Does Response Modality Influence Conflict? Modelling Vocal and Manual Response Stroop Interference

Alex Fennell, Roger Ratcliff

Funded by:

- National Institutes of Health, National Institute on Aging

Does Response Modality Influence Conflict? Modelling Vocal and Manual Response Stroop Interference

Alex Fennell, Roger Ratcliff

The Ohio State University

Address correspondence to:

Alex Fennell
Fennell.50@osu.edu
The Ohio State University,
Department of Psychology,
Columbus, OH, 43210 USA

Author Note

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number R01AG041176. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abstract

In the Stroop task, color words are presented in colored fonts and the task of the subject is to either name the word or name the color. If the word and font color are in agreement, then the stimulus is said to be congruent (e.g. RED in red font color), however if the word and font color are not in agreement, the stimulus is said to be incongruent (e.g. RED in blue font color). Conflict in the Stroop task is measured by both response time and accuracy. In prior research, the amount of conflict differs depending on the response modality, vocal vs. manual. We applied a model for multichoice decision-making (and confidence), the RTCON2 model (Ratcliff & Starns, 2013) to the data from four experiments, two with two-choice manual responses, one with four-choice manual touch screen responses, one with both four-choice vocal responses and four-choice manual keyboard responses. Changes in the rate of information accumulation captured conflict effects for the manual-response versions, but not for the vocal-response version. Adding an extra non-decision time parameter allowed RTCON2 to account for the data patterns in the vocal-response version. However, in order to fully understand conflict in the vocal-response Stroop task, a model of conflict processing in the vocal word production system must be developed that would explain the additional processing time in the nondecision time parameter.

Keywords: Stroop task, response modality, cognitive modelling, diffusion model, RTCON2, sequential sampling models

Does Response Modality Influence Conflict? Modelling Vocal and Manual Response Stroop Interference

Selectively attending to relevant aspects of the environment, while ignoring distracting information, is key to carrying out goal-oriented actions. The critical importance of selective attention has made it an important area of investigation within psychology. There are a number of tasks that have been used to investigate this, and in this article, we focus on the Stroop task (Stroop, 1935). There are many versions of the task, but the general idea is that individuals are presented stimuli that vary on two dimensions, and make a response based on one. In the classic paradigm, the stimulus is a color word presented in a font color. If the word and font color are in agreement, then the stimulus is said to be congruent (e.g. RED in red font color), however if the word and font color are not in agreement, the stimulus is said to be incongruent (e.g. RED in blue font color). Neutral stimuli have one dimension that is relevant to the decision while the other dimension does not contribute to the decision. Given this, the kind of neutral stimulus will depend on the task (e.g. XXXXX in red font color, when making a response based on font color). The participant is asked to make a response based either on what the word says, or on the font color in which it is presented. When making decisions based on the font color, incongruent stimuli produce slower responses compared to neutral stimuli. This is referred to as Stroop interference and is a hallmark effect within this task. It is also sometimes observed that individuals produce faster responses when responding to congruent versus neutral stimuli, a phenomenon referred to as Stroop facilitation (MacLeod, 1991).

The amount of Stroop facilitation and interference are influenced by a number of factors, but one that has been examined over the years, with several different theoretical accounts, is that of response modality. A number of studies have explicitly examined the effect of vocal responses versus manual keyboard responses, and there is a general finding that interference is less for manual responses compared to vocal responses, herein referred to as the modality effect (Neill, 1977; Redding & Gerjets, 1977; Sharma & McKenna, 1998; White, 1969). Furthermore, there are some instances in which the

facilitation effect is greater for manual responses compared to vocal responses (Redding & Gerjets, 1977). This suggests that there is a difference that depends on the response modality, but it is unclear what difference in processing produces this difference in the patterns of responses.

In this article we use a sequential sampling model, the RTCON2 model (Ratcliff & Starns, 2013), to account for behavioral data from the Stroop task. RTCON2 can be viewed as an extension of the two-choice diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008). Both models integrate reaction time (RT) and response proportions to produce psychologically interpretable parameters, which provide an account of processing. The primary aim of this study is to use RTCON2 to identify and understand differences in processing between these two different response modalities.

There have been numerous attempts to provide a theoretical account of Stroop phenomena with the use of mathematical models. Each of the models presents a different perspective on how information is processed in this task and how it differs as a function of response modality. Perhaps the most well-known model of Stroop phenomena is the one developed by Cohen, Dunbar, and McClelland (1990). This model is a feedforward network which implemented the assumption that automaticity is not a discrete all-or-nothing phenomenon, but rather a continuous one (Logan, 1985; MacLeod & Dunbar, 1988). The model contains two processing pathways, one for word and one for color, and each one contains an input layer of nodes, a layer of hidden nodes, and an output layer of nodes. The model is implemented with two colors and, within each pathway, two input nodes that correspond to the different colors used in the simulation. The connection weights between the input and hidden nodes, and those between the hidden and output nodes, differ between the color and word pathways. The word pathway has stronger connection weights than the color pathway, in order to reflect the fact that word reading is a more practiced skill than color naming (MacLeod & Dunbar, 1988; Posner & Snyder, 1975). In addition to these pathways, there are two attentional nodes, which are activated according to whether the task is font color naming or word reading. The font naming node is connected to the hidden

layer of the color pathway, and the word reading node is connected to the hidden layer of the word pathway. These attentional nodes bias activation of the hidden layer nodes, depending on task, to make units in one pathway more likely to be activated than the other.

With this structure, the model is able to simulate mean RTs of Stroop interference similar to those observed in empirical data. But, in its current form, the model in unable to account for the modality effect. However, it may be possible for the model to account for this effect by incorporating a node that reflects the response modality within which the task is presented. This modification has not been explicitly implemented, but if it were successful in capturing the modality effect, it would suggest that differences in conflict processing between these tasks is due to vocal responding being a more automatic process than manual responding. The response modality node would make connections to the word reading node that increase the activation of this node. Therefore, more inhibition would be required when incongruent stimuli are presented, resulting in more Stroop interference. On the other hand, keyboard responses would require less inhibition, resulting in less Stroop interference. This idea is supported by RT distributional analyses of the Stroop task (Spieler, Balota, & Faust, 2000; Steinhauser & Hübner, 2009). These studies show that the main difference in RTs between interference in the manual and vocal Stroop tasks is that the manual response tasks produce smaller shifts in the leading edge of the RT distributions between neutral and incongruent conditions when compared to vocal response versions.

One problem with the Cohen et al. (1990) model is that it produces RT distributions of the incorrect shape (Mewhort, Braun, & Heathcote, 1992). Thus, the model does not provide a complete description of the data and cannot address the issue of the effect of manual versus vocal on the leading edge of RT distributions. Also, there is still the question of whether this additional conflict in the vocal Stroop task is due to additional processing during the decision-making process, or whether this is due to additional time taken to encode or initiate a response.

The dimensional overlap model (Kornblum, Hasbroucq, & Osman, 1990; Kornblum, Stevens, Whipple, & Requin, 1999; Zhang, Zhang, & Kornblum, 1999) is a computational model that accounts for phenomena across a wide variety of tasks, including the Stroop task. The dimensional overlap model posits that Stroop interference arises from incompatibility of the stimulus dimensions (i.e. font color, and color word) as well as incompatibility between the stimulus and response. The main underlying assumption of this model is that processing occurs in two stages. The first stage is perceptual and captures the (in)compatibility of the stimulus dimensions. The second stage is the response production stage, which captures stimulus-response (in)compatibility. The model itself is represented by two layers of modules, an input layer, and an output layer. The input layer consists of modules that correspond to the relevant stimulus dimensions (e.g. font color), while the output layer consists of modules that correspond to the dimensions related to the response options. Additional modules can be added to these layers to capture effects from irrelevant stimulus dimensions, or irrelevant response dimensions.

Excitatory pathways connect the relevant stimulus modules to the response modules (e.g. red font color to red response option) while inhibitory pathways connect the irrelevant stimulus modules to the relevant stimulus modules. In addition to this, processing is divided into a stage-like structure in which processing in the input layer must cross a certain threshold, before processing can begin in the output layer. Thus, the processing time in the input layer can be attributed to perceptual processing, whereas the processing time in the output layer can be attributed to response production.

With this structure the dimensional overlap model has simulated Stroop facilitation and Stroop interference, along with several other Stroop phenomena (Kornblum et al., 1999; Zhang et al., 1999). It is also able to simulate RT distributions that approximate those seen across a wide variety of Stroop and related paradigms (Zhang et al., 1999). The dimensional overlap model has not been specifically applied to capture the modality effect, but it would be feasible to modify the model to do so. If another module representing the task dimension was added to the model (e.g. different modules for manual responses

and vocal responses) with inhibitory and excitatory connections to both the stimulus dimensions and the response modules, then the modality effect could possibly be simulated. The main shortcoming of this model is that it does not predict response accuracy and so it does not provide a complete description of the data.

The RACE/A model (van Maanen & van Rijn, 2007; van Maanen, van Rijn, & Taatgen, 2012) is a model of the Stroop task that offers a more complete explanation of the processing that occurs during the task. RACE/A operates within the cognitive architecture of ACT-R (Anderson, 2007), while also integrating decision mechanisms from sequential sampling models. Before we explain how the RACE/A model accounts for interference in the Stroop task, we provide a brief overview of sequential sampling models.

