 2000), have accounted for response set effects by proposing that attention is selectively allocated to the restricted set of eligible colours. This ensures that their activation levels are greater than those to colours not in the response set. Thus, when an eligible colour concept is denoted in the irrelevant dimension, it would be harder to ignore and lead to greater interference. In the Cohen et al. (1990) model, the eligible colour concepts are identified by task demand units where a bias is set such that those particular colours are more likely to guide attention. However, there is no description of the specific processes involved in establishing the colours as response set colours.

In the WEAVER++ model, the nodes of response set colours are flagged as goal concepts, which allows for subsequent selection and processing of information gleaned from a stimulus. Colours that are not part of the response set are not flagged and thus are less likely to be processed as a potential response or interfere with response selection (although see Caramazza & Costa, 2000; 2001 for evidence against the flagging component). Non-response set trials interfere only through their connections to the flagged response set nodes in the conceptual network. Given this connection, any manipulation that affects performance of incongruent trials would indirectly affect the performance of non-response set trials in tandem, but likely to a smaller degree since second-order activations would be smaller due to being further along the activation pathway. Similar to the Cohen et al. model, there is no description of the development of this process although Roelofs (2001) stated that achieving response set status would likely require repetition to achieve response-level salience.

Lamers, Roelofs, and Rabeling-Keus (2010) tested competing accounts of response set effects, with one account held by Roelofs (2003) and Cohen et al., (1990) arguing that response set effects arise due to the selective allocation of attention to eligible responses. They contrast this account with one based on greater inhibition of non-response set colours. In one experiment, they manipulated response set membership on a trial-by-trial basis by cuing the possible responses before each trial. They also manipulated response set size, reasoning that doing so would make it more difficult to inhibit individual responses under the inhibition account. The results showed that response set effects were independent of response set size (additive effect). In their second experiment, the distractor colour was cued before each trial, which resulted in facilitation on both incongruent and congruent trials (they did not use non-response set trials in their

second experiment). They concluded that the facilitation on congruent trials was evidence that pre-exposure to the distractor does not result in greater inhibition. These findings were argued to be consistent with the selective allocation of attention account.

Potential contextual modulation of the response set effect

As stated earlier, response competition, as measured by the response set effect, is an important component of interference in the Stroop task, but a cursory review of the literature on studies reporting the use of non-response trials will indicate that the magnitude of the response set effect varies considerably from study to study, independent of response mode. Our reading of this literature is that the experiment's presentation type (whether trials are presented in a random, mixed order or in blocks containing each type of trial) is a possible moderator of the size of the response set effect (see Table 1 detailing these studies, their presentation type and measured response set effects). Studies that present trials in a mixed order seem to show much smaller response set effects compared to studies that present trials in pure blocks containing only one type of trial each. Given that the response set effect is the key index of response competition, this suggests that the contribution of response competition to Stroop interference varies by experimental context. This is of theoretical importance because prominent models of the Stroop task have heavily drawn from the results of early classic studies (see MacLeod, 1991) that favoured pure block presentation due to technological limitations (RTs of each block were recorded using a stopwatch), while presenting trials in random, mixed order has now become standard in the field. The models described earlier are unable to account for this apparent pattern because even though they account for the response set effect in different ways, they assume it is a fixed and natural

Table 1: Response-set effects from studies that have used non-response set trials

Study	Response-set Effect (ms)	Presentation Type	Response Type	Notes
Caramazzaand Costa. (2000)	-1*	Mixed	Vocal	Picture-word naming task. Each block mixed neutral (unrelated) and either response set or non-response set trials
Hasshim and Parris (2014)	-13.65	Mixed	Manual	Two-to-one response Stroop task, Non-significant
Klein (1964)	241**	Pure	Vocal	List method*** (not computerized)
La Heij (1988)	24	Mixed	Vocal	Used picture-word naming task
	12	Mixed	Vocal	Used picture-word naming task
Lamers et al. (2010)	11	Mixed	Vocal	Response membership established trial-by-trial
	19			
Milham et al.(2001)	6*	Mixed	Manual	Each block mixed neutral and either response set or non-response set trials
Proctor (1978)	111	Pure	Vocal	Experiment 1 - List method (not computerized)
	29.0	Mixed	Vocal	Experiment 2
	23.7	Mixed	Vocal	Experiment 3
Risko et al. (2006)	8	Mixed	Vocal	Used colour associates
	6	Mixed	Manual	Used colour associates
Scheibe et al. (1967)	205	Pure	Vocal	List method (not computerized)
Sharma and McKenna (1998)	96.7	Pure	Manual	
	63.6	Pure	Vocal	
Stirling (1979)	17	Mixed	Vocal	Non-Significant
	11	Mixed	Vocal	Non-Significant
West et al. (2004)	34	Mixed	Manual	Digit counting task
	12	Mixed	Manual	Digit counting task, Non-Significant

* Response set effect was calculated by the difference between the interference effects of the incongruent block and the non-response set block. Note that in Milham et al. the RTs to response set trials were slower than non-response set trials, the RTs of neutral trials in the latter block was faster as well.

** Response set effect was calculated by subtracting RTs of non-response set trials from incongruent trials. In cases where different types of non-response set trials were used, we chose the trial type that resembled standard non-response set trials the most.

*** The RTs for the list method experiments were calculated by dividing the overall time taken to go through the list, by the number of words in the list.

consequence of the selective attention process, and thus the effect should not be affected by experimental context.

Moreover, demonstrating that the semantic and response based components of Stroop interference can be manipulated would have implications for recent work discussing the uncontrollable nature of semantic processing in the Stroop task (e.g. Augustinova & Ferrand, 2014) as they work on the assumption that response conflict make up the bulk of Stroop interference. As such, these studies have argued that certain experimental manipulations that reduce Stroop interference affect mainly response conflict and not semantic conflict. However, given that all of these studies have employed random, mixed presentation of trial types it is possible that semantic conflict is in fact being reduced.

Mixing effects in related literatures

The difference between presenting different trial types in pure versus mixed blocks has been explored using other paradigms and in different literatures. Two general patterns of results have been reported, both describing the size of an effect being smaller in mixed blocks compared to pure blocks (referred to as *mixing effect)*. For example, Los (1996) reviewed possible strategic and stimulus-driven accounts that differentiate performance between the two presentation types. He described the effect of a *mixing cost*, which is a general slowing of responses in mixed blocks compared to pure blocks. In various studies using perception and memory tasks, the mixing cost was shown to be greater when the relationship between the stimulus and response is more compatible, which shows that it is not simply due to a general difference in task demand between the two presentation types.

Pertinent to the current study, one way to determine if a mixing cost can explain the difference in performance between our presentation formats is to

observe an asymmetry in mixing cost. This is when a relatively slow trial is not slowed down as much as a faster trial in mixed blocks, which would result in effects being smaller in mixed blocks compared to pure blocks and is the reason smaller effects are observed in the former (Los, 1996).

Another observation from studies presenting trials in mixed and pure blocks is the effect of *homogenisation* described by Lupker, Brown and Colombo (1997) and Lupker, Kinoshita, Coltheart and Taylor (2003) in research on word naming. Unlike the mixing cost described above, compared to pure blocks, the RTs of trials in mixed blocks tend to move towards the overall mean RT of the different trial types in the block (i.e. the slower trials become faster while faster trials become slower). This effect is driven by the averaging of the threshold for the decision making process towards the mean of the all trial types in each block (see Lupker et al., 2003 for a more comprehensive explanation), which results in the RT of the faster trials increasing while the RTs of slower trials decreasing in the mixed blocks.

We report two within-subjects experiments that tested the prediction that trial type mixing reduces the response set effect (and thus response competition) in the Stroop task. In the first experiment we compared the response set effect in mixed vs. pure blocks and show that the response set effect is indeed substantially reduced in the mixed block context. This result indicated that while the majority of Stroop interference is composed of response competition in pure blocks, in the mixed blocks interference is composed of roughly equal amounts of response and semantic competition. In the second experiment we tested a prediction regarding how the mixing effect might operate.

Experiment 1

The goal of Experiment 1 was to determine whether the response set effect is smaller when trials are presented in mixed blocks compared to compared to pure blocks as suggested by the observation in Table 1, and if so, whether the pattern of results is consistent with either account of mixing costs from the literature.

Method

Participants

40 participants (9 male) were recruited from the student population in exchange for course credit or £5. They had a mean age of 21.7 (*SD* = 4.38).

Design

A 3x2 within subjects, multifactorial design with the two independent variables being trial type (neutral, non-response set & response set) and presentation format (mixed & pure).

Apparatus and Materials

Stimuli were presented on a PC using Experiment Builder software (SR Research Ltd.) with responses recorded via pressing one of the assigned keys on a Cedrus response pad (RB 740, Cedrus Corporation). Three response keys were used with each key assigned one of the three possible colour responses. Participants were free to use fingers from either one or both hands to respond.

Stimuli

To control for possible effects of different colours being in the response and non-response set, participants went through one of two versions of the experiment where the non-response colours of one version served as the response-set colours of the other. The colours used were yellow (RGB: 255; 255; 0), pink (255; 20; 147) and green (0; 255; 0) in one version, and blue (0; 112; 192), purple (204;

0; 255) and orange (255; 127; 0) in the other. The words *wall, marvel* and *story* were used for the neutral trials and had been matched for frequency and length using the English Lexicon Project (Balota et al., 2007). All the stimuli were presented on a black background. Stimuli were presented in size 20, Courier New font, on a black background. Participants sat approximately 50cm away from the screen.

Because only two of the three response options are possible correct responses for different response trials (the third response button would correspond to a congruent trial, which are not involved in the experiment), the same limitation was imposed on each colour stimulus to ensure that regardless of trial type, each word stimulus had the same probability (50%) of its correct response being one of two response options. This was done by never pairing each word stimulus (neutral and colour word) to one specific colour each. The specific colour omitted was counterbalanced across the words in each trial type (e.g. the word *wall* never appeared in blue while *story* never appeared in green).

Procedure

At the start of each trial, participants were presented with a grey fixation cross in the centre of the screen for 500ms. This was followed by the Stroop stimulus, which remained at the centre of the screen until a response was executed. Participants were instructed to press the assigned key corresponding to the colour of the text as quickly and accurately as possible while ignoring the word's meaning. Upon committing an error, an additional auditory tone and a visual error message were presented. The error message lasted for 1500ms followed by a blank screen of 100ms.