One of the more successful ways to account for data in simple decision-making tasks has been with sequential sampling models. These models assume that information used in making decisions is noisy and so this information must be accumulated over time. The models also assume that a decision is made once a decision boundary is crossed, one boundary for each response choice. The accumulation process can either be absolute, with evidence accumulating separately for each response option towards a separate decision boundary for each choice (Vickers, 1970). Or the process can be relative, with a single accumulator and with a response being made when evidence for one response exceeds evidence for the others by a certain criterial amount (Ratcliff, 1978). The quality of information, or drift rate, in conjunction with the amount of information required to make a response, or decision criterion, allow this class of models to make predictions about RT, error rates, and their distributions. Drift rate is dependent on stimulus quality, with easier stimuli having higher drift rates, and difficult stimuli having lower drift rates. Decision boundaries on the other hand, reflect response conservativeness and are set by the individual. For a more in-depth review and comparison of sequential sampling models see Ratcliff

and Smith (2004) and Teodorescu and Usher (2013). Thus, this class of models attempts to give an account of the hypothesized processes that underlie the accuracy and RT measures.

In the RACE/A model, processing is distributed across a number of different systems (e.g. perceptual, motor response, vocal, memory) and produces response times based on the summation of time spent in these systems. The most successful application of RACE/A has been to the picture word interference paradigm. Specifically, RACE/A was used to test the hypothesis put forward by Dell'Acqua, Job, Peressotti, and Pascali (2007), that processing during picture word interference occurred during perceptual encoding, as opposed to response selection. This encoding view is one way in which Stroop interference has been hypothesized to occur (Fagot & Pashler, 1992). Using RACE/A, van Maanen, van Rijn, and Borst (2009) demonstrated that picture word interference and Stroop interference can be accounted for within a single framework. This is accomplished by dividing the processing into three stages, perceptual, decision, and response, with interference occurring unequally in each stage. It is conceivable that RACE/A could also capture the modality effect, however, the model has not been applied to the modality effect. The modality effect could be captured by changing the manner in which processes in the response stage occur between modalities. For example, it may be that motor responses associated with button pressing are simply faster than those associated with vocal responding. It could also be the case that interference occurs differently across the stages of processing for manual responses and vocal responses. The main drawback of the RACE/A model is that although the model provides predictions about RT distributions, these have not been explicitly examined or compared with data, and so it is unclear whether the differences in RT distributions between the vocal and manual response Stroop task could be captured by this model.

We have discussed three models that offer different accounts of how Stroop interference can be accounted for and how this differs as a function of response modality. The Cohen et al. (1990) model proposes that interference in the Stroop task is a result of response competition, and responding vocally

is more automatic than responding manually, which accounts for the modality effect. The dimensional overlap model attributes Stroop interference to occurring on both a semantic and response level, and thus the modality effect arises from differences in how inhibition interacts at both these stages. On the other hand, RACE/A suggests that interference is spread over many different stages of processing and the modality effect may be accounted for by different processing in the manual and vocal Stroop tasks. We propose RTCON2 as an alternative model to these previous models to provide a comprehensive description of the behavioral data, as well as psychologically interpretable parameters that reflect processing during the decision-making process. However, this model does not describe sources of conflict or how attention contributes to performance on the Stroop task. However, we suggest that the model might be a meeting point between it and the theories just reviewed, theories that provide little information about the decision process.

The RTCON2 model is a sequential sampling model that was developed to account for confidence judgments in decision-making tasks and other multi-choice paradigms (Ratcliff & Starns, 2013). One of the key assumptions of RTCON2 is that the degree of match between a stimulus and memory is conceptualized as a distribution over the degree of match between a test item and memory rather than a single value (Beck et al., 2008; Gomez, Ratcliff, & Perea, 2008; Jazayeri & Movshon, 2006; Ratcliff, 1981, 2018; Ratcliff & Starns, 2009). In the current experiment, however, evidence will be conceptualized as a single value, as was done in the motion discrimination experiment in Ratcliff & Starns (2013). RTCON2 employs a constant summed evidence algorithm in which evidence for one option is evidence against the others. The evidence against the other alternatives is equally divided among the alternatives, so the increment in the winning accumulator is equal to the sum of the decrements in the others. Furthermore, drift rates across the different accumulators add to one, evidence may go below zero, and there is no decay in evidence over time.

The most important parameters for examining the effects of Stroop interference are drift rate (v), non-decision time (T_{er}), and decision boundary settings (b). These correspond to separate and readily interpretable psychological constructs. The model is unable to divide up time allocated to encoding information, accessing memory, and initiating a response, and combines these into the single non-decision time parameter. The primary focus of the model is on the decision process, with drift rate and decision boundary settings being centrally important. Drift rate is a measure of the quality of evidence from a stimulus. Difficult stimuli have low drift rates which produces slower and less accurate responses. For example, incongruent stimuli in the classic Stroop paradigm would have lower drift rates than congruent stimuli, as these stimuli produce conflicting information, which makes decision about them harder. Decision boundary settings reflect differences in response style that vary across individuals. Lower boundaries indicate a less conservative response style in which less evidence is required for a response, which produces responses that are faster and less accurate than the average. Higher boundaries produce responses that are slower and more accurate.

In addition to these parameters, there are trial-to-trial variability parameters that reflect the assumption that the system cannot produce identical parameter settings from trial-to-trial. Across-trial variability in drift rate (s_v) reflects the assumption that the same stimulus will not be encoded in an identical manner when encountered on different occasions. This model also assumes there is variability in decision boundaries (s_b) across trials. This assumption is derived from the notion that individuals are assumed to be unable to hold fixed boundaries across trials, and instead have fluctuations in how much evidence is required to make a decision from trial-to-trial. Finally, there is across-trial variability in non-decision time (s_t) which represents trial–to-trial variability in processes outside the decision process. The first two of these variability parameters have been found to accommodate differences in the RT distributions of correct versus error responses (Ratcliff & Tuerlinckx, 2002). Perhaps most importantly,

the model includes within-trial variability in the accumulation process (σ), which accounts for fluctuations in evidence accumulation around its mean, during the accumulation process.

Figure 1 presents RTCON2 as applied to a color identification task with four response options, “red”, “green”, “blue”, “yellow. The black lines under the “accumulators” column correspond to evidence accumulation paths for a given trial. This is a noisy process, so the response option with the largest drift rate does not always win, and the response times will vary from trial to trial even with identical model parameters. In the example in the figure, the evidence accumulation process terminates first for the A response category resulting in a “red” response.

The RTCON2 model has several advantages for application in this domain. First, the model is able to deal with tasks that have more than two response options. Second, it is able to use all of the behavioral data (choice proportions and RT distributions for all the choices). Furthermore, the model does this for data from individual participants which allows it to account for the typically large differences in response patterns across participants. This allows this model-based analysis to examine whether an individual processes information similarly across tasks (Ratcliff, Thapar, & McKoon, 2010, 2011). The model does not accomplish this by being overly flexible: the RTCON2 model (as well as the standard diffusion model) is constrained to produce RT distributions that are positively skewed. Finally, RTCON2 provides fits that can be quantitatively assessed. In doing so, conclusions drawn from the model can either be supported or refuted according to the degree of match between the model and data.

The aim of the current article is to apply the RTCON2 model to data from the Stroop task in order account for decision processes involved in the task. We begin by comparing the model fits of RTCON2 to that of the standard diffusion model for data from two-choice Stroop tasks to show that it provides comparable fits to the diffusion model. As was previously mentioned, the interference effects in vocal response variants of the Stroop task are larger than those seen in manual response variants. We

can test the same individuals in both of these tasks and use RTCON2 to examine what processing components differ between the two tasks. The Stroop models reviewed earlier present some hypotheses for what processes may differ between response modalities in the RTCON2 model. The Cohen et al. model suggests that the modality effect arises from vocal responses being more automatic than manual responses. The dimensional overlap model suggests that the strength of the connection between input and output layers differs as a function of response modality. The hypotheses proposed by the Cohen et al. model and the dimensional overlap model would be reflected as a difference in drift rates across the tasks. On the other hand, RACE/A suggests that interference is distributed among all stages of processing, and this would be reflected in a combination of differences in drift rates, decision boundaries and non-decision times across tasks.

Experiment 1

The first two experiments were designed to assess whether the multi-alternative RTCON2 model can fit behavioral data from a two-choice Stroop conflict paradigm and whether these fits are comparable to those of the diffusion model. The diffusion model was chosen for comparison as it is constrained, and has been able to successfully account for data from a wide variety of experimental paradigms. Additionally, the RTCON2 model can be seen as an extension of the diffusion model, and as such, its parameter values should be consistent with those of the diffusion model across fits to individual participant data. Although similar in structure, RTCON2 does differ from the diffusion model. The main difference between the models is that RTCON2 must be fit by simulation, while the diffusion model has a closed form solution. It should be noted that in the context of two-choice tasks, RTCON2 is equivalent to the diffusion model. Previously, RTCON2 has been shown to produce similar parameter values, and model fits to other sequential sampling models, including the diffusion model, in the context of associative recognition, item recognition, and motion discrimination (Ratcliff & Starns, 2013; Voskuilen

& Ratcliff, 2016). It has yet to be determined whether the RTCON2 model can account for the patterns of behavioral data typically observed in the Stroop task, which is the aim of this study.

The first experiment is the classic Stroop experiment, presented in a manual response two-choice format. This experiment and all subsequent experiments were approved by The Ohio State University Social and Behavioral Institutional Review Board.

Method

Participants

Eighteen college-aged participants with normal or corrected to normal vision, were recruited from the introductory psychology course at The Ohio State University and were given course credit for their participation.

Materials

The stimuli consisted of four color words (“red”,” green”, “blue” and “yellow”) and the letter string “XXXXX”. These were presented in one of the four character colors, or the color white. Combining all of the color words and the letter string with all of the colors yielded 20 stimuli. All stimuli were presented in the center of the screen on a black background. The task included word identification and color identification blocks whose order was counterbalanced across participants. Each task included 32 blocks with 20 congruent items (color words presented in the matching character color); 20 incongruent items (color words presented in a nonmatching character color); and 20 neutral stimuli (color words presented in white font color for word identification blocks, or a string of X’s presented in a character color for color identification blocks). The stimuli were randomized so that no stimulus type, no stimulus dimension (character color, or color word), or response key mapping was repeated more than two times in a row. All stimuli were presented on a CRT display using a real-time computer system.

Procedure

The task took approximately 50 minutes to complete. The first two blocks of the task were practice blocks, one for word identification, and the other for color identification. Participants made their responses according to the name of the word in the word identification blocks, or the color in which the word was presented for the color identification blocks. The practice blocks consisted of 18 items, 6 congruent, 6 incongruent, and 6 neutral stimuli. Responses were collected using the "/" and "z" keys on a PC keyboard, and participants were instructed to use their right and left index fingers respectively to make responses. In this variant of the Stroop task, only two colors were present in each block. The two colors presented varied from block to block, but the response key mappings stayed consistent (i.e. red and blue were always the '/' key and green and yellow were always the 'z' key so the blocks had either red or blue and either green or yellow). In a given block, the color associated with the '/' could be presented with either of the two colors associated with the 'z' key (e.g. red could be presented with green in one block, and yellow in another). Instructions were presented at the beginning of each block that informed the participants the two colors that would be presented during the block, what type of block it was (color identification, or word identification), and with directions to make responses as quickly and accurately as possible. Items were presented on screen until a response was made, and were followed by blank screen that lasted 300 ms. Responses that took longer than 1250 ms. were followed by a "TOO SLOW" message that was displayed for 500 ms. Responses faster than 250 ms. had a "TOO FAST" message that was displayed for 1500 ms. An "ERROR" message was displayed on screen for 300 ms. when an error was made.

Analyses

Statistical analysis. RT latencies less than 250 ms. and greater than 1500 ms. were excluded from statistical analysis (<2% of the data). Mean RT and accuracy were submitted to a 2 (Block) x 3 (Stroop Condition) repeated measures ANOVA. Interactions were further examined with planned paired

t-tests. Specifically, we examined Stroop facilitation (i.e., congruent minus neutral) and interference (i.e., incongruent minus neutral) within each block type.

Model Fitting. There were 2 response choices and 6 conditions (word identification: congruent, incongruent, neutral, and color identification: congruent, incongruent, neutral). The RTCON2 and diffusion models were fit to the response proportions and .1, .3, .5, .7, and .9 RT quantiles for each individual participant for each response choice in all conditions. Optimal parameter values were obtained by first selecting initial parameter values that provided predictions similar to those observed in the data, (see Voskuilen & Ratcliff, (2016)) and then a simplex minimization routine (Nelder & Mead, 1965) was used to search for the parameters that provided predictions close to the observed data. A chi-square (χ^2) statistic was used to quantify the degree of match between the model predictions and empirical data. This was computed using the observed quantiles to produce the cumulative proportions between quantiles, and the frequencies by multiplying by the number of observations. The χ^2 statistic was computed for each of the six quantile bins (Ratcliff & Tuerlinckx, 2002). χ^2 for bins with less than seven observations used a single value from observed and expected proportions in the χ^2 calculation.

For the RTCON2 model, simulations were used to generate predicted values from the model as there are no exact solutions for this model. The accumulation process was simulated using Euler's method with 1 ms. steps (Brown, Ratcliff, & Smith, 2006; Usher & McClelland, 2001). At each time step, a randomly chosen accumulator was incremented or decremented, and the other accumulator received an equal and opposite amount of evidence (for further discussion of the equations see (Ratcliff & Starns, 2013). 20,000 iterations of the decision process were used to generate the response proportions and RT quantiles for each response option. The resultant simulated data comprise what will be referred to as model predictions, which are specific to a set of fixed parameters, rather than being predictions about future data based on fits of the current data.

The RT quantiles divide into 6 bins for each response option, which, for 2 choices, gives 12 degrees of freedom for each condition of the experiment. However, given that the response proportions must add to one, the degrees of freedom for each condition becomes 11. With 6 conditions, this gives a total of 66 degrees of freedom. There are 12 free parameters in the RTCON2 model for this experiment leading to 54 degrees of freedom. There are two drift rate means for each condition (which must add to one), each corresponding to one accumulator; this provides one independent parameter per condition. Examination of the data showed no bias towards one response option or the other, so data were collapsed into correct and error responses. As a result, only one decision boundary was free to vary, and the other was set equal to it. Furthermore, within-trial noise in the diffusion process in the RTCON2 model was fixed at .1, in order to directly compare parameter values of RTCON2 and the diffusion models. Thus, 7 parameters are used to model the stimulus representations used in the decision process (6 mean drift rates, and 1 between-trial variability in the mean of the drift distribution). The other 5 parameters represent the decision process (one decision boundary, across-trial variability in the decision boundaries, non-decision time, variability in non-decision time, and the scaling parameter on drift).

The diffusion model has 12 free parameters, with 7 parameters corresponding to the stimulus representation (6 drift rates, and 1 across-trial variability in drift rate). The other 5 parameters, similar to the RTCON2 model, correspond to the decision process (non-decision time, variability in non-decision time, boundary separation, starting point variability, and a parameter for contaminant responses, the last of which is not present in RTCON2). Starting point was fixed to be half of boundary separation because there is no bias for one response option over the other. Taking into account the number of free model parameters, there are 54 degrees of freedom for the diffusion model for this experiment.

Results

RT and Accuracy

The statistical analysis suggested that there was a substantially significant block x Stroop condition interaction for both mean RT, $F(2,34) = 14.26$, $p < .001$, $\eta^2 = .02$, and accuracy, $F(2,34) = 27.71$, $p < .001$, $\eta^2 = .09$. Responses in the color identification block were significantly slower than in the word identification block, $F(1, 17) = 6.76$, $p < .05$, $\eta^2 = .04$, but error rates did not substantially differ, $F(1,17) = 1.87$, $p > .05$, $\eta^2 = .02$. Specifically, in the word identification block, as shown in Table 1, participants produced significantly more errors, $t(17) = -2.87$, $p < .01$ $d = .68$, but did not differ in their response latencies, $t(17) = 1.08$, $p > .05$, $d = .26$, for the incongruent compared to neutral condition. Similarly, the facilitation effect for RTs was not significant within the word identification block, $t(17) = 1.83$, $p > .05$, $d = .43$, and participants were significantly less accurate in the congruent compared to neutral condition, $t(17) = -3.57$, $p < .01$, $d = .84$. On the other hand, in the color identification block, responses for the incongruent condition produced significantly longer response latencies, $t(17) = 3.43$, $p < .01$, $d = .81$, and more errors, $t(17) = -5.96$, $p < .001$, $d = 1.40$, than the neutral condition. There was also Stroop facilitation present in the color identification trials, with participants producing significantly faster responses, $t(17) = -3.20$, $p < .01$, $d = .75$, and fewer errors, $t(17) = 2.24$, $p < .05$, $d = .53$, in the congruent compared to neutral condition.

Model Fits

This section examines whether the diffusion and RTCON2 models fit the data and if the RTCON2 model provides predictions and parameter values similar to those of the diffusion model. We show that both RTCON2 and the diffusion model fit the data well, and that the parameter estimates of the decision process are strongly correlated across the models. Both models were fit to each participant's data individually which provided parameter estimates that, in turn, were used to generate predicted RT quantiles (.1, .3, .5, .7, .9) and response proportions for each condition.

The difference between the conditions is accounted for by allowing drift rate to vary among conditions. As can be seen in in Table 2 and 3, drift rate for the incongruent stimuli in the color

identification condition is smaller than that in the neutral condition. Thus, the degree of match between the stimulus and mental representation is smaller for incongruent stimuli, resulting in less accurate responses.

The model fits were assessed quantitatively by comparing the model predictions to the empirical data using a χ^2 statistic. For some participants, there were few error observations in both the neutral and congruent conditions for both tasks, and so the χ^2 was based on a single value for each error response category. The average parameter values are shown in Tables 2 and 3 for the RTCON2 and diffusion model respectively. The mean χ^2 value was 96.2 with a SD of 22.6 for the diffusion model and 119.4 with a SD of 26.8 for the RTCON2 model, which are both more than the critical value (72.2 for both models as within-trial noise in the diffusion process was fixed in RTCON2) suggesting a mismatch between the models' predictions and the data. For the diffusion model, 3 participants out of 18 had χ^2 values lower than the critical value, and RTCON2 had zero participants with χ^2 values lower than the critical value. Although the χ^2 values are larger than the critical value, the degree of mismatch is similar to that seen in other fits of the diffusion model. The mismatch is attributable in part to the sensitivity of the χ^2 test to large numbers of observations. With an average of 250 responses per condition per participant, even small deviations between the empirical data and model predictions will result in inflated χ^2 values (Ratcliff, Thapar, Gomez, & McKoon, 2004).

A complete assessment of model fit requires not only a quantitative comparison, but also a qualitative assessment, in which the model predictions for each condition are compared to the observed data. Quantile-probability plots averaged over all participants for the three Stroop conditions in the word identification block and color identification block are presented in Figure 2. Each column shows one experimental condition, with the .1, .3, .5, .7, and .9 RT quantiles being plotted as a function of the proportion responses (Ratcliff & Smith, 2004). Only correct responses are presented because there were too few error observations to compute quantiles. The different stimulus conditions, neutral, incongruent

and congruent, are presented in separate columns. x's represent the empirical data, while the diffusion model predictions are represented by D's , and the RTCON2 model predictions are represented by R's. A proportion scale is centered on the data and shows the 1 percent deviation in response proportion.

The predictions for both the RTCON2 and diffusion models match the data quite well. Predicted quantile RTs are within 30 ms. of the data for correct responses and predicted accuracy values are within two percentage points of the data. Both models predict shorter .9 quantiles than is observed in the data for the color identification task. In the word identification condition, it can be seen that the RT quantiles are nearly identical for correct responses across all conditions. The color identification condition on the other hand, shows a different pattern between conditions, with incongruent correct trials resulting in a slightly longer leading edge and a longer tail compared to the neutral condition. The congruency effect results in fewer responses in the tail of the RT distribution for congruent compared to neutral trials (i.e., a shorter tail). This pattern of results is consistent with results from other analyses of RT distributions for manual response variants of the Stroop task (Aarts, Roelofs, & van Turennout, 2009; Steinhauser & Hübner, 2009).