Before the experimental trials participants went through a practice block of 60 trials made up of hash symbols (#) of three to six characters in length. For the experimental blocks, participants went through a total of 576 trials, made up of 96 trials of each trial type (neutral, non-response set and response set), presented in the two presentation formats (mixed and pure; i.e. 96 trials x 3 trial types x 2 presentation format). Thus the proportion of neutral, non-response set and response set (different-response) trials were equal throughout each version and presentation format.

During the experiment, trials were presented in blocks of 96 trials and the order of presentation format presented was counterbalanced (i.e. participants either did all the pure or all mixed blocks first), as were the trial types within the pure blocks presentation. At the end of each block of 96 trials participants initiated a keypress to move on to the next block.

Results

Only correct responses within 200ms and 2500ms were included in the analyses. The proportions of valid responses for the mixed and pure blocks were .967 (*SD* = .027) and .965 (*SD* = .021) respectively. Table 2 lists the descriptive statistics for all four trial types.

A 3x2 repeated measures ANOVA revealed a significant trial type (neutral, non-response set or response set) by presentation format (mixed or pure) interaction, $F(2,78) = 3.56$, $p = .033$, $r = .209$. A follow-up repeated measures one-way ANOVA measuring differences across the neutral, non-response set, and response set trial types was significant for both mixed and pure blocks ($F(2,78) = 20.589$, $p < .001$, $r = .457$, and $F(2,78) = 13.119$, $p < .001$, $r = .379$, respectively).

Planned comparisons between the response set and non-response set trials for each presentation format showed that the response set (different-response) trials were slower than non-response trials in both mixed (17.49ms, *t(*39) = 2.80, *p* = .008, *r* = .409), and pure block presentations (46.22ms, *t*(39) = 4.15, *p* < .001, *r* = .553) which meant that the response set effects for both presentation formats were statistically significant. Follow up comparisons of the size of the effect showed that the response set effect was larger in pure blocks compared to the mixed blocks (28.73ms, *t*(39) = 2.76, *p* =.009, *r* = .553).

To determine whether the mixing effect fits into either the homogenisation or mixing cost patterns described earlier, the three trial types were compared across the presentation formats. Non-response set and neutral trials were non-significantly different across presentation format (-8.49ms, *t*(39) = -0.92, *p* = .365, *r* = .146; and 7.37ms, *t*(39) = 0.902, *p* = .373, *r* = .143, respectively) but response set (different-response) trials were slower in pure blocks (20.24ms, *t*(39) = 2.15, *p* = .038, *r* = .326).

Table 2: Mean RTs in ms (and SEs) of all trial types and mean response set effect of Experiment 1

	Mixed	Pure
Neutral	586.62 (13.12)	593.99 (13.47)
Non-response	610.01 (13.41)	601.52 (13.75)
Response set	627.50 (15.29)	647.74 (17.88)
Response set effect (Response set – Non-response)	17.49 (11.13)	46.22 (6.24)

Error analysis

The 3x2 repeated measures ANOVA on the error rates revealed a non-significant interaction F(2,78) = 1.55, p = .218, r = .140. The main effect of trial type was also

non-significant ($F(2,78) = 0.62$, $p = .543$, $r = .089$) as was the main effect of presentation format ($F(1,39) = 1.13$, $p = .295$, $r = .168$).

Discussion

Consistent with our predictions, the results showed that presentation format modulated the size of the response set effect in the Stroop task; with larger effects observed when trials were presented in pure blocks compared to when presented in mixed blocks. In terms of the Stroop task, this means that when trials are presented in a random order, both semantic and response competition contribute roughly equally to Stroop interference, but when presented in pure blocks, Stroop interference is mainly made up of response competition. Interestingly, this mixing effect was driven by the difference in performance to response set (different-response) trials, with slower RTs to response set (different-response) trials in the pure block condition compared to in the mixed block condition. The RTs to non-response set trials were larger in mixed blocks presentation compared to pure blocks, although this difference was statistically non-significant. Nonetheless, this pattern of results resembles the homogenisation account more than the mixing cost account of mixing effects.

Worse performance to response-set trials in pure blocks suggest that it is more difficult to establish response-level salience in the pure blocks context. This goes against the predictions of models such as WEAVER++ (Roelofs, 2003) and the PDP model of Cohen et al. (1990), where concepts that are relevant to the task (i.e. response set colours) are identified via top-down processes of flagging or selective allocation of attention and thus, the identification of such concepts should not be affected by experimental design. Although Roelofs (2001) suggested that establishing this salience requires some repetition in the opening trials, such a

‘salience learning phase’ account should instead predict an effect in the opposite direction, with better performance in the pure blocks where only relevant colour concepts are presented.

A similar account is described in other conflict adaptation accounts that investigate congruency effects. Studies manipulating the proportion of incongruent and congruent trials within blocks of trials (e.g. Kane & Engle, 2003; Lindsay & Jacoby, 1994; Logan & Zbrodoff, 1979; Lowe & Mitterer, 1982; West & Baylis, 1998) typically report larger Stroop interference when there are more congruent trials. This phenomenon is attributed to a strategic shift towards word reading strategies due to the probability of encountering a congruent trial. However, such conflict adaptation accounts have been contested by research (e.g. Egner, 2014; Mordkoff, 2012; Schmidt & Besner, 2008) showing that the effects can be accounted for unintended learning biases brought up by experimental design. Although congruent trials were not used in the current research, having only one trial type in a block means 100% proportion of one type of trial leading to the same type of interference is encountered throughout each block. Hence, conflict adaptation accounts would indicate that it would be easier to strategically allocate attention to deal with the conflict compared to the situation in mixed blocks, where different types of conflict is present within the different trials of each block, leading to better performance to trials in pure blocks. Our finding of faster RTs to response set (different-response) trials in the mixed blocks compared to pure blocks is in the opposite direction of what conflict adaptation accounts predict.

The fact that strategic allocation of attention and conflict adaptation accounts cannot explain the performance of response set (different-response) trials led us to explore the role of bottom-up mechanisms instead. In Experiment 2

we introduce and test a hypothesis that the number of colour concepts activated in the task affects the ability to inhibit distracting information.

A further challenge posed by our results concerns the performance of non-response trials, which had slower RTs, albeit statistically non-significant, when presented in mixed blocks compared to pure blocks. This is inconsistent with models of the Stroop effect, such as the WEAVER ++, that predict the performances of non-response set and response set (different-response) trials would be affected in tandem (i.e. non-response set trials produce interference due to their conceptual-associative links to response set (different-response) trials).

Experiment 2

Postulated account based on a limited resource, labile and transient association process

The results from Experiment 1 show that the response set effect is smaller in a mixed block context, which suggests that it is harder to establish which colours are relevant to the task in such contexts, and thus reducing the response set effect. In this experiment we proffer and test an account of this effect based on exposure to colour concepts in the irrelevant dimension.

While participants are exposed to the same trials in both presentation formats, when trials are presented in pure blocks containing only one trial type each, they are exposed to a restricted set of colour concepts within each block of neutral and response set (different-response) trials. The absence of exposure to the non-response colour words may likely to result in the increased activation of the concepts of the response set colours (even more so in the response set trial block since the distractor words and font colour activate a task-relevant colour). When the restricted set of colours is repeatedly presented without any intervening non-response set colour or non-colour words, it is likely that all attentional

resources would be allocated to that small set of concepts, making them more accessible and thus more likely to interfere when they are presented in the irrelevant dimension.

In mixed blocks, however, the presence of non-response set trials (along with neutral trials) results in a greater number of colour concepts being involved in the task at any one time. In the example of Experiment 1, twice the number of colours was activated in the mixed blocks than in a pure response set (different-response) trials block. With more active colour concepts, attentional resources would be distributed such that each colour has relatively a smaller amount of activation, which would result in them being easier to inhibit when activated as the irrelevant word dimension (i.e. better performance to response set (different-response) trials).

Response set effects were observed in the mixed block condition suggesting that salience is still established, just not quite as strongly. Unlike the previously mentioned accounts, this account does not assume only a strategic top-down mechanism establishes certain colour concepts as more salient. Rather, saliency is also established through exposure to concepts in the irrelevant dimension through a presumably implicitly learned process.

To test this hypothesis we manipulated the number of non-response set colours participants were exposed to, and consequently the proportion of response set to non-response set colour exposure in each block. It was predicted that the response set effect would be smaller in the mixed blocks than in the pure blocks and that it would be smallest in the mixed block with the larger number of non-response set colours.

Method

Participants

40 students (4 male, age: *M* = 19.03, *SD* = 1.12) participated in exchange for course credit.

Design

A 4x 2 within subjects repeated measures design was used. The two independent variables were trial type (neutral, 6 non-response set colours, 2 non-response set colours, & response set) and presentation format (mixed & pure).

Apparatus and Materials

The apparatus and materials used were the same as those in the previous experiment with the only difference being an additional mixed block condition in which the number of non-response set colours was larger (6 colour-words) than in the other (2 colour-words; referred to here as *6NR* and *2NR* respectively).

Stimuli

As with Experiment 1, two versions of the experiment were administered. The response set colours were purple (204; 0; 255), yellow (255; 255; 0) and green (0; 255; 0); in one version, and white (255; 255; 255), blue (0; 112; 192), and orange (255; 127; 0) in the second version. For the non-response set trials, the irrelevant words used in the 2NR condition were ‘pink’ and ‘blue’; and ‘green’ and ‘yellow’ in the respective versions, while the 6NR contained the additional words ‘red’, ‘brown’, ‘white’, ‘orange’ for version one and ‘pink’, ’red’, ‘brown’, ‘purple’ in the other version. Neutral trials were included but only to keep to the original design as closely as possible.

Procedure

Each participant completed three sets of blocks: one set of pure blocks and two separate sets of mixed blocks. The pure blocks set contained blocks of each of the four trial types (neutral, 6NR, 2NR and response set) while each set of mixed blocks consisted of three blocks of neutral, response set (different-response) trials, and non-response set trials with either 6 or 2 non-response set colours, with each block containing an equal number of trials of each trial type randomly presented. In other words, participants went through 10 experimental blocks (4 pure blocks: neutral, 6NR, 2NR, and 3 mixed blocks of neutral, 2NR and response set (different-response) trials, and 3 mixed blocks: neutral, 6NR and response set (different-response) trials) with 72 trials in each block. A practice block made up of 48 trials preceded the experimental blocks, which resulted in a total of 768 trials performed by each participant. The order of the sets (6NR, 2NR) was counterbalanced across participants, as was the order of the presentation format and trial types in the pure block format.

Results

Using the same criteria as Experiment 1, the total number of valid responses in the pure, mixed 2NR and mixed 6NR sets were .967 (*SD* = .022), .964 (.014) and .965 (.018) respectively. The mean RTs of each trial type are detailed in Table 3.

The magnitudes of the response set effects were calculated in the following ways: For the two mixed blocks, the effects were calculated by taking the difference between the RTs to response set (different-response) trials and the corresponding non-response trials of the block, while in the pure block set, two response set effects were obtained by taking the difference between the response set (different-response) trials block and each of the two non-response blocks. This

led to four measures of the response set effect, one in each of the mixed block conditions and two in the pure block presentation condition.

To determine the effect of presentation format, a one-way ANOVA on the four response set effects yielded a significant effect ($F(3,117) = 7.95$, $p < .001$, $r = .252$). Planned comparisons revealed a non-significant difference between the two response set effects in the pure blocks ($t(39) = 0.18$, $p = .855$, $r = .029$), but larger response set effects in the pure blocks compared to the corresponding response set effect in the mixed blocks (6NR: $t(39) = 3.34$, $p = .002$; $r = .472$, 2NR: $t(39) = 2.58$, $p = .014$, $r = .382$). These analyses showed a general mixing effect where response set effects were larger in pure blocks compared to mixed blocks, which is consistent with the findings of Experiment 1.

The effect of having different number of activated colour concepts in the irrelevant dimension was investigated by comparing the magnitude of the response set effect in the two mixed blocks. A pairwise comparison between them showed that as predicted, the response set effect was larger when there were less non-response colours ($t(39) = 2.62$, $p = .013$, $r = .387$).

To compare the pattern of results to the mixing cost and homogenisation accounts, separate one-way ANOVAS were conducted on the RTs of each trial type was across the all the blocks. The effects of presentation format was non-significant for Neutral trials ($F(2,78) = 2.52$, $p = .087$, $r = .177$) and non-response (both 2 and 6NR) trials ($F(3,117) = 2.54$, $p = .060$, $r = .146$), but statistically significant for response-set trials ($F(2,78) = 3.28$, $p = .043$, $r = .201$). Pairwise comparisons within the response set (different-response) trials showed only the difference between the trials in the pure and mixed (6 NR colours) blocks to be statistically significant ($t(39) = 2.82$, $p = .008$, $r = .412$) while the difference between the response set (different-response) trials in the two mixed blocks ($t(39)$

= 1.95, p = .058, r = .298); and mixed (2 NR colours) and pure blocks (t(39) = 0.59, p = .558, r = .094) were statistically non-significant.

Table 3: Mean RTs in ms (and SEs) of all trial types and mean response set effect of Experiment 2

		Mixed (2 Non-resp)	Mixed (6 Non-resp)	Pure
Neutral		637.77 (13.85)	628.23 (14.56)	619.01 (13.00)
2NR		652.19 (15.29)	-	629.25 (15.08)
6NR		-	651.59 (18.74)	631.02 (14.29)
Response set		660.40 (15.64)	641.91 (15.70)	667.61 (18.68)
Response set effects	6NR	-	-9.68 (7.18)	36.59 (11.97)
	2NR	8.21 (6.42)	-	38.36 (10.96)

Error rates

The one-way ANOVA on the error rates for the four response set effects was statistically non-significant (F(3,117) = 1.58, p = .198, r = .115)

Discussion

The results from this experiment replicated the effect of trial type mixing on the magnitude of the response set effect. Furthermore, a comparison of the response set effects in the two sets of mixed blocks revealed a larger effect in the 2NR blocks compared to the 6NR blocks. This finding shows that the size of the response set effect is smaller when more of non-response set colours is present in the irrelevant dimension, which is consistent with the notion that the magnitude of response set effect is influenced by the number of colour concepts activated in any experimental block or at any one time. In other words the response competition is diluted when more colour concepts are active in a block of trials and is heightened

when both the relevant and irrelevant dimensions of the Stroop stimulus in a block only contain colours concepts that are potential response options.

It should be noted that the observed significant difference between the response set effects of 2NR and 6NR mixed blocks might have benefitted from the raw RTs to non-response set trials being longer than those to response set (different-response) trials in the 6NR block. A similar finding was observed in a previous study from our lab (Hasshim & Parris, 2014, Experiment 2b). Nothing in our presented theory predicts longer RTs to non-response set trials and thus this potentially represents a challenge to our theory. However, given this is a null effect we shall not interpret it further.

The comparisons of each trial type across the 2NR mixed, 6NR mixed and pure blocks revealed that the only statistical significant effect was faster RTs to response-set trials in the pure blocks compared to 6NR mixed blocks, which fits with the previous finding that the mixing effect is driven by facilitation to response set (different-response) trials when presented in pure blocks. The direction of effects were also more in line with the homogenisation account of mixing, although, like the results of Experiment 1, the difference in the RTs of non-response trials did not reach statistical significance.

General Discussion

The experiments in this study set out to investigate the effect of presentation format on response set effects in the Stroop task. Data from both experiments showed response set effects to be smaller when the trials were presented in pure blocks that contained only one trial type each, compared to mixed blocks that contained trials from all trial types, randomly presented. Although only response set (different-response) trials were significantly affected by the mixing effect, the

overall pattern of results were more consistent with the homogenisation account with response set (different-response) trials being impeded and non-response set trials being facilitated, in pure blocks compared to when presented in mixed blocks.

Experiment 2 was conducted to test a proposed limited resources account of the observed pattern by varying the number colour words appearing in the irrelevant dimension of non-response set trials. This manipulated the number of colour concepts activated within a block, with results showing a negative relationship between the number of colour words and the size of the resulting response set effect. It should be noted that since only the number of non-response colour words were manipulated, the results do not allow us to identify whether the effect is limited to variation in non-response colour concepts or whether manipulating the number of colours in the response set would have the same effect. We attempted to address this by conducting another experiment manipulating the number of response set colours, but adding an additional response option had a general effect of increasing RTs in all trials by ~100ms, which potentially occluded any expected experimental effect. Another possible explanation for the mixing effect is that having a lower number of irrelevant words in the pure blocks makes it easier to ignore them, and thus facilitating responses. The comparison between the two pure non-response set trial blocks is a direct test of this and the difference between them were statistically non-significant.

The present research offers important insights into the processes involved in the mechanisms of selective attention. Our results suggest that response set effects are not the result of the ability to ignore colour concepts that have not been identified as task relevant via a fixed, pre-set top-down bias or flagging. Although being part of the response set makes a distractor more difficult to inhibit, as shown

by response set (different-response) trials having slower RTs compared to non-response set trials, the amount of interference is modulated by the number of other non-response colours in the same block. The negative relationship between the number of colour concepts in a block and the size of the response set effects along with the non-significant difference between the 2NR and 6NR trials (within pure and mixed block presentations) suggest that the amount of interference a response set colour-word elicits depends on its level of exposure. If task relevant colours were somehow identified and fixed according to task instructions or even after a few trials, there would be no effect of presentation format. Our results indicate the presence of a bottom-up process that helps establish concept salience (which in turn determines the amount of interference they elicit), and that salience can be diluted by the presence of more colour concepts in the to-be-ignored dimension.

The results found in the current study do not conform to those of Lamers et al. (2010) who showed the benefits of being able to predict the distractor dimension of a Stroop stimulus. In their study's second experiment, they showed that cuing the irrelevant colour word facilitates RTs to incongruent (response set) trials, indicating a benefit to processing the irrelevant dimension. However, they cued the irrelevant colour word 2000 ms prior to target presentation, which is not typical of the Stroop task where both relevant and irrelevant information is presented simultaneously, which likely gave the participants the chance to inhibit the irrelevant word by the time the Stroop stimulus appeared. Also, the trial-by-trial cuing reliably indicated the identity of the irrelevant colour for the specific trial but did not affect the overall activation levels of the colours at the block level, which is a departure from the manipulations of the current study.

Another implication that the current finding has on current models of Stroop interference is that our experimental manipulation did not significantly affect RTs to non-response set trials. Post-hoc analysis of the effects of increasing the number of non-response set colours also shows no effect on non-response set effects (difference between non-response and neutral trials; $F(3,117) = 0.66$, $p = .581$, $r = .075$), which is inconsistent with predictions from models suggesting they should be affected in tandem (Roelofs, 2003). However, this finding is a null results and thus should be interpreted with caution.

Implications for the debate on the automaticity of reading

The inability to prevent the irrelevant colour word from interfering with colour naming has been taken as evidence for word reading being an automatic (happening without intent and not requiring attentional resources) and ballistic (cannot be stopped once started) (Brown, Gore & Carr, 2002; Neely & Kahan, 2001; and Posner & Snyder, 1975). However, the demonstration that Stroop interference can be reduced using manipulations such as the narrowing of spatial attention (e.g. Besner, 2001; Besner, Risko, & Sklair, 2005; Besner, Stolz, & Boutilier, 1997; Labuschagne & Besner, 2015, Stolz & McCann, 2000) social priming (Goldfarb, Aisenberg, & Henik, 2011) and a post-hypnotic suggestion (e.g. MacLeod & Sheehan, 2003; Parris, Dienes & Hodgson, 2012; Raz & Campbell, 2011; Raz, Moreno- Iñiguez, Martin, & Zhu, 2007; Raz, Kirsch, Pollard, & Nitkin-Kaner, 2006; Raz et al., 2003; Raz, Sharipo, Fan & Posner, 2002) has been taken as evidence against the notion of that word reading is automatic.

In their reviews of these studies, Augustinova and Ferrand (2014) and Flaudias and Llorca (2014) pointed out that Stroop interference is made up of both semantic and response based processes. Augustinova and Ferrand (2014) argued

that only the former is assumed to be automatic, and as such studies need to show that their manipulations affect semantic processes before a claim for control over ‘automatic’ processes can be made. They also argued that the use of manual responses, which are the norm for such studies, is not appropriate for measuring semantic processes since they have been shown to mainly index response conflict in the Stroop task (Sharma & McKenna, 1998). Therefore, they argued that even when these studies showed an elimination of Stroop interference, they were unlikely to have demonstrated a reduction in semantic processing and instead were affecting response based processes only.