The under-prediction of the .9 quantile produces a large part of the numerical mismatch in χ^2 between the models and the data. This result is similar to that of Ratcliff and Starns (2009) who found that small shifts in the RT quantiles could produce large increases in χ^2. Furthermore, even small misses (<.1) in response proportions have been shown to produce χ^2 values double the critical value (Ratcliff et al., 2004). Given these caveats of the χ^2 fitting method, it can be seen that both models do a good job of capturing overall trends in the data with reasonable precision.

To illustrate the variability of individual responses and the ability of both the diffusion and RTCON2 models to fit this variety of data, we plotted accuracy and RT quantile data for all conditions in the color identification block. Figure 3 shows the observed accuracy and RT quantiles plotted against the predicted values for the diffusion (left) and RTCON2 (right) model. Error responses are represented by

crosses, while correct responses are unfilled circles. A perfect correspondence between the empirical data and predictions would be shown by the points lying on the line with a slope of one (indicated by the diagonal line on each plot). Error bars were constructed using a bootstrap method. A bootstrap sample was created by sampling with replacement from all of the responses in a condition for each participant. This was repeated 100 times to create the bootstrap samples which were used to generate the SDs of the RT quantiles for each participant, condition, and response option. The SDs for the three conditions were averaged across participants separately for correct and error responses to create the error bars. Error bars depicting one standard deviation are shown in the bottom right of each plot, for both errors (top) and correct responses (bottom). An error bar depicting 2 standard deviations is shown intersecting the reference line. Examination of the response proportions shows that the data and predictions from both models match each other quite well for all participants. There is some variability in the lower quantiles, but most of the differences between theory and data occur in the .9 quantiles. However, both models provide predictions that are within 2 standard deviations, except in a few cases, in the 0.9 quantile RTs.

Although the diffusion and RTCON2 models do not have the same parameters, there are some parameters that are common across models, namely drift rate, decision boundary setting, and non-decision time. These should be consistent across the different conditions for each participant. Figure 4 displays comparisons of drift rate, decision boundary and non-decision time, for all individual participants between the RTCON2 and diffusion model. The leftmost plot displays the drift rates for each participant across all conditions, with the diffusion model estimates on the y-axis, the RTCON2 estimates on the x-axis, and a best fitting linear regression line. The high correlation between the parameters demonstrates that there is a good correspondence in drift rate estimates between the two models. Therefore, when the diffusion model produces large drift rate estimates that suggest a fast rate of

information accumulation, the RTCON2 model does the same. Similar results are obtained for decision boundaries and nondecision time which are shown in the other two panels of Figure 4.

Experiment 2

The goal of this experiment was the same as that for Experiment 1. In this experiment, four colors were presented throughout the experiment (instead of two per block in Experiment 1) with two colors consistently mapped to each of two response keys across the entire experiment. Again, we will examine whether RTCON2 and the diffusion model produce similar parameters and predictions for this as in Experiment 1.

Method

Participants

Nineteen college-aged participants with normal or corrected to normal vision, were recruited from the introductory psychology course at The Ohio State University and were given course credit for an introductory psychology course upon completion.

Materials

The materials used in Experiment 2 are adapted from (De Houwer, 2003; James R Schmidt & Cheesman, 2005; Steinhauser & Hübner, 2009) and are the same as those used in Experiment 1, with a few exceptions. Neutral stimuli are color words presented in white characters on a black background for word identification blocks, and a letter string of X's presented in a color for color identification blocks (as in Experiment 1). Identical stimuli have the color words presented in the matching character color, while congruent stimuli have different color words and character colors, but they are from the pair assigned to the same response (i.e. the word red in blue font color, because red and blue are mapped to the '/' key). Incongruent stimuli have color words and character colors that are mapped to different keys (i.e. the word green in red font color, with green mapped to the 'z' key and red mapped to the '/' key). The use of the word congruent may seem confusing as this condition does entail semantic incongruity, however

we use this terminology to remain consistent with the rest of the experiments in this paper. There were 16 stimuli of each type (Identical, Congruent, Incongruent, and Neutral) per block with a total of 36 blocks, with half the blocks color identification blocks and half word identification blocks.

Procedure

The procedure was identical to Experiment 1 with a few exceptions. The most notable is that the response key mappings are slightly different, so that two colors are mapped to one response key (red and blue for the '/' key and green and yellow for the 'z' key) for all of the experiment. In addition to this, the practice block consisted of 24 trials, evenly divided among the Stroop conditions.

Analyses

Statistical analysis. RT latencies less than 250 ms. and larger than 1500 ms. were excluded from statistical analysis (less than 2% of the data was eliminated). Mean RT and accuracy were submitted to a 2 (Block) x 4 (Stroop Condition) repeated measures ANOVA. Interactions were further explored planned paired t tests. Specifically, we examined Stroop facilitation (i.e., identical minus neutral) and interference (i.e., incongruent minus neutral) for each block type.

Model Fitting. Model fitting was the same as in Experiment 1 except there were 8 conditions (word identification: congruent, incongruent, neutral, identical, and color identification: congruent, incongruent, neutral, identical) and 2 response options. Because there were 8 conditions instead of 6, there were 14 free parameters in the RTCON2 model, and 14 free parameters in the diffusion model, with a total of 74 degrees of freedom.

Results

RT and Accuracy

Mean RT and accuracy from Experiment 2 are presented in Table 1. In both blocks, participants took longer to respond to incongruent stimuli, and made more errors compared to neutral stimuli. In

the color identification block only, responses to identical stimuli were faster and more accurate when compared to neutral stimuli.

Statistical analyses show that there is a significant block x Stroop condition interaction for accuracy, $F(3 , 54) = 4.01$, $p < .05$, $\eta^2 = .03$, but not mean RT, $F(3, 54) = 2.01$, $p = .12$, $\eta^2 = .01$. There was no significant main effect of block for accuracy, $F(1, 18) = 2.76$, $p = .11$, $\eta^2 = .01$, or mean RT, $F(1, 18) = .018$, $p = .9$, $\eta^2 < .01$. However, there were differences between the incongruent versus neutral conditions for accuracy, $F(3, 54) = 29.22$, $p < .001$, $\eta^2 = .28$, and for mean RT, $F(3, 54) = 45.9$, $p < .001$, $\eta^2 = .10$. Planned pairwise comparisons indicate that for the color identification block, participants made significantly slower responses, $t(18) = 5.69$, $p < .001$, $d = 1.3$, and more errors, $t(18) = -4.15$, $p < .001$, $d = 1.4$, for the incongruent compared to neutral condition. Similarly, this interference effect was present in the word identification blocks for both mean RT, $t(18) = 4.07$, $p < .001$, $d = .93$, and accuracy, $t(18) = -5.70$, $p < .001$, $d = .68$. As with Experiment 1, there was significant Stroop facilitation in the color identification blocks, with participants producing faster responses, $t(18) = 3.87$, $p < .05$, $d = .55$, and more accurate responses, $t(18) = 3.87$, $p < .01$, $d = .53$, in the identical compared to neutral condition. This was not the case in the word identification blocks, with response latencies, $t(18) = 1.52$, $p = .15$, $d = .35$, and accuracy, $t(18) = -1.09$, $p = .29$, $d = .84$, not significantly differing between identical and neutral conditions.

Model Fits

In the section that follows, the fits of RTCON2 to those of the diffusion model are compared. We do this in the same manner as Experiment 1. The mean parameters and standard deviations for the parameters from RTCON2 and the diffusion model are presented in Tables 2 and 3 respectively. To preview, both models fit the data well, and provide similar parameter estimates. We first begin the comparison quantitatively, by looking at the results from the χ^2 statistic. There were a handful of participants who had too few error observations in both the congruent and neutral conditions for both

tasks, and so the χ^2 was based on a single value for each error response category. The mean χ^2 for the RTCON2 model is 132.1 with an SD of 25.7. This is more than the critical value for 88 degrees of freedom and $\alpha = .05$ (110.9), indicating a mismatch between the data and the model predictions. 2 out of 19 participants had χ^2 values lower than the critical value for RTCON2. Similarly the diffusion model has a mean χ^2 of 116.1 (df = 88) and SD of 26.7, which also indicates a mismatch between theory and data. 4 out of 19 participants had χ^2 values lower than the critical value for the diffusion model. However, as was discussed earlier, these values are within the range of those typically observed for the diffusion model and the values represent adequate fits.

We begin the qualitative assessment of the model fits by first examining the quantile probability plots. Figure 5 right column shows RT quantiles plotted against response proportion both averaged across all participants for the 4 Stroop conditions in the color identification block. As before, each column represents the RT distributions for the correct responses one condition. Observed data are indicated by x's, the diffusion model predictions are indicated by D's, and the RTCON2 model predictions are represented by R's. A proportion scale is centered on the data and shows a plus and minus 1 percent range in the response proportion.

Both models fit the data well. For correct responses in both tasks, the predictions of both models differed by no more than 20 ms. from observed data, across all RT quantiles. The models' predicted response proportions differed by less than 2% from the empirical data for both the color identification (right side of Figure 5) and word identification (left side of Figure 5) tasks. Overall the patterns in the data are captured quite well by the models, with both RTCON2 and the diffusion model making similar predictions.