The findings from the present research suggest that semantic conflict is involved in manual response Stroop tasks, findings that are consistent with those from Brown and Besner (2001) who presented a reanalysis of the Sharma and McKenna (1998) paper on which Augustinova and Ferrand’s argument is based. Indeed, the present results show that in mixed blocks a meaningful portion of interference can be semantic. Thus the argument that these studies report manipulations only affecting response conflict is inaccurate, although we do agree with Augustinova and Ferrand (2014) that to convincingly show that semantic processing is affected, reduction in interference on a trial type that isolates response from semantic conflict in the Stroop task, such as non-response trials, is necessary. In our estimation, this is best achieved using pure block presentation of trial types.

Conclusion

By demonstrating the modulation of the response set effect, a well-established component of Stroop interference, the present study highlights how the make-up of Stroop interference is not fixed and is instead, to some extent at least,

dependent on experimental context. Future studies investigating the contributions of semantic and response conflict to the Stroop task will have to take heed of the present findings. Finally, we have argued that response sets are established by computing relevant and irrelevant perceptual components, and that irrelevant components can, somewhat ironically, dilute those selective attention mechanisms responsible for facilitating goal-oriented behaviour. The mere computation of this irrelevant content represents a failure of selective attention indicating it is not the result of optimal selective mechanisms, but rather a consequence of a mechanism computing goal-related, but not goal-relevant information.

Experiment 3

The bigger response set effect in pure blocks in the previous experiments was hypothesised to be due to having fewer active colours in the response set (different-response) trials pure block compared to the mixed block, resulting in increased activation of the colour concepts in the pure block (and therefore more difficulty in inhibiting them when they were in the irrelevant dimension). However, in Experiment 2 the number of non-response set colours was manipulated while the number of response set colours were held constant and it was shown that there is a negative relationship between the number of non-response colours and the size of the response set effect. The current experiment investigates whether this property of response set effects is driven by the proportion of non-response set colours compared to response set colours since there was a greater proportion of non-response set colours in Experiment 2, or whether it is a more general effect of having more colour concepts involved in the task. To achieve this, the current experiment varied the total number of concepts activated in the task while always keeping the number of response and non-response colours equal.

If the absolute number of active concepts affects the performance to response set (different-response) trials, possibly due to lesser cognitive resources available to activate the response set concepts (which are sometimes irrelevant) when more colours are involved, the difference between the response set effects of pure and mixed blocks is expected to be smaller in the four-response (eight-colours) version. Furthermore, the response set effects in the four-response condition are expected to be smaller compared to the three-response (six-colours) version.

As with Experiment 1, participants performed the task in both pure and mixed block presentations. In addition they also performed the task with three and

four response options. In the former, three response colours (and thus three response buttons) and three non-response colours were involved while the latter had four colours of each set (four response buttons). This means that in the three-response condition, the number of colours activated while performing response set (different-response) trials were three in pure blocks and six in mixed blocks. In the four-response condition, the number of colours activated was four and eight respectively. Thus in each version, the proportion of colour words in the non-response set is always 50% of the total number of colour concepts used, but there were more active colour concepts in the four-response version.

Method

Participants

40 students (5 male, age: *M* = 20.7, *SD* = 3.90) participated in exchange for course credit or £5.

Design

A 2x2x2 within subjects repeated measures design was used. The independent variables were: number of responses (3 & 4), presentation type (pure & mixed), and effect type (response set effect and non-response set effect).

Apparatus and Materials

The three trial types used were: 1) neutral trials, where the words were not associated with any colour; 2) non-response set trials, where the words spelt out a colour not part of the response set; and 3) response set (different-response) trials, where the word spelt out an incongruent colour that was part of the response set. Three response buttons were used as using only two would mean that each word stimulus would appear in the incongruent colour 100% of the time, allowing for the possibility of responses to be due to the learnt association between the word

stimuli and colour. To control for contingency, each word stimulus appears in only two of the three colours for the three-response condition and three of the four colours in the four-response condition. The omitted colour for each stimulus type in each condition (3 and 4 response) was counterbalanced to ensure that each colour appeared equally frequently across all conditions. Thus each word stimulus was mapped to two and three possible response buttons in the three-response and four-response conditions respectively with equal probability of either being correct.

Two versions of each experiment were administered, counterbalanced across participants. Words spelling out the possible colour responses in one version acted as the word stimuli in the other version's non-response set trials. For the three-response condition, one version used the colours yellow (RGB: 255; 255; 0), pink (255; 20; 147) and green (0; 255; 0), while the other used blue(0; 112; 192), purple(204; 0; 255) and orange(255; 127; 0). The neutral words used were STORY, WALL, MARVEL. In the four response condition, the colour white (255; 255;255) and red (255; 0; 0) were respectively added along with the neutral word DUE. The words in each condition were matched in frequency and length.

Procedure

On each trial, participants were presented with a grey fixation cross in the centre of the screen for 500ms followed by the Stroop stimulus which remained on the screen until a response was made. They were instructed to press the corresponding key to the colour of the text as quickly as possible while ignoring the word's meaning. An auditory tone and a visual error message were given on an error. The message lasted for 1500ms followed by a 100ms blank screen.

At the start of each session participants went through a practice block of 60 trials made up of hash symbols of three to six in length. On the experimental

blocks participants did 96 trials of each condition for each presentation format, resulting in 1152 trials altogether. The experiment was administered in blocks of 96 trials and participants went through all the pure blocks either before or after all the mixed blocks. The order of the pure block presentation was counterbalanced, as was whether they performed the pure or mixed blocks first and the order they went through three or four colour versions.

Results

Only correct responses that were within 200ms and 2500ms were included in the analyses. The proportion of valid responses for the mixed and pure blocks were .952 (*SD* = .037) and .949 (.037) respectively in the four-response version and .965 (.026) and .962 (.020) in the three-response version. The mean RTs of each condition are presented in Table 4. Since the effect of the experimental

Table 4: Mean RTs in ms (and SEs) of all trial types and mean response set and non-response effect of Experiment 3

	Mixed (3 responses)	Pure (3 responses)	Mixed (4 responses	Pure (4 responses)
Neutral	571.85 (11.98)	573.97 (12.61)	700.34 (18.20)	671.10 (17.02)
Non-response set	602.88 (12.41)	595.04 (14.13)	722.28 (17.70)	699.71 (17.43)
Response set	610.94 (14.15)	626.22 (15.86)	726.44 (18.33)	725.50 (17.41)
Non-Response set effect	31.03 (5.11)	21.07 (11.0)	21.94 (7.18)	28.61 (9.66)
Response set effect	8.06 (5.74)	31.19 (8.61)	4.16 (8.21)	25.79 (10.03)

manipulation on the size of response set and non-response set effects were being evaluated, these were measured by calculating the mean difference between response set and non-response trials, and non-response trials and neutral trials, respectively.

The 2x2x2 (number of responses, presentation format and effect type) repeated measures ANOVA interaction was non-significant, $F(1,39) = 0.259$, $p = .614$, $r = .081$. The presentation format by effect type interaction was significant, $F(1,39) = 4.90$, $p = .033$, $r = .334$, but the presentation format by number of responses interaction ($F(1,39) = 0.80$, $p = .377$, $r = .142$) and the number of responses by effect type interaction ($F(1,39) = 0.08$, $p = .785$, $r = .045$) were non-significant.

Analyses on the presentation format by effect interaction showed that the response-set effect was larger in the pure block compared to mixed block presentation, $t(39) = 2.78$, $p = .008$, $r = .407$, while the non-response set effect was non-significantly different across presentation type, $t(39) = 0.25$, $p = .804$, $r = .040$.

Discussion

The three-way (number of responses, presentation format and effect type) interaction was statistically non-significant, which means that the number of response options did not affect the magnitude of the ‘mixing effect’, and thus does not support the prediction that the absolute number of active concepts is a factor determining the mixing effect. The presentation type by effect type interaction replicated the mixing effect found in Experiment 1 where the response set effect was reduced in mixed blocks compared to when presented in pure blocks.

An interesting finding from the present study is that, unlike the previous experiments showing the mixing effect, the smaller response set effect in mixed

blocks seemed to be driven by slower RTs to non-response set trials in mixed blocks and not faster RTs for response set (different-response) trials in pure blocks. However as noted earlier, comparisons made across blocks are not strictly valid and the only conclusions that can be made are on the sizes of the effects. The findings do corroborate earlier results that show the response set effect is reduced in mixed block presentation compared to pure blocks.

One potential weakness of the current design is the number of colours in each condition is actually close (6 vs. 8). The lack of an effect of the number of active colours could be a result of the difference in the number of active colours not being larger. However, RTs are about 100ms faster in the three-response condition compared to the four-response condition (see Table 4). This suggests that the additional response option in the four-response condition increases the difficulty of the task; an effect that would be exacerbated by having a larger difference in the number of active colours. This increase is relatively large compared to the size of response set effects that are expected to be around 20-30ms. The increased difficulty might occlude the actual magnitude of the effect and suggests that varying the number of responses might not be the best way to study the response set effect. Since an additional response option affects RTs by a large degree relative to the expected size of the actual effect being investigated, future studies wishing to investigate the research question may want to consider using vocal responses instead. The set of responses in that modality is not limited by memory for colour location, which means that varying the number of responses may not have as large an effect as it does with manual responses.

Chapter 6: Thesis Discussion

The overarching aim of this thesis was to further our understanding of how the cognitive system deals with irrelevant information that may be detrimental to task performance, while processing information that is relevant to the task at hand. This was done within the context of the debate between multi-stage, pre- response conflict resolution models of processing (e.g. De Houwer, 2003; Klopfer, 1996; Kornblum et al. 1990, 1995), and late selection / single-stage response level conflict resolution models (e.g. Cohen et al. 1990; Roelofs, 2003) in the Stroop task. The difference between these two accounts is that while both acknowledge that informational conflict can occur at the earlier semantic level as well as the later response level, the multi-stage models posits that semantic conflict can be resolved at the semantic level before processing goes on to the response level, where response conflict is then resolved. On the other hand single-stage models assume that that all conflict (including those at the semantic level) is passed on, unresolved, to the response level, which is the only stage where conflict resolution occurs.

The arguments informing this debate hinge on the ability to accurately measure semantic and response based conflict as individual constructs. Teasing apart these two processes is challenging since the performance of a cognitive task, such as the Stroop task, requires processing at both semantic and response levels to occur before the appropriate behaviour can be executed. Any effect measured would have gone through both processing stages before a behavioural response can be observed and thus it would be difficult to ascertain how much of each process contributed to the overall effect.