Color naming trials produce RT quantiles with a similar pattern of results as in Experiment 1. The interference effect produces a longer tail of the RT distribution, and a slightly longer leading edge for incongruent compared to neutral trials. Identical trials result in shorter .9 quantiles but similar .1

quantiles relative to neutral trials. There is a difference between the results from Experiment 1 and 2 because the word identification blocks also produced interference. This reverse Stroop effect emerges in the .9 quantile for the incongruent compared to neutral condition, just as is the case in the color identification blocks. These results replicate those observed by Steinhauser and Hübner (2009). As in Experiment 1, differences in drift rate were mainly responsible for differences across Stroop conditions in the models. Incongruent stimuli in both color identification and word identification blocks produced lower drift rates than neutral stimuli, indicating poorer quality of evidence for incongruent stimuli.

The correspondence between parameters of the two models can be examined by plotting model parameters for the two models against each other as for Experiment 1 and by examining correlations between the main parameters of interest, drift rate, response boundary, and non-decision time. Figure 6 shows plots of these three parameters, with diffusion model parameters on the y-axis, and RTCON2 parameters on the x-axis. The plot on the left displays the drift rates of the RTCON2 and diffusion models for all participants in all conditions, with a best-fitting linear regression line. This plot demonstrates that the rate of information accumulation is in agreement between the models. The middle plot shows the decision boundary parameters for all participants for both models, with a best-fitting regression. It can be seen that there is close correspondence between the models for decision boundaries. The plot on the right in Figure 6 shows the relationship between the non-decision parameter of both models. The close clustering around the regression line suggests that both models produce similar estimates of non-decision time.

Experiment 3

Experiment 3 was designed to test the ability of RTCON2 to account for the behavioral data of four-choice manual and vocal response Stroop tasks. The experiment also allowed us to examine whether participants' pattern of results differed as a function of response modality. As was discussed in the introduction, RT distributions behave differently depending on whether participants make responses

vocally or make them manually on a keyboard. By having the same participants complete both, model parameters can be compared across tasks.

Method

Participants

Twenty eight college-aged participants with normal or corrected to normal vision, were recruited from the introductory psychology course at The Ohio State University and were given course credit for their participation.

Materials

The materials were the same as those used in Experiment 1, except there was no word identification block, and participants made responses based only on the character color. The word identification block was removed as it was not central to our research questions, and we wanted to collect more observations in each condition. This design has 4-choices, in which there were equal numbers of congruent, incongruent and neutral stimuli. Incongruent stimuli were presented in all possible color combinations (e.g. the word red in green, blue and yellow ink). Thus, out of the total 1440 trials, 480 were congruent, 480 were incongruent, and 480 were neutral stimuli. This does create a design in which a color word is more likely to be presented in a congruent font color than in one of the other incongruent for colors. For example, the word “green” will be presented in a green font color three times in order to balance presenting it in red, blue, and yellow font for the incongruent condition. Due to this balancing, the facilitation effect was subject to a contingency bias (Lorentz et al., 2016; James R. Schmidt & Besner, 2008). However, our main interest lies in modelling the Interference effect, and thus the same randomization procedure was used.

CMU sphinx 2 (Huang et al., 1993) voice recognition software was used to record and identify vocal responses. The database of recognized words was created using lmtool. Lmtool is software that takes user input words and builds a set of lexical and language decoder files which CMU sphinx 2 uses to

decipher responses. The corpus of recognized words we used was restricted to only the words “red”, “blue”, “green”, and “yellow”. For the vocal response session, a research assistant recorded the responses manually in order to verify the accuracy of responses produced by the speech recognition software.

Procedure

Participants completed two sessions each lasting 50 minutes. One session was a manual four-choice Stroop task, and the other was a four-choice vocal response Stroop task. Order of sessions was counterbalanced across participants. For the vocal response session, the microphone responsiveness was calibrated by having the participant say the 4 color words 5 times. If accuracy of the identification of the words was less than 95 percent, adjustments were made and the participant completed this calibration phase again. This was repeated until accuracy was 95 percent or greater. All participants were able to achieve this accuracy, so no participants were discarded. For the vocal response task, participants made responses into a microphone, with “red”, “green”, “blue”, and “yellow” being the only responses accepted. An experimenter was present during the entire experiment to manually record the responses and check these against the microphone responses. If there was a discrepancy, the manual record replaced the microphone response. The manual response task was similar to the two-choice version presented in Experiment 1, except four colors were present in all blocks. Red was mapped to ‘/’, blue was mapped to ‘.’, green was mapped to ‘x’, and yellow was mapped to ‘z’. The keys were covered with color patches to aid identification by the participants. A practice block of 18 trials was completed at the beginning of the task. Participants were given feedback in the same way as in Experiment 1.

Statistical analysis. RT latencies less than 250 ms. and greater than 1500 ms. were excluded from statistical analysis (corresponding to less than 2% of the data). Mean RT and accuracy were submitted to a one-way repeated measures ANOVA, with the 3 Stroop conditions being the independent

variable. Interactions were further explored with planned paired t-tests. Specifically, we examined Stroop facilitation (i.e., congruent minus neutral) and interference (i.e., incongruent minus neutral) separately in each task.

Model Fitting. RTCON2 was fit using the same procedure described in Experiment 1, except within-trial noise was allowed to vary freely across participants. In the previous experiments, within-trial noise was held constant to provide a balanced comparison of the diffusion model and RTCON2. Since we are not conducting such a comparison, we wanted to use the model in its freest form. There were 3 conditions and 4 response options which gave 59 degrees of freedom with 10 free parameters.

Results

RT and Accuracy

Accuracy and mean RT for both tasks in Experiment 3 are presented in Table 1. The congruency effect was obtained as before, with participants taking longer to respond to incongruent compared to neutral trials in both tasks. Also, participants were faster in the congruent compared to the neutral condition for both manual and vocal response variants of the task.

There was a significant main effect of Stroop condition on RT for both the manual response task, $F(2, 28) = 45.12$, $p < .001$, $\eta^2 = .15$, and vocal response task, $F(2,28) = 97.50$, $p < .001$, $\eta^2 = .29$. For the vocal response task participants produced both significantly faster responses for congruent compared to neutral trials, $t(14) = -2.60$, $p < .05$, $d = .67$, as well as slower responses to incongruent compared to neutral trials, $t(14) = 11.05$, $p < .001$, $d = 2.85$. Similarly, for the manual response task, participants produced a significant facilitation effect, $t(14) = -3.38$, $p < .01$, $d = .87$, and congruency effect, $t(14) = 5.92$, $p < .001$, $d = 1.53$.

Statistical analysis also showed a significant main effect of Stroop condition on accuracy in the manual response, $F(2, 28) = 4.89$, $p < .05$, $\eta^2 = .05$, and vocal response tasks, $F(2, 28) = 22.72$, $p < .001$, $\eta^2 = .46$. Planned comparisons showed that accuracy did not significantly differ between incongruent and

neutral trials in the manual response task, t(14) = -1.14, $p > .05$, $d = .30$, although they were less accurate on congruent compared to neutral trials, t(14) = 2.60, $p < .05$, $d = .67$. In the vocal response task, participants made less errors for congruent compared to neutral trials, t(14) = 4.78, $p < .001$, $d = 1.23$, and more errors for incongruent compared to neutral trials, t(14) = -4.56, $p < .001$, $d = 1.18$.

Model Fits

This experiment was designed to build upon the previous experiments and explore how processing differs between manual and vocal response variants of the Stroop task. To preview, the results suggest that the main difference in processing between the two tasks lies outside the decision-making process. The RTCON2 model was able to provide adequate fits to the empirical data for the manual response task, however the model missed some important aspects of the data in the vocal response task, namely, large shifts in the leading edge of RT distributions between neutral and incongruent conditions. Results also showed weak correlations between the parameters related to the decision-making process between the manual response and vocal response tasks.

Table 4 contains the average parameter values and standard deviations derived from model fits to individual participants' data. The mean χ^2 values for this experiment were 73.1, and 84.3, with standard deviations of 41.3, and 41.0 for the vocal and manual response tasks respectively. The χ^2 value for the manual response task is greater than the critical χ^2, 77.9, with 59 degrees of freedom and $\alpha = .05$, but is within the range of values typically observed in fits of the diffusion model to data. This indicates that RTCON2 provides an adequate fit to the data in the manual response task. There were 13 out of 28 participants in the manual response task whose χ^2 value exceeded the critical value.

There were few error observations in both the congruent and neutral conditions for the vocal response task, and so the χ^2 was based on a single value for each error response category. Thus, there were 29 degrees of freedom for which the critical χ^2 was 42.6 ($\alpha = .05$), indicating a mismatch between the predictions of RTCON2 and the data in the vocal response task. The χ^2 values exceeded the critical

value in 22 of the participants. However, the χ^2 value is within the range of values typically observed in fits of the diffusion model that are accepted as adequate fits.

To examine where the mismatch in the manual and vocal response Stroop tasks comes from, the empirical response proportions, .1, .5, and .9 RT quantiles, were plotted against the model predictions, for all participants in all conditions in Figure 7. Only correct responses are shown because few errors were observed, in fact none for most participants in some conditions. A reference 1 SD error bar is presented in the bottom right of each plot, and a 2 SD error bar is shown intersecting the reference line. Points lying on the reference line indicate perfect agreement between model predictions and the observed data, while points below the line indicate model under predictions, and points above the line are model over predictions. The far left plot shows that RTCON2 predicts the .1 RT quantile for the incongruent condition (filled diamonds) to be shorter in many cases than the observed value, although the model did adequately fit the .5 and .9 quantiles. A similar pattern is observed for the manual response task, (the middle right plot) with the .1 quantile under predicted in the incongruent condition, although the misprediction is not as systematic as in the vocal response task. In the manual response task the largest mismatch between model prediction and observed data in the .1 quantile is 70 ms., whereas it is 100 ms. in the vocal response task. Mispredictions of this size are observed for 3 participants in the manual response task and 8 participants in the vocal response task.

To account for this systematic under prediction of the .1 quantile in the incongruent condition, we decided to implement a variant of RTCON2 with an additional non-decision time parameter (2-Ter model) that is free to vary in the incongruent condition (Ratcliff & Smith, 2010). This addition is inspired by a study by Ratcliff and Frank (2012) in which an extra non-decision time parameter was added to the diffusion model and was allowed to differ between conditions in order to account for large shifts in the leading edge of the RT distribution (this was one of two possibilities in their article). This modification was based on constraints provided by a neurally plausible basal ganglia model (Frank, 2005, 2006). The

extra non-decision time parameter merely reflects extra processing time occurring outside of the decision process during the incongruent condition (and might be seen as ad hoc in this application).

The model parameters for RTCON2, as well as the 2-Ter model, averaged across all participants for each task, are shown in Table 4. The χ^2 values for the 2-Ter model are 50.6, and 75.5, with 28 and 58 degrees of freedom for the vocal response and manual response tasks respectively. The nested model comparison indicates that allowing non-decision time to vary in the incongruent condition offers a significant improvement over the initial RTCON2 model in both the vocal response task, χ^2 (1) = 22.5, $p < .01$, and manual response task, χ^2 (1) = 8.8, $p < .01$.

For the vocal response task, 14 participants had χ^2values that exceeded the critical value, with the 2-Ter model compared with 22 for RTCON2 (1-Ter model). A smaller improvement in fit was seen for the manual response task with 10 participants with χ^2 values that exceeded the critical value in the 2-Ter model compared with 13 for the original model.

Generally, the parameters were consistent across 1-Ter and 2-Ter RTCON2 models. In the 2-Ter model, there was a difference of 52 ms. between the two Ter values in the vocal response task. This allowed the model to capture the shifts in the .1 RT quantile in the incongruent condition that the 1-Ter model under predicted. A plot of model predictions against the empirical data for the 2-Ter model is presented in Figure 7, for both the vocal and manual response tasks. Both models provide predictions of response proportions that closely correspond to the empirical data; the points are closely clustered on the reference line. With the addition of an extra non-decision time parameter, the 2-Ter model provides predictions across all RT quantiles that are in close agreement with the data in both the vocal and manual response tasks.

To examine whether the model analyses suggest that the processes involved in decision-making are consistent across the two response modalities, we examined the correlations of model parameters across the vocal and manual response tasks for the 1-Ter and the 2-Ter models. Scatter plots of the

parameters in the 1-Ter RTCON2 model are presented in Figure 8. Table 5 contains the parameter correlations across both tasks for the 1-Ter and 2-Ter RTCON2 models. The models have similar correlations of boundary separation, drift rate, non-decision time, and across-trial variability in non-decision time. However, in the 2-Ter model, the correlation of across-trial variability in drift is much higher relative to the 1-Ter model. The average correlations across individuals were .05 for drift rate, .60 for boundary separation, and .08 for non-decision time. Thus, boundary setting is the only parameter with a correlation greater than the critical value for 26 degrees of freedom, and $\alpha = .05$ (.37). This suggests that participants who were conservative on one task, were conservative on the other as well. The lack of correlation of non-decision time across the two tasks, indicates that individuals had different amounts of processing time devoted to encoding, translation, and response execution across the tasks. Similarly, evidence used in the decision process in the vocal response task was not related to that in the manual response task (drift rates were not correlated across tasks). This then raises the question of what it is about vocal responding that makes processing so different from manual responding, given the similarity of the tasks.

Experiment 4

The purpose of this experiment was to examine if a limitation on the number of motor response options resulted in the additional interference observed in the vocal response Stroop task. In the vocal response task, there is one motor response pathway that a vocal response must traverse, while in a manual response version, there are multiple motor response pathways, one for each finger and hand used in the task (e.g., two fingers on each of two hands for Experiment 3). It is possible that having only one response pathway creates a bottleneck in which competing information creates more interference than if there are multiple response output options. To address this, we used a manual task in which a single finger was used to make the response. This used a touch screen in which one finger on one hand is used to make all of the responses. If the number of motor response pathways is an important factor in

Stroop interference, a pattern of behavioral data similar to that in the vocal response Stroop task should emerge in this task. On the other hand, a pattern of RTs and error rates as observed in the manual response task, would indicate that the vocal response system produces additional interference in a manner that is different from manual tasks.

Participants

Twenty college-aged right handed participants with normal or corrected to normal vision were recruited from the introductory psychology course at The Ohio State University and were given course credit for their participation.

Materials

The same stimuli and randomization as in Experiment 1 was used. Responses were collected using a 17-inch CRT with serial resistive touchscreen (Elo-Touchsystems screen), which consists of a resistive coating and a coversheet with conductive coating. When the screen is touched, the flexscreen makes contact with the glass' coating, which generates an ultra-sonic wave, and a response is recorded.

Procedure

This task lasted approximately 50 minutes. As in Experiment 3, participants made responses based upon the character color of the stimulus for the entire task. A practice block consisting of 24 trials was completed to make sure the participant understood the task. Response options were presented on screen as four color squares (red, green, blue, and yellow) in a 180 degree arc around a white square. Each response option was 7.5 cm away from the white square, with the color squares being equally spaced from each other (at 0, 60, 120, and 180 degrees). The order and position of these color squares was fixed across blocks and participants. Participants were instructed to press and hold the white square with their right index finger to make a stimulus appear. They were also instructed to wait to lift their finger until a decision had been made. When the decision was made, participants were to lift the finger and move it to touch the color square corresponding to their choice. A "TOO SLOW PLACING" message

was displayed for 500 ms. if the time from lifting the finger from the white square to placing it on the color square was greater than 400 ms. This was done to ensure that responses were ballistic and the movement time was separate from the time allocated to decision-making processes. A "TOO SLOW LIFTING" message was displayed for 500 ms. if a participant held their finger on the white square for more than 1500 ms. If the finger was lifted earlier than 200 ms. after presentation of the white square, a "TOO FAST" message was displayed for 1500 ms. Error feedback was given in the form of an "ERROR" message displayed for 500 ms. in the center of the screen. RTs were the amount of time the finger was held on the white square before lifting it to move it to the color square corresponding to their choice.

Analysis

Statistical analysis. The movement time between lifting the finger off the white square and placing it on the color square was examined to ensure responses were ballistic. If there were any significant deviations in mean movement time between the various conditions (>40 ms.), then the plan was to discard that participant's data. But this did not occur, which indicated participants followed instructions. Statistical analyses are the same as those carried out in Experiment 3.

Model Fitting. Model fitting was the same as described in Experiment 3.

Results

RT and Accuracy

Proportion correct and mean RT for Experiment 4 is presented in Table 1. Participants were not any faster or more accurate for congruent compared to neutral trials. However, interference was observed with slower and less accurate responses for the incongruent compared to neutral condition.

There was a main effect of Stroop condition on accuracy, $F(2, 38) = 11.05$, $p < .001$, $\eta^2 = .26$, with participants producing significantly more errors on incongruent compared to neutral trials, $t(19) = -3.43$, $p < .01$, $d = .77$. However, error rates did not significantly differ between congruent and neutral trials, $t(19) = .37$, $p > .05$, $d = .08$. Stroop condition also produced a main effect on mean RT, $F(2, 38) = 11.95$, p

< .001, η^2 = .01. As was the case with accuracy, a significant interference effect emerged, t(19) = 3.43, *p* < .05, *d* = .77, but facilitation was not present, t(19) = -1.14, *p* > .05, *d* = .26.

Model Fits

In this section we address whether limiting motor response output options results in more interference, as is observed in the vocal Stroop task. Best fitting model parameters for individuals are presented in Table 4. As was the case with the vocal response Stroop task in Experiment 3 there were few error observations in all conditions, so only one value was used for the χ^2 estimation for each of the error response options. The mean χ^2 value is 30.8 with a SD of 18.6, which is larger than the critical value, 23.7, with 14 degrees of freedom and α = .05. Thus, the model generates predictions that mismatch the observed data. This is the case for 11 of the 20 participants. As for the previous experiments, although the χ^2 values were larger than the critical value, it was within an acceptable range typically observed with the two-choice diffusion model.

The best fitting parameter values for each participant are used to generate predicted RT quantiles and response proportions for each condition. A qualitative comparison can be made by comparing these predicted values to the empirical data. Figure 9 shows the quantile response times plotted as a function of response proportions for each condition averaged over the participants. The x's represent the empirical data and R's to the model predictions. The model produces response proportions and RT distributions that closely correspond to the data, aside from some slight misses in the .9 quantile. Of central interest is how RT distributions in this task relate to those in the vocal and manual response Stroop tasks. The RT distributions in this task are similar to those seen in Experiment 1, except participants are more accurate in this task. There is little interference (slowing) observed in the .1 quantile, with most occurring in the .9 quantile. This is in contrast to the vocal response task, in which interference was mainly present in the .1 quantile. RTCON2 can capture the change in the RT distribution in the experiment with a change in drift rate across conditions. As was the case in

Experiments 1-3, drift rate was smaller for the incongruent condition compared to the neutral condition for most participants.