To circumvent this problem, researchers have come up with experimental manipulations designed to selectively affect only one of these processes. The

studies in this thesis were concerned with the construct validity of some of these methods and paradigms because of their gaining popularity as tools to study inhibitory processes in a variety of settings. Thus it is important to be sure that they are actually affecting and measuring the specific processes being investigated. While doing so, the studies in the thesis also examined the nature of response conflict and how it arises.

The focus of this thesis was on evaluating the measurement of semantic and response conflict in the Stroop task, specifically by considering the use of same-response trials (chapters 2, 3 and 4) and non-response set trials (chapters 2 and 5). The studies provided several findings, summarised below, that have both theoretical and methodological relevance and make important contributions to our understanding of interference in the Stroop task and the measurement of its different components.

Same-response and non-colour word neutral trials have identical performance

The studies in Chapters 2, 3 and 4 evaluated the same-response trials as a way of dissociating response and semantic conflict. Same-response trials require the use of the two-to-one response-mapping paradigm (De Houwer, 2003), and have seen increased use in a wide variety of research settings (e.g. Berggren & Derakshan, 2014; Chen, Bailey, Tiernan & West, 2011; Chen, Lei, Ding, Li & Chen, 2013; Wendt, Heldmann, Munte & Kluwe, 2007). Even though De Houwer (2003) indicated that observed RT differences between same-response and congruent trials are due to facilitation, researchers have regularly used it as a measure of semantic based competition (stimulus-stimulus conflict).

The initial concern identified with studies utilising this paradigm was the use of congruent trials as a baseline. Since congruency between the word and colour

of Stroop stimuli have been shown to improve performance through facilitation (a process that is qualitatively different from interference) there is a need to demonstrate that the measurement of semantic conflict using same-response trials has not been inflated by the use of congruent trials as a baseline. Thus one of the aims of the studies in Chapters 2 and 3 was to evaluate whether same-response trials are an accurate measure of semantic conflict by comparing their performance to neutral trials, which have been recommended (Brown, 2011; Laeng et al., 2011) as a better baseline for measuring interference, compared to congruent trials. Since neutral trials do not involve any facilitative or colour-related (semantic) components, any difference between the two trial types would help elucidate the nature of the same-response trials.

The results from Chapters 2 and 3 showed that the performance of same-response trials was not significantly different to that of neutral trials. This implies that after controlling for the effects of facilitation, same-response trials do not capture any additional interference effect compared to the neutral baseline, making them an inaccurate measure of semantic competition. Chapter 4 utilised eye-tracking and pupillometry techniques to follow up on the investigation of same-response trials. The better sensitivity and reduced sampling error of the eye tracker compared to the manual keyboard-based response meant that any small difference between the two conditions was more likely to be detected. Moreover, pupillometry, which is a well-established measure of effort, could potentially be used to reflect the processing difference between the two conditions. Since more processes are involved in the performance of same-response trials compared to neutral trials (due to potentially both semantic interference and response facilitation) it might be possible to dissociate them by the difference in effort required in performing them. However, the results showed that the pupillary

measurements echoed the results from the manual button-press and saccadic responses, which further strengthens the idea that same-response trials do not measure additional conflict compared to neutral trials.

The results of both pupillometry and saccadic responses were thus consistent with the findings of Chapters 2 and 3 in that the differences between same-response and neutral trials were non-significant, while the Bayes factors for both measures showed evidence for no difference between the two conditions. This convergence of findings using two other measures strengthens the conclusion that same-response trials are not a reliable measure of semantic competition as they do not index additional interference compared to neutral trials. Prior studies using congruent trials as the baseline have not been indexing interference, but rather facilitation on congruent trials.

Implications for research in the field

The lack of an observable difference between neutral and same-response trials has crucial consequences for studies that use the two-to-one paradigm as a means of isolating semantic competition. Effects that have been attributed to semantic conflict using same-response and congruent trials are instead likely due in part, if not wholly, to facilitation. Conclusions to those studies need to be re-evaluated in light of this and it was recommended that future studies using the paradigm should instead use neutral trials as their baselines. Below the implications of the present results for this findings is discussed.

Chen et al. (2011) observed ERPs in the parietal regions of the brain while participants performed the two-to-one response Stroop task and concluded that different areas were involved in resolving response and semantic conflict as determined by the differences in activation between responding to congruent,

same-response and different-response trials. They reported that the medial frontal negativity, and the conflict slow potential over the left lateral frontal region to be sensitive to response conflict but not semantic conflict, while parietal and right lateral frontal regions were sensitive to both semantic and response conflict. The authors concluded that the parietal region was primarily involved in response selection in the Stroop task, and the lateral frontal regions may be involved in response monitoring and conflict adaptation. In light of the investigations in this thesis one cannot definitively conclude that different neural regions have differential sensitivities to types of interference. The differential sensitivity of the medial and left lateral frontal regions to different-response (standard incongruent) trials is more likely sensitive to a greater amount of interference on these trials compared to same-response trials which are, according to the findings from this thesis, more like neutral trials. Moreover, since they compare same-response trials to congruent trials as their index of semantic conflict, specific sensitivity to this comparison might represent facilitation, not interference. That is not to say that ERPs could not detect differences that the measures in this thesis (i.e. reaction times and pupillometry) cannot, but that Chen et al.'s (2011) interpretation of their results are certainly now more in doubt, given other potential interpretations.

Similarly, van Veen and Carter's (2005) fMRI study utilised the two-to-one paradigm to weigh in on the inconclusive findings in the literature on the neural activity in the DLPFC and ACC during conflict. They identified separate regions of the ACC, prefrontal, and parietal areas of the brain that are distinctively activated by semantic and response conflict. Specifically, they found semantic conflict to engage more of the superior DLPFC while response conflict engages relatively more inferior areas, while in the ACC, semantic conflict engaged more posterior and more dorsal areas. van Veen and Carter noted that the observation of distinct

activation regions to semantic and response conflict is not common in similar neuroscience studies. However, if the contrast between same-response and congruent trials reflect facilitation instead of conflict, it would be consistent with their finding of non-overlapping activation since facilitation and conflict are different processes. van Veen and Carter (2005, p. 501) even indicated that their study differed from other similar studies in that they used a congruent baseline instead of a non-colour neutral word baseline of the other studies (e.g. Milham et al., 2001; and West et al., 2004). They stated that it is possible that the measurement of response and semantic conflict in their study might not be accurate and that more empirical work needs to be done to check this. The requested empirical investigation has been presented in this thesis and suggests that their experimental results are not easily interpreted in favour of different neural regions underpinning semantic and response conflict detection/resolution.

Using distributional analysis on behavioural data, both Chen et al. (2011), and Steinhauser and Hübner (2009) showed that interference captured by same-response trials is consistent across the RT distribution (affecting the overall mean) while different-response trials affect the slower RTs (skewness of the distribution). They argue that distributional analysis is therefore a useful method for identifying the different processes contributing to Stroop interference. Similar to the preceding section on neuroscience data, showing that the paradigm measures facilitation instead of semantic conflict means that the conclusions related to semantic conflict should be attributed to facilitation. Furthermore, it indicates that distributional analysis techniques, such as ex-Gaussian analysis, have not yet been shown to be able to differentiate different types of interference.

There are many other studies cited in previous chapters utilizing same-response trials and as previously noted these are becoming more common. The

reinterpretation of the studies above equally applies to these studies. Future research might yet successfully differentiate same-response and neutral trials; for example, it is possible that if overall interference were generally increased, a difference between the two conditions might be observed. One approach would be to modify congruency ratio by increasing the number of congruent trials, a technique which has been shown to increase the Stroop interference effect (Logan & Zbrodoff, 1979). However, by using this method there would be issues related to response contingency, which need to be avoided. Another approach might be to test individuals who show naturally greater Stroop interference, such as those with low working memory (Kane & Engle, 2003). Again, though Kane and Engle did this using the congruency ratio manipulation, they observed no difference between individuals with low and high working memory when congruency ratio was 1:1. A potential option that does not involve modifying congruency ratio is a response-stimulus interval (RSI) manipulation. Both De Jong et al. (1999) and Parris (2014) have shown that interference is greater at longer RSIs. De Jong et al. used a long RSI condition of 2000ms while Parris used an even longer RSI of 3500ms. In the present experiments the RSIs were 500ms, which is closer to their shorter RSI conditions (200ms for both). Hence a study using an RSI manipulation to increase overall interference might show a difference between same-response and neutral trials. Nevertheless, until a method is identified to increase interference without contingency issues the results from this thesis strongly suggest that the use of same-response trials should be avoided and that another alternative is required.

Non-response set trials and the Response Set Effect

Since the suitability of same-response trials as a way to dissociate semantic and response conflict was cast in doubt by the initial experiments of Chapters 2, 3

and 4, the latter experiments in these chapters evaluated the suitability of an alternative trial type non-response set trials. The response set membership effect has been used as a measure of response competition in many studies (see Table 1 in Chapter 5) with Macleod (1991) listing it as one of the 18 well-established findings that models of the Stroop task need to explain. Studies looking to disentangle semantic and response conflict should consider utilising non-response trials. While the additional interference captured by non-response trials is small, every consideration should be taken to allow for semantic processing to be fully expressed before categorically dismissing its involvement and/or reduction following experimental manipulations.

However, investigating the appropriateness of this alternative trial type led to another important finding from this thesis, that the make-up of Stroop interference is affected by task design, which has clear implications for studies utilising the Stroop task. Researchers need to consider how the presentation type chosen affects the construct that they intend to measure. Crucially, the interpretation of past research might need to be reconsidered. An example of this, as highlighted in Chapter 5, is in the on-going debate on the automaticity of reading since the results from the chapter clearly show that the assumptions regarding the make-up of Stroop interference are based on studies employing pure block presentation while the studies reporting on the debate typically use mixed blocks.

In the investigation of the role of presentation format (either mixed or pure blocks) on the make-up of Stroop interference, it was observed that response conflict, as measured by the response set effect, is significantly smaller when trials are presented in mixed blocks. This mixing effect on the magnitude of response set effects was largely due to a reduction in RT to response set (different-

response) trials while non-response set trials were largely unaffected by trial type mixing. If anything, RTs to non-response set trials increased when RTs to response set (different-response) trials were decreasing, which is inconsistent with the notion that interference on non-response set trials is the result of their connections to response set colours (cf. Roelofs, 2003). If this were the case, the two trial types would have been affected in tandem.