Discussion

This study was conducted with two main aims. The first was to determine if the diffusion model and RTCON2 could account for data patterns observed in the Stroop task. The second was to understand how information is processed in the manual and vocal response variants of the Stroop task. In Experiments 1 and 2, we fit the diffusion model and RTCON2 model to the empirical data and these provided predictions for response proportions and correct and error RT distributions. Both models were able to provide adequate fits to the data. In addition to this, the best fitting parameter estimates for the two models correlated strongly across participants. For Experiments 3 and 4 we used the RTCON2 model because the two-choice diffusion model is not able to account for data from four-choice tasks. The RTCON2 model fit the data from the manual response task in Experiment 3 quite well, but fit the vocal response task in Experiment 3 somewhat poorly with systematic misses in the 0.1 quantile RTs. Conflict in this task produced a large shift in the leading edge of RT distributions for incongruent compared to neutral trials for most participants and the model was unable to fit this pattern. We modified the RTCON2 model to allow an additional non-decision time parameter for the incongruent condition, and this model provided much better fits to the quantile RTs and response proportions. The RTCON2 model fit the data from the touch screen task in Experiment 4 quite well. The pattern of results was similar to that observed in the manual response task in Experiment 3 with smaller interference effects than those of the vocal response task. Conflict in the majority of these experiments was explained by differences in the rate of evidence accumulation between conditions. Specifically, there was a smaller drift rate for incongruent compared to neutral stimuli across all experiments. The only exception was the vocal response task in Experiment 3, in which conflict occurred outside of the decision process.

A second purpose of Experiment 3 was to determine how processing is related between vocal response and manual response Stroop tasks. RTCON2 provides parameter estimates for each participant, and these allowed individual differences in components of processing to be compared across the two tasks. We found that decision boundary values, but not drift rate or non-decision time were correlated across these tasks. Thus, individuals maintain similar decision boundaries for the amount of information required for a decision across tasks. However, an individual who had a small drift rate on the vocal response Stroop task did not necessarily have a small drift rate on the manual response Stroop task. Similarly, an individual who had a long non-decision time on the vocal response Stroop task did not necessarily have a long non-decision time on the manual response Stroop task.

This RTCON2 model had difficulty accounting for the large systematic differences in the .1 quantile between the neutral and incongruent conditions that were observed in the data. To account for these shifts, we used a second RTCON2 model (2-Ter model) with an additional non-decision time parameter for the incongruent condition. With this modification, results showed similar patterns of correlations of drift rate, non-decision time, and decision boundary across response modalities as for the 1-Ter model. Given the agreement in individual differences in model parameters between these two models, the results suggest that participants process information differently in the two tasks. The lack of correlation of drift rates across the two Stroop modalities is surprising because correlations in drift rates have been observed between quite different simple cognitive tasks such as numerosity discrimination, recognition memory, and lexical decision (Ratcliff et al., 2010). These are arguably more different from one another than the manual and vocal response Stroop tasks. Results also showed that accuracy was lower in the manual response task, and RT distributions differed across the tasks, supporting the notion that there are processing differences depending on whether a response is given manually or vocally.

Experiment 4 examined the possibility that the difference between these tasks is due to a limited number of response pathways in a vocal versus manual response task. The experiment had

participants respond on a touch screen with only one hand. The pattern of empirical data was similar to that observed in the other manual response Stroop tasks, suggesting that the difference in the processing of conflicting information in a vocal versus manual response Stroop task is not due to limited response output options.

In the introduction of this article we described models that provide different theoretical perspectives on Stroop phenomenon. RACE/A is a production rule model that incorporates sequential sampling models into the architecture of ACT/R. Within this structure, processing is divided across a number of stages (e.g. visual, procedural, vocal, and retrieval) and the model is able to simulate response proportions and mean RT. Although it has not been explicitly tested, RACE/A could possibly account for the modality effect by allocating all of the additional interference of the vocal response Stroop task to the response stage. This would suggest that non-decision time is the main difference between the two tasks. This is in agreement with findings from a study by Gomez, Ratcliff, and Childers (2015) in which they fit the diffusion model to data from a letter discrimination task that manipulated response modality (key press, eye movement, and touch screen). They found that non-decision time changed as a function of response modality, while the other parameters related to the decision-making process did not. However, the results presented here suggest that the differences between the response modalities were due to more than just nondecision time and that there are other differences in processing which suggests that this simple version of RACE/A would be inadequate.

As was mentioned previously, there important differences between the vocal response and manual response Stroop tasks. The most notable are that non-decision time was longer in the manual response task than in the vocal response task, and drift rate did not correlate across the two response modalities. This suggests that differences in conflict processing between the vocal response and manual response tasks in part lies outside of the decision-making process. This is in line with the other version of

the RACE/A model we described earlier, which predicts differences across response modalities would be due to differences in both non-decision and decision processes.

The Cohen et al. (1990) model is a feedforward multilayer neural network that accounts for Stroop interference word naming by modeling it as a more automatic process than color naming. This is done by using separate nodes for word information and color information. Information travels from these nodes through a hidden layer, and then to an output/response layer. Word information is weighted more heavily than color information and so it is processed more quickly than color information resulting in conflict when incongruent stimuli are presented. The model initially did not capture the modality effect, but with the addition of a node that represents the response modality in which the task is presented, the model would be able to account for mean RT differences as a function of response modality. This would attribute differences in conflict processing between these tasks to vocal responding being a more automatic process than manual responding. The main shortcoming of this model is that the RT distributions it produces are of an incorrect shape (Mewhort et al., 1992).

In a similar vein, RACE/A only provides mean RT estimates. Although it can produce RT distributions, an analysis of RT distribution predictions has yet to be done. The third model, the dimensional overlap model, can produce RT distributions, but not response proportions. Finally, none of these models have been shown to provide parameter estimates for single participants. In the analyses presented above, without such information about individual differences, it would not have been possible to determine whether individuals process information similarly across response modalities. Thus, these previous theoretical accounts are incomplete. At this point, our results are best explained in the context of RTCON2 and the two-nondecision time extension.

One possible explanation for the difference in processing between the vocal and manual response Stroop task is that processing is not purely stimulus driven. Instead, processing is influenced by the goal of the task. In a vocal response Stroop task, a participant is naming the color, whereas in a

manual response Stroop task, the task involves categorizing the color into one of several response options. Given that these tasks require different processes, it is likely that the interaction of reading and color identification processes differ according to the task. To further explain, when an incongruent stimulus is presented, the sublexical information from the word is processed and this will interfere with vocalizing a response, but not when a keyboard response is required. Thus, an assumption is that information travels down different pathways as a function of whether the response is manual or vocal. This view is supported by results from a recent study by Kinoshita, De Wit, and Norris (2016) in which RT distributions were found to differ between various neutral stimuli and incongruent stimuli in both manual and vocal response Stroop tasks. They found results similar to ours from Experiment 3, with the difference in RT distributions between incongruent and neutral stimuli, differing according to response modality. In addition to this, they also found that neutral stimuli that are more "wordlike" produced more interference in the vocal response task, but not the manual response task. Similarly, conflict increased as a function of length of words and pseudowords in the vocal response task, but not the manual response task. These analyses were based on aggregate data, so there was no analysis of individual differences as was the case in our study. However, their results provide further support for the notion that information is processed differently in these two tasks.

In addition to behavioral experiments, there are also studies using EEG that have identified neural differences in processing according to response modality. The most common neural EEG marker in Stroop task processing is the N450. This ERP component has been attributed to semantic incongruity and is prominent in the Stroop task. Specifically, this component is more negative for incongruent compared to neutral trials and is present in both manual and vocal response variants of the Stroop task (Badzakova-Trajkov, Barnett, Waldie, & Kirk, 2009; Larson, Clayson, & Clawson, 2014; Liotti, Woldorff, Perez Iii, & Mayberg, 2000). The N450 for the manual response task is located over medial electrodes (Badzakova-Trajkov et al., 2009; Liotti et al., 2000), whereas it is located at frontal electrodes for the

vocal response task (Liotti et al., 2000; Rebai, Bernard, & Lannou, 1997). Source localization methods have determined that the N450 is generated by the anterior cingulate cortex (Badzakova-Trajkov et al., 2009), although there are likely different neural generators in the anterior cingulate cortex for the manual versus vocal response N450 (Liotti et al., 2000). The spatial separation of this component across response modality suggests that processing conflicting information takes a different trajectory depending on whether vocal responses or manual responses are to be made, providing a neural context for our current results.

RTCON2 provides a theoretical account of Stroop phenomenon which is based on all aspects of the empirical data unlike previous models that provide mean RTs (RACE/A), RT distributions of the wrong shape (Cohen et al. model), or no information about accuracy (Dimensional overlap model). In doing so, RTCON2 provides a detailed account of the decision process in the Stroop task. Furthermore, it provides fits to the individual participant data, which can be used in examining individual differences. As a result, the model provides meaningful parameters that are associated with how information is processed during decision-making. This enabled us to see that individuals set a similar decision boundary for the amount of information required before making a decision across response modalities, even though their encoding, motor responses, and rate of information accumulation differed across participants for vocal and manual response tasks.

The RTCON2 model produces response proportions, as well as correct and error RT distributions across all conditions. It usually must do this by only allowing drift rate to vary across conditions and this produces the different shapes of RT distributions across conditions. A decrease in drift rate, such as is the case between neutral and incongruent trials, results in a slight slowing of responses in the leading edge of the distribution, and substantially more slowing in the tail. For the data from the vocal response task in Experiment 3, we observed a large slowing of responses in the leading edge of the distribution, in addition to more slow responses in the tail. Even with across-trial variability in non-decision time (s_t),

which can shift the leading edge of RT distributions by as much as 10% of s_t for large values of drift rate (Ratcliff & Tuerlinckx, 2002), RTCON2 is unable to provide adequate predictions for the incongruent RT distributions in the vocal response task. On the other hand, in the manual response Stroop task, there is only a small slowing of responses in the leading edge of the distribution with many more slow responses in the tail of the distribution for incongruent compared to neutral trials. This is an effect that can be accommodated by changes in drift rate across conditions, which is why the model predictions in the manual response tasks are much closer to the empirical data.