The first experiment of Chapter 5 showed that the processes involved in performing non-response trials are more complex than is widely thought, while the second experiment attempted to elucidate the mechanisms involved in their performance and how it differs from incongruent (different response) trial performance. Based on the results from the first experiment, a theory based on a limited resources of attention account was put forward and tested. Consistent with the hypotheses, the results showed that the proportion of activation/attention received by the distractor colour concept is a critical factor influencing the amount of interference on a trial. It is more difficult to inhibit a distractor colour the more activated it is. When there are fewer colour concepts from the irrelevant dimension active during the performance of a block, more attentional resources can be allocated to each colour and thus it is harder to inhibit when that colour happens to be the irrelevant dimension in a trial. In contrast, when there are more active colour concepts, a smaller amount of attentional resources is allocated to each and thus it is relatively easier to overcome the less activated colour. Experiment 3 in Chapter 5 aimed to test whether it was the total number of active colour concepts as opposed to the number of colour concepts in the irrelevant dimension. By keeping the proportion of colours in both dimensions equal the design tested this possibility. The results were inconclusive due to an unexpected effect of the number of response options on overall reactions times. Reaction times were

~100ms longer in the four response condition which means that any interesting affects could have been hidden by this manipulation.

Theoretical implications

Showing that response conflict can be modulated by trial type mixing is a significant contribution to the Stroop literature. Popular models such as Cohen et al. (1990) and Roelofs (2003) have worked on the theoretical assumption that an irrelevant colour has to be part of the response set for it to directly interfere with processing and that the interference on non-response trials is simply due to the indirect activation of the response set colours via their connection to the non-response set colour concepts. The demonstration that presentation format affects the RTs to response set and non-response set trials in different ways is indicative of different processes being involved in the performance of the two types of trials. However, caution must be applied when make this interpretation because the effect of mixing on non-response set trials was null.

Extant models also describe the identification of information relevant to the task to occur via a top-down mechanism. However, in investigating the cognitive processes involved in dealing with response conflict, the results from Chapter 5 suggest that the number of concepts in the irrelevant dimension (non-response set colours) is important in determining the magnitude of the response set effect. It is not simply the operation of preset strategic flagging or biasing mechanism, but rather the relative activation levels of the distractor colour that determines the amount of interference. The reason why classic studies that compare non-response set trials to response set (different-response) trials typically show response competition (response set effects) making up the bulk of Stroop interference is that such studies inadvertently caused greater activation to the

response set colours compared to non-response set colours. This is due to there being no alternative colour concepts using up resources that are otherwise fully attached to the response set colours.

Models of Stroop interference typically do not consider the negative influence of activation of irrelevant concepts on task performance. The research here shows that higher levels of activation to a concept not only speeds up processing when it is in the relevant dimension, but it also slows it down when the same colour concept is in the irrelevant dimension due to greater difficulty in inhibiting them. This highlights the need to consider the influence of irrelevant concepts has on Stroop task performance. Ironically, it is processing information in the irrelevant dimension that confers a benefit on Stroop task performance.

Pre-response pupillometry

Another notable contribution comes from the research presented in Chapter 4, which is the first study to utilise a pre-response measure of pupil dilation. Using this measure, the typical effects of Stroop interference, namely Stroop facilitation and interference, were distinguishable and thus demonstrating its utility in measuring effects in the Stroop task. This is potentially a major methodological contribution to the use of eye-tracking in experimental psychology research as it means that there is potential for greater flexibility in the use of pupillometry in such research. Although less sensitive than the more typically used post-response measures, the option to use pre-response data means that research utilising pupillometry is not limited to experimental designs with long trial durations or long RSIs. Of course, since this is the first time such a technique has been used, much more research needs to be done to ascertain the limitations and boundary conditions of its use. At a theoretical level, demonstrating the utility of pre-

response pupil information also weighs in on an on-going discussion about whether pupillary effects are simply a reflection of a behavioural response since pupillary changes have typically been measured only after a behavioural response is made, when the greatest pupil dilation occurs. Demonstrating that pupillary Stroop effects are measurable even before a behavioural response argues against the notion that it is dependent on a behavioural response.

Further future directions

Although the findings from Chapter 5 have important empirical and theoretical implications regarding the measurement of semantic and response interference in the Stroop task, there are still several theoretical questions that require further investigation before the mechanisms of the mixing effect can be fully understood. As mentioned in the chapter, it is unclear whether the number of non-response colour concepts specifically influences the magnitude of the response set effect or if the effect can be observed by manipulating the number and proportion of response set concepts or even the number of task-irrelevant words in general, including neutral non-colour related words. A series of studies systematically manipulating each of these factors, using the Stroop or non-colour Stroop-like tasks (that allow for a wider selection of response options) is a possible avenue for future research.

Vocal Responses

Since the studies in this thesis have mainly focused on manual responses, it is important to consider whether the results would be applicable to vocal responses as well since it has been shown that the makeup of Stroop interference is influenced by response modality (Sharma & McKenna, 1998). A cursory look at the data from Chapter 5 reveals effect sizes comparable to what would be

expected in a pure block vocal response task experiment (e.g. Sharma & McKenna, 1998), where both semantic and response based processes had large and notable contributions to overall Stroop interference. It is possible that since the number of possible responses is not restricted for vocal responses as it is for manual responses (i.e. any colour can be vocalised), utilising a vocal response would be more fruitful in revealing the effects of increasing the number of overall colour concepts on the response set effect, and would escape the main effect of number of response colours observed Experiment 3 of Chapter 5.

Semantic associates and non-response trials

To differentiate semantic from response competition, the studies in this thesis have used non-response set trials instead of the more popular semantic associates (e.g. see Augustinova & Ferrand, 2014; Flaudias & Llorca, 2014, for reviews advocating the importance of controlling for semantic processes by including semantic associates). Non-response set trials were chosen because they are a more conservative measure of response processes compared to semantic associates as demonstrated by Sharma and McKenna (1998), a frequently cited study for their measurement of the magnitude of semantic and response based components of the Stroop task (e.g. Augustinova & Ferrand, 2012; 2014; Ferrand & Augustinova, 2013, Flaudias & Llorca, 2014). In their study, Sharma and McKenna (1998) used different trial types that tap into different levels of lexical, semantic and response processing. While semantic associates were used (with the difference in performance between them and the slower neutral words being labelled as semantic relatedness), they also identified a further level of semantic processing, semantic relevance, which was indexed by non-response set trials. They attributed the difference between response set (standard incongruent) trials

and non-response set trials (response set membership effect) to be due to interference at the late response selection stage, unlike the other processes that occur earlier in the semantic and lexical stages. Thus, even though semantic associates do capture a portion of semantic processing, taking the difference between them and incongruent (different response) trials as a measurement of response conflict might not be the most accurate account since there is a difference between semantic associate trials and non-response set trials.

A more conservative boundary for measuring response processes, such as non-response set trials, might be required to more accurately differentiate semantic from response processes. For example, Risko, Schmidt, and Besner (2006) illustrated the effect of the additional process caused by response set membership of semantic associates. They manipulated whether the colour associated with the semantic associate was part of the response set or not and found slower RTs to trials where the associated colour was part of the response set in both vocal and manual response modalities. However, it has yet to be clearly established whether non-response set trials do or do not involve some level of response competition. Although the results from this thesis go some way to supporting the notion they do not, more work is needed to better understand non-response set trials. Notably, models of the Stroop task have tended to account for interference on semantic associate and non-response set trials in similar ways; through their connections with response set (different-response) trials.

Timing accounts

The descriptions of mixing effects in other literatures (i.e. mixing cost and homogenisation) are of different timing accounts. That is to say, the mechanisms that result in the difference in performance is described as a change in the

decision-making thresholds for a response to be made. Such an account can be investigated using formal decision making threshold models (e.g. diffusion models and linear ballistic accumulator models) although they typically deal with binary choice tasks (e.g. see Ratcliff & McKoon, 2008; and Brown & Heathcote, 2008). It would be worthwhile to adapt an appropriate task that manipulates presentation format to see how the models can be fit to such data.

How it all fits in with formal models

The findings from this thesis provide insight into the measurement of semantic and response competition and how formal models of Stroop interference do not adequately account for the processes involved in performing these trials. Specifically, the findings from Chapter 5, have notable implications for models of Stroop task performance. Firstly, it is clear that the way current models account for response set effects need to be rethought. These models typically attribute performance on non-response trials to be a consequence of performance on response set (different-response) trials. Since non-response set colours are not relevant to the task, models such as Cohen et al.'s PDP and the WEAVER ++ assume that they do not have any direct influence on processing since a top down process in the system (task demand unit and flagging, respectively) identifies only the relevant colours that should be processed. Any interference to non-response set colours is seen as a by-product of interference to response set (different-response) trials. This is purportedly due to the non-response colours being connected to the relevant response colours; in other words performance to non-response trials is viewed as simply watered down Stroop interference. However, the results from the studies in this thesis consistently show that response set and non-response set trials were not similarly affected by the different presentation

formats and crucially, the effects on the two trial types were, numerically at least, in opposite directions. These are compelling indicators that different processes are involved in their performance, contrary to the accepted wisdom. However, since the conclusions on effects of trial type mixing on the non-response set trials are based on null results, they cannot be taken as strong evidence in favour of different processes although this is certainly a compelling avenue for further research in the future.

The findings from the studies on non-response trials and presentation format also highlight the important influence of the stimulus and experimental design on performance at a more macro level. Although steps have been made to include bottom-up processes into models (e.g. Cohen & Huston, 1994) much is still to be done to identify how such processes actually affect the processing and performance on the task.

Models of the Stroop task have typically concentrated on the benefits conferred as a result of flagging or biasing response set colours and the resultant ease of processing relevant information. Clearly the opposite effect is also important; activation also makes such information difficult to be inhibited when in the irrelevant dimension. A highly activated distractor is more difficult to inhibit, which is reflected by worse performance. So a top-down flagging or biasing or response set colour would represent in essence a failure of selective attention in situations where the response colours can also be in the irrelevant dimension. The present results suggest that selective attention mechanisms do not actually work this way and that response set membership is established during the task and is labile in a way that top-down mechanisms would not be. Computing response level salience involves the irrelevant dimension. When there are more colours in the irrelevant dimension it is harder to establish response level salience and thus the

response set effect disappears. This describes the operation of a mechanism that computes the occurrence of colour concepts independent of top-down goals.

Conclusion

The studies in this thesis set out to evaluate same-response and non-response trials, experimental manipulations that have been used to measure and dissociate semantic and response conflict in the Stroop task. Same-response trials were found to be unsuitable for this endeavour while non-response trials were shown to be a possible alternative. Notably, investigation into non-response set trials led to novel findings relating to the structure of Stroop interference and how it is modulated by task design.

The research presented in this thesis provides several methodological insights into measuring Stroop processes. Of course, while these methodological findings are novel they should be further developed and explored in future research. Along with the identification of important gaps in the theoretical accounts of the Stroop task, the work presented in this thesis has been informative to researchers studying the processes involved in the Stroop task and also wider implications such as for the literature on the automaticity of reading. The present research has highlighted how the mechanisms producing Stroop interference are not well understood and that there is still much more research to be done. It is hoped that the work presented in this thesis will be a useful initial step towards this endeavour.

References

Augustinova, M., & Ferrand, L. (2012). Suggestion does not de-automatize word reading: Evidence from the semantically based Stroop task. *Psychonomic Bulletin & Review, 19*(3), 521-527.

Augustinova, M., & Ferrand, L. (2014). Automaticity of Word Reading Evidence From the Semantic Stroop Paradigm. *Current Directions in Psychological Science, 23*(5), 343-348.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445-459.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychological Bulletin, 121*(1), 65.

Bauer, L. O., & Hesselbrock, V. M. (1999). Subtypes of family history and conduct disorder: Effects on P300 during the Stroop test. *Neuropsychopharmacology, 21*(1), 51-62.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276-292.

Berggren, N., & Derakshan, N. (2014). Inhibitory deficits in trait anxiety: Increased stimulus-based or response-based interference?. *Psychonomic Bulletin & Review, 21*(5), 1339-1345.

Besner, D. (2001). The myth of ballistic processing: Evidence from Stroop's paradigm. *Psychonomic Bulletin & Review, 8*(2), 324-330.

Besner, D., Risko, E. F., & Sklair, N. (2005). Spatial attention as a necessary preliminary to early processes in reading. *Canadian Journal of Experimental Psychology, 59*(2), 99-108.

Besner, D., Stolz, J. A., & Boutilier, C. (1997). The Stroop effect and the myth of automaticity. *Psychonomic Bulletin & Review, 4*(2), 221-225.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review,108*(3), 624.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153-178.

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*(04), 308-319.

Brown, M., & Besner, D. (2001). On a variant of Stroop's paradigm: Which cognitions press your buttons? *Memory & Cognition, 29*(6), 903-904.

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*(1), 85-99.

Brown, T. L., Gore, C. L., & Carr, T. H. (2002). Visual attention and word recognition in Stroop color naming: Is word recognition "automatic?". *Journal of Experimental Psychology: General, 131*(2), 220-240.

Caramazza, A., & Costa, A. (2000). The semantic interference effect in the picture-word interference paradigm: Does the response set matter? *Cognition, 75*(2), B51-B64.

Caramazza, A., & Costa, A. (2001). Set size and repetition in the picture–word interference paradigm: Implications for models of naming. *Cognition, 80*(3), 291-298.

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*(2), 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.

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

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*(3), 332.

Cohen, J.D. & Huston, T.A. (1994) Progress in the use of interactive models for understanding attention and performance. In C. Umilta & M. Moscovitch (Eds.) *Attention and Performance XV* pp. 453–456, Cambridge, MA: MIT Press.

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*, 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*(3), 353-359.

De Jong, R., Berendsen, E., & Cools, R. (1999). Goal neglect and inhibitory limitations: Dissociable causes of interference effects in conflict situations. *Acta Psychologica, 101*(2), 379-394.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193-222.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64,* 135-168.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*(3), 274–290.

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

Donders, F. C. (1969/1858). On the speed of mental processes. *Acta Psychologica, 30*, 412-431.

Dyer, F. N. (1973). The Stroop phenomenon and its use in the stlldy of perceptual, cognitive, and response processes. *Memory & Cognition, 1*(2), 106-120.

Egner, T. (2014) Creatures of habit (and control): a multi-level learning perspective on the modulation of congruency effects. *Frontiers in Psychology 5*:1247.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of Learning and Motivation, 44,* 145-200.

Flaudias, V., & Llorca, P. M. (2014). A brief review of three manipulations of the Stroop task focusing on the automaticity of semantic access. *Psychologica Belgica, 54*(2), 199-221.

Fox, L. A., Shor, R. E., & Steinman, R. J. (1971). Semantic gradients and interference in naming color, spatial direction, and numerosity. *Journal of Experimental Psychology, 91*(1), 59-65.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science, 17*(2), 172-179.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance, 8*(6), 875-894.

Glaser, W. R., & Glaser, M. O. (1989). Context effects in stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118*(1), 13-42.

Goldfarb, L., & Henik, A. (2006). New data analysis of the Stroop matching task calls for a reevaluation of theory. *Psychological Science, 17*(2), 96-100.

Goldfarb, L., Aisenberg, D., & Henik, A. (2011). Think the thought, walk the walk–Social priming reduces the Stroop effect. *Cognition, 118*(2), 193-200.

Goldfarb, L., & Henik, A. (2007). Evidence for task conflict in the Stroop effect. *Journal of Experimental Psychology: Human Perception and Performance, 33*(5), 1170-1176.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology, 33*(4), 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.

Henik, A., & Salo, R. (2004). Schizophrenia and the stroop effect. *Behavioral and Cognitive Neuroscience Reviews, 3*(1), 42-59.

Hock, H. S., & Egeth, H. (1970). Verbal interference with encoding in a perceptual classification task. *Journal of Experimental Psychology, 83*(2p1), 299-303.

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*(5), 569-574.

Kahneman, D. (1973). *Attention and effort* (p. 246). Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 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*(1), 47-70.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience, 23*(1), 315-341.

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *The American Journal of Psychology, 77*(4), 576-588.

Klopfer, D. S. (1996). Stroop interference and color-word similarity. *Psychological Science, 7*(3), 150-157.

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*(4), 855-875.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: cognitive basis for stimulus-response compatibility--a model and taxonomy. *Psychological Review, 97*(2), 253-270.

La Heij, W. (1988). Components of Stroop-like interference in picture naming. *Memory & Cognition, 16*(5), 400-410.

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

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary stroop effects. *Cognitive processing, 12(*1), 13-21.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18-27.

Lamers, M. J., Roelofs, A., & Rabeling-Keus, I. M. (2010). Selective attention and response set in the Stroop task. *Memory & Cognition, 38*(7), 893-904.

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*(2), 219-234.

Loewenfeld, I. E. (1993). The pupil. In *Anatomy, Physiology, and Clinical applications* (Vol. 1, pp. 695-707). Iowa State and Wayne State University Press Ames and Detroit.

Logan, G. D., & Irwin, D. E. (2000). Don't look! Don't touch! Inhibitory control of eye and hand movements. *Psychonomic Bulletin & Review, 7*(1), 107-112.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition, 7*(3), 166-174.

Los, S. A. (1996). On the origin of mixing costs: Exploring information processing in pure and mixed blocks of trials. *Acta Psychologica, 94*(2), 145-188.

Lowe, D. G., & Mitterer, J. O. (1982). Selective and divided attention in a Stroop task. *Canadian Journal of Psychology 36*(4), 684-700.

Luo, C. R. (1999). Semantic competition as the basis of Stroop interference: Evidence from color-word matching tasks. *Psychological Science, 10*(1), 35-40.

Lupker, S. J., Brown, P., & Colombo, L. (1997). Strategic control in a naming task: Changing routes or changing deadlines? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(3), 570-590.

Lupker, S. J., Kinoshita, S., Coltheart, M., & Taylor, T. E. (2003). Mixing costs and mixing benefits in naming words, pictures, and sums. *Journal of Memory and Language, 49*(4), 556-575.

MacKinnon, D. P., Geiselman, R. E., & Woodward, J. A. (1985). The effects of effort on Stroop interference. *Acta Psychologica*, 58(3), 225-235.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163-203.

MacLeod, C. M. (1992). The Stroop task: The" gold standard" of attentional measures. *Journal of Experimental Psychology: General, 121*(1), 12-14.

MacLeod, C. M. (2005). The Stroop task in cognitive research. In A. Wenzel & D. C. Rubin (Eds.) *Cognitive Methods and Their Application to Clinical Research*, 17-40 Washington, DC: American Psychological Association

MacLeod, C. M. (2007). The concept of inhibition in cognition. In D. S. Gorfein & C. M. MacLeod (Eds.), *Inhibition in Cognition*, 3-23 Washington, DC: American Psychological Association.

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*(1), 126-135.

MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences, 4*(10), 383-391.

MacLeod, C. M., & Sheehan, P. W. (2003). Hypnotic control of attention in the Stroop task: A historical footnote. *Consciousness and Cognition, 12*(3), 347-353.

Mathews, A., & MacLeod, C. (1985). Selective processing of threat cues in anxiety states. *Behaviour Research and Therapy, 23*(5), 563-569.

Melara, R. D., & Algom, D. (2003). Driven by information: a tectonic theory of Stroop effects. *Psychological Review*, 110(3), 422-471.

Milham, M. P., Banich, M. T., Webb, A., Barad, V., Cohen, N. J., Wszalek, T., & Kramer, A. F. (2001). The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Cognitive Brain Research, 12*(3), 467-473.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100.

Mordkoff, J. T. (2012). Observation: Three reasons to avoid having half of the trials be congruent in a four-alternative forced-choice experiment on sequential modulation. *Psychonomic Bulletin & Review, 19*(4), 750-757.

Neely, J. H., & Kahan, T. A. (2001). Is semantic activation automatic? A critical re-evaluation. In H.L. Roediger, J.S. Nairne, I. Neath, & A.M. Surprenant

(Eds.), *The Nature of Remembering: Essays in Honor of Robert G. Crowder* (pp. 69–93). Washington, DC: American Psychological Association.

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.

Parris, B. A., & Dienes, Z. (2013). Hypnotic suggestibility predicts the magnitude of the imaginative word blindness suggestion effect in a non-hypnotic context. *Consciousness and Cognition, 22*(3), 868-874.

Parris, B. A., Bate, S., Brown, S. D., & Hodgson, T. L. (2012). Facilitating goal-oriented behaviour in the Stroop task: When executive control is influenced by automatic processing. *PloS one, 7*(10), e46994.

Parris, B. A., Dienes, Z., & Hodgson, T. L. (2012). Temporal constraints of the word blindness posthypnotic suggestion on Stroop task performance. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 833-837.

Parris, B. A., Dienes, Z., Bate, S., & Gothard, S. (2014). Oxytocin impedes the effect of the word blindness post-hypnotic suggestion on Stroop task performance. *Social Cognitive and Affective Neuroscience, 9*(7), 895-899.

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal of Experimental Psychology, 60*(2), 211-229.

Posner, M. I., & Snyder, C. R. R. (1975). Facilitation and inhibition in the processing of signals. *Attention and Performance V*, 669-682.

Proctor, R. W. (1978). Sources of color-word interference in the Stroop color-naming task. *Perception & Psychophysics, 23*(5), 413-419.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873-922.

Raz, A., & Campbell, N. K. (2011). Can suggestion obviate reading? Supplementing primary Stroop evidence with exploratory negative priming analyses. *Consciousness and Cognition, 20*(2), 312-320.

Raz, A., Kirsch, I., Pollard, J., & Nitkin-Kaner, Y. (2006). Suggestion reduces the Stroop effect. *Psychological Science, 17*(2), 91-95.

Raz, A., Landzberg, K. S., Schweizer, H. R., Zephrani, Z. R., Shapiro, T., Fan, J., & Posner, M. I. (2003). Posthypnotic suggestion and the modulation of Stroop interference under cycloplegia. *Consciousness and Cognition, 12*(3), 332-346.

Raz, A., Moreno-Íniguez, M., Martin, L., & Zhu, H. (2007). Suggestion overrides the Stroop effect in highly hypnotizable individuals. *Consciousness and Cognition, 16*(2), 331-338.

Raz, A., Shapiro, T., Fan, J., & Posner, M. I. (2002). Hypnotic suggestion and the modulation of Stroop interference. *Archives of General Psychiatry, 59*(12), 1155-1161.

Richer, F., & Beatty, J. (1985). Pupillary dilations in movement preparation and execution. *Psychophysiology, 22*(2), 204-207.

Risko, E. F., Schmidt, J. R., & Besner, D. (2006). Filling a gap in the semantic gradient: Color associates and response set effects in the Stroop task. *Psychonomic Bulletin & Review, 13*(2), 310-315.

Roelofs, A. (2001). Set size and repetition matter: Comment on Caramazza and Costa (2000). *Cognition, 80*(3), 283-290.

Roelofs, A. (2003). Goal-referenced selection of verbal action: modeling attentional control in the Stroop task. *Psychological Review, 110*(1), 88-125.

Scheibe, K. E., Shaver, P. R., & Carrier, S. C. (1967). Color association values and response interference on variants of the Stroop test. *Acta Psychologica, 26*, 286-295.

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*(3), 514-523.

Schmidt, J. R., & Cheesman, J. (2005). Dissociating stimulus-stimulus and response-response effects in the Stroop task. *Canadian Journal of Experimental Psychology. 59*(2), 132-138.

Schmidt, J. R., Crump, M. J., Cheesman, J., & Besner, D. (2007). Contingency learning without awareness: Evidence for implicit control. *Consciousness and Cognition. 16*(2), 421-435.

Sharma, D., & McKenna, F. P. (1998). Differential components of the manual and vocal Stroop tasks. *Memory & Cognition*, 26(5), 1033-1040.

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*(1), 63-76.

Simpson, H. M. (1969). Effects of a task-relevant response on pupil size. *Psychophysiology, 6*(2), 115-121.

Steinhauser, M., & Hubner, 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*(5), 1398-1412.

Stelmack, R. M., & Siddle, D. A. (1982). Pupillary dilation as an index of the orienting reflex. *Psychophysiology, 19*(6), 706-708.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method*. Acta psychologica, 30*, 276-315.

Stirling, N. (1979). Stroop interference: An input and an output phenomenon. *The Quarterly Journal of Experimental Psychology, 31*(1), 121-132.

Stolz, J. A., & McCann, R. S. (2000). Visual word recognition: Reattending to the role of spatial attention. *Journal of Experimental Psychology: Human Perception and Performance, 26*(4), 1320-1331.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology,18*(6), 643-662.

Sugg, M. J., & McDonald, J. E. (1994). Time course of inhibition in color-response and word-response versions of the Stroop task. *Journal of Experimental Psychology: Human Perception and Performance, 20*(3), 647-675.

Sullivan, K., & Edelman, J. (2009). An oculomotor Simon effect. *Journal of Vision, 9*(8), 380-380.

Unsworth, N., Spillers, G. J., Brewer, G. A., & McMillan, B. (2011). Attention control and the antisaccade task: A response time distribution analysis. *Acta Psychologica, 137*(1), 90-100.

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*(1), 158-169.

van Veen, V., & Carter, C. S. (2005). Separating semantic conflict and response conflict in the Stroop task: A functional MRI study. *Neuroimage, 27*(3), 497-504.

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*(7), 1104-1112.

West, R., & Baylis, G. C. (1998). Effects of increased response dominance and contextual disintegration on the Stroop interference effect in older adults. *Psychology and Aging, 13*(2), 206-217.

West, R., Bowry, R., & McConville, C. (2004). Sensitivity of medial frontal cortex to response and nonresponse conflict. *Psychophysiology, 41*(5), 739-748.

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*(1), 3-19.

Zhang, H. H., Zhang, J., & Kornblum, S. (1999). A parallel distributed processing model of stimulus–stimulus and stimulus–response compatibility. *Cognitive Psychology, 38*(3), 386-432.

Appendices

Appendix 1: Proportion contingency for experiments in Chapter 2

Experiment 1	Number of trials				contingency (%)	
	RED	GREEN	YELLOW	BLUE	button 1	button 2
red	24	24	12	12	67	33
green	24	24	12	12	67	33
yellow	12	12	24	24	33	67
blue	12	12	24	24	33	67
Experiment 2a						
red	0	24	12	12	50	50
green	24	0	12	12	50	50
yellow	12	12	0	24	50	50
blue	12	12	24	0	50	50
wall	6	6	6	6	50	50
due	6	6	6	6	50	50
marvel	6	6	6	6	50	50
story	6	6	6	6	50	50
Experiments 2b & 3						
red	0	24	12	12	50	50
green	24	0	12	12	50	50
yellow	12	12	0	24	50	50
blue	12	12	24	0	50	50
wall	6	6	6	6	50	50
due	6	6	6	6	50	50
marvel	6	6	6	6	50	50
story	6	6	6	6	50	50
purple	6	6	6	6	50	50
white	6	6	6	6	50	50
blue	6	6	6	6	50	50
grey	6	6	6	6	50	50

Appendix 2: Proportion contingency for experiments in Chapter 3
(note that counterbalanced versions used different colour and word stimuli but proportions remain the same)

Experiment 1	Number of trials				contingency (%)	
	RED	GREEN	YELLOW	BLUE	button 1	button 2
red	24	24	12	12	67	33
green	24	24	12	12	67	33
yellow	12	12	24	24	33	67
blue	12	12	24	24	33	67
Experiment 2a						
red	12	12	12	12	50	50
green	12	12	12	12	50	50
yellow	12	12	12	12	50	50
blue	12	12	12	12	50	50
wall	3	3	3	3	50	50
due	3	3	3	3	50	50
marvel	3	3	3	3	50	50
story	3	3	3	3	50	50
Experiment 2b & 3						
red	0	12	6	6	50	50
green	12	0	6	6	50	50
yellow	6	6	0	12	50	50
blue	6	6	12	0	50	50
wall	3	3	3	3	50	50
due	3	3	3	3	50	50
marvel	3	3	3	3	50	50
story	3	3	3	3	50	50
purple	3	3	3	3	50	50
white	3	3	3	3	50	50
blue	3	3	3	3	50	50
grey	3	3	3	3	50	50

Appendix 3: Proportion contingency for experiments in Chapter 4
(note that counterbalanced versions used different colour and word stimuli but proportions remain the same)

	Number of trials				contingency (%)	
	RED	GREEN	YELLOW	BLUE	response 1	response 2
red	12	12	12	12	50	50
green	12	12	12	12	50	50
yellow	12	12	12	12	50	50
blue	12	12	12	12	50	50
wall	3	3	3	3	50	50
due	3	3	3	3	50	50
marvel	3	3	3	3	50	50
story	3	3	3	3	50	50

Appendix 4: Proportion contingency for experiments in Chapter 5
(note that counterbalanced versions used different colour and word stimuli but proportions remain the same)

Experiment 1	Number of trials			contingency (%)		
	YELLOW	PINK	GREEN	button 1	button 2	button 3
yellow	0	16	16	0	50	50
pink	16	0	16	50	0	50
green	16	16	0	50	50	0
blue	0	16	16	0	50	50
purple	16	0	16	50	0	50
orange	16	16	0	50	50	0
wall	0	16	16	0	50	50
marvel	16	0	16	50	0	50
story	16	16	0	50	50	0

Experiment 2 (2NR)						
	YELLOW	PURPLE	GREEN			
yellow	0	12	12	0	50	50
purple	12	0	12	50	0	50
green	12	12	0	50	50	0
pink	12	12	12	33	33	33
blue	12	12	12	33	33	33

Experiment 2 (6NR)						
	YELLOW	PURPLE	GREEN			
yellow	0	12	12	0	50	50
purple	12	0	12	50	0	50
green	12	12	0	50	50	0
pink	3	3	3	33	33	33
blue	3	3	3	33	33	33
red	3	3	3	33	33	33
brown	3	3	3	33	33	33
white	3	3	3	33	33	33
orange	3	3	3	33	33	33

Experiment 3 (3 responses)						
	YELLOW	PINK	GREEN			
yellow	0	16	16	0	50	50
pink	16	0	16	50	0	50
green	16	16	0	50	50	0
blue	0	16	16	0	50	50
purple	16	0	16	50	0	50
orange	16	16	0	50	50	0

Experiment 3 (4 responses)

					Contingency (%)			
	YELLOW	PINK	GREEN	WHITE	button 1	button 2	button 3	button 4
yellow	0	8	8	8	0	33	33	33
pink	8	0	8	8	33	0	33	33
green	8	8	0	8	33	33	0	33
white	8	8	8	0	33	33	33	0
blue	0	8	8	8	0	33	33	33
purple	8	0	8	8	33	0	33	33
orange	8	8	0	8	33	33	0	33
red	8	8	8	0	33	33	33	0