Ratcliff and Frank (2012) fit the diffusion model to simulated data from a model of the corticostriatal network and reinforcement learning in the basal ganglia system (Frank, 2005, 2006). Data showed large shifts in the leading edge of RT distributions for conflict conditions and the application of the diffusion model had the aim of determining whether allowing drift rates to vary as a function of conflict was enough to capture conflict effects, or whether other assumptions would be needed such as collapsing decision boundaries instead of fixed boundaries. They also fit the diffusion model to empirical data to examine whether differences in drift rates could explain conflict effects. Experimental results were similar to those from the vocal response task in Experiment 3, in which a high conflict condition resulted in a large shift in the leading edge of the RT distribution compared to the low conflict condition (Ratcliff & Smith, 2010). The diffusion model was not able to accommodate this shift with drift rate only varying across conditions. The authors added a second non-decision time parameter that varied as a function of conflict and they also created another diffusion model in which the decision boundaries collapsed over time in the high conflict condition (this was derived from the Frank basal ganglia model). These alterations allowed the model to account for this shift in the leading edge of the RT distributions, as well as provide a quantitatively better fit. These alterations are motivated by the neurally plausible operations within the basal ganglia model. This approach is similar to our modifications to the RTCON2 model that incorporated a second non-decision time parameter. This modification allowed the RTCON2

model to successfully account for data in the vocal response Stroop task. However, to fully understand the differences in processing between these two response modalities, a model of vocal responding must be developed.

This study provides a theoretical account of interference in the Stroop task using the RTCON2 multichoice decision model. The model accounts for response proportions as well as correct and error RT distributions and is able to do this for individual participant data. Although RTCON2 in its current form is unable to account for a large shift in the leading edge of RT distributions across incongruent and neutral conditions in the vocal responding task, adding an additional non-decision time parameter that differs as a function of conflict offers one solution to this issue. This new model points to a fundamental difference in processing that no model currently explains; however, the new model does not explain why the difference occurs. RTCON2 provides fits to empirical data comparable to that of the two-choice diffusion model, and provides parameter estimates that are consistent between the models. Our results also show that processing in the Stroop task differs as a function of task demands, namely manual versus vocal responding. Thus, RTCON2 offers a method in which to decompose behavioral data and discriminate between decision related processes and non-decision processes, and in turn provides a more complete view of processing in the Stroop task.

References

Aarts, E., Roelofs, A., & van Turennout, M. (2009). Attentional control of task and response in lateral and medial frontal cortex: brain activity and reaction time distributions. *Neuropsychologia, 47*(10), 2089-2099.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Badzakova-Trajkov, G., Barnett, K. J., Waldie, K. E., & Kirk, I. J. (2009). An ERP investigation of the Stroop task: The role of the cingulate in attentional allocation and conflict resolution. *Brain Research, 1253*, 139-148. doi: http://dx.doi.org/10.1016/j.brainres.2008.11.069

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., . . . Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron, 60*(6), 1142-1152.

Brown, S. D., Ratcliff, R., & Smith, P. L. (2006). Evaluating methods for approximating stochastic differential equations. *Journal of mathematical psychology, 50*(4), 402-410.

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.

De Houwer, J. (2003). On the role of stimulus-response and stimulus-stimulus compatibility in the Stroop effect. *Memory & Cognition, 31*(3), 353-359.

Dell'Acqua, R., Job, R., Peressotti, F., & Pascali, A. (2007). The picture-word interference effect is not a Stroop effect. *Psychonomic Bulletin & Review, 14*(4), 717-722.

Fagot, C., & Pashler, H. (1992). Making two responses to a single object: Implications for the central attentional bottleneck. *Journal of Experimental Psychology: Human Perception and Performance, 18*(4), 1058.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience, 17*(1), 51-72.

Frank, M. J. (2006). Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks, 19*(8), 1120-1136.

Gomez, P., Ratcliff, R., & Childers, R. (2015). Pointing, looking at, and pressing keys: A diffusion model account of response modality. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1515.

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: a model of letter position coding. *Psychological Review, 115*(3), 577.

Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., Lee, K.-F., & Rosenfeld, R. (1993). The SPHINX-II speech recognition system: an overview. *Computer Speech & Language, 7*(2), 137-148.

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience, 9*(5), 690-696.

Kinoshita, S., De Wit, B., & Norris, D. (2016). The Magic of Words Reconsidered: Investigating the Automaticity of Reading Color-Neutral Words in the Stroop Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: 10.1037/xlm0000311

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

Kornblum, S., Stevens, G. T., Whipple, A., & Requin, J. (1999). The effects of irrelevant stimuli: 1. The time course of stimulus–stimulus and stimulus–response consistency effects with Stroop-like stimuli, Simon-like tasks, and their factorial combinations. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 688.

Larson, M. J., Clayson, P. E., & Clawson, A. (2014). Making sense of all the conflict: A theoretical review and critique of conflict-related ERPs. *International Journal of Psychophysiology, 93*(3), 283-297. doi: 10.1016/j.ijpsycho.2014.06.007

Liotti, M., Woldorff, M. G., Perez Iii, R., & Mayberg, H. S. (2000). An ERP study of the temporal course of the Stroop color-word interference effect. *Neuropsychologia, 38*(5), 701-711. doi: http://dx.doi.org/10.1016/S0028-3932(99)00106-2

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology/Revue canadienne de psychologie, 39*(2), 367.

Lorentz, E., McKibben, T., Ekstrand, C., Gould, L., Anton, K., & Borowsky, R. (2016). Disentangling genuine semantic stroop effects in reading from contingency effects: On the need for two neutral baselines. *Frontiers in Psychology, 7*.

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

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.

Mewhort, D., Braun, J., & Heathcote, A. (1992). Response time distributions and the Stroop task: A test of the Cohen, Dunbar, and McClelland (1990) model. *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 872.

Neill, W. T. (1977). Inhibitory and facilitatory processes in selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 3*(3), 444.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308-313.

Posner, M. I., & Snyder, C. R. (1975). *Attention and Cognitive Control*. Hillsdale, NJ: Erlbaum.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59.

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88*(6), 552.

Ratcliff, R. (2018). Decision making on spatially continuous scales. *Psychological Review, 125*(6), 888.

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural Computation, 24*(5), 1186-1229.

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

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*(2), 333.

Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: the temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General, 139*(1), 70.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*(1), 59.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. *Psychological Review, 120*(3), 697.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging, 19*(2), 278.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology, 60*(3), 127-157.

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General, 140*(3), 464.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9*(3), 438-481.

Rebai, M., Bernard, C., & Lannou, J. (1997). The Stroop's Test Evokes A Negative Brain Potential, the N400. *International Journal of Neuroscience, 91*(1-2), 85-94. doi: 10.3109/00207459708986367

Redding, G. M., & Gerjets, D. A. (1977). Stroop effect: Interference and facilitation with verbal and manual responses. *Perceptual and Motor Skills, 45*(1), 11-17.

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. doi: 10.1037/0278-7393.34.3.514

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

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

Spieler, D. H., Balota, D. A., & Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 506.

Steinhauser, M., & Hübner, R. (2009). Distinguishing response conflict and task conflict in the Stroop task: evidence from ex-Gaussian distribution analysis. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1398.

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

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological Review, 120*(1), 1.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review, 108*(3), 550.

van Maanen, L., & van Rijn, H. (2007). An accumulator model of semantic interference. *Cognitive Systems Research, 8*(3), 174-181. doi: 10.1016/j.cogsys.2007.05.002

van Maanen, L., van Rijn, H., & Borst, J. P. (2009). Stroop and picture—word interference are two sides of the same coin. *Psychonomic Bulletin & Review, 16*(6), 987-999.

van Maanen, L., van Rijn, H., & Taatgen, N. (2012). RACE/A: An architectural account of the interactions between learning, task control, and retrieval dynamics. *Cognitive Science, 36*(1), 62-101. doi: 10.1111/j.1551-6709.2011.01213.x

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13*(1), 37-58.

Voskuilen, C., & Ratcliff, R. (2016). Modeling confidence and response time in associative recognition. *Journal of Memory and Language, 86*, 60-96. doi: 10.1016/j.jml.2015.09.006

White, B. W. (1969). Interference in identifying attributes and attribute names. *Perception & Psychophysics, 6*(3), 166-168.

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.

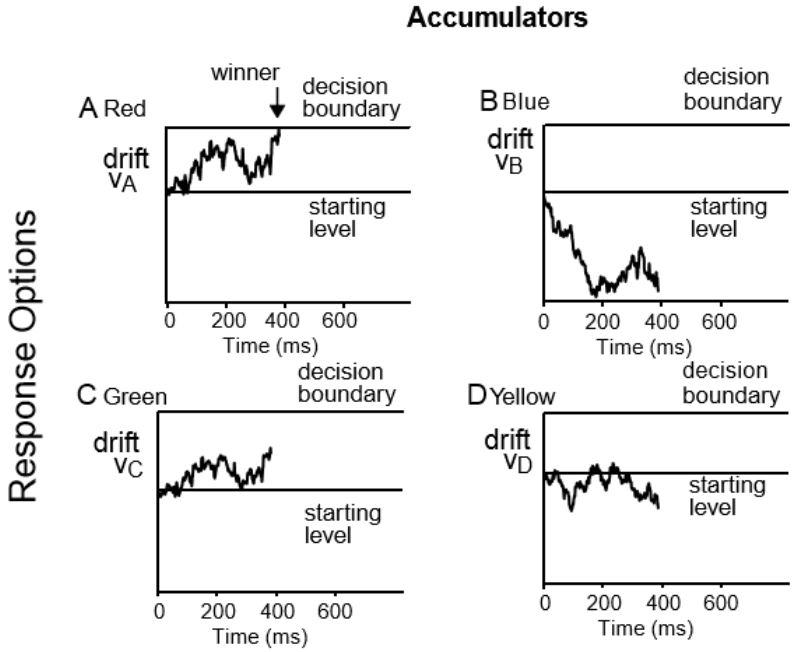

Figure 1. The plots under "Accumulators" show paths of evidence accumulation in the accumulators. Drift rates for the various response options are calculated using constant summed evidence, in which evidence for the winning drift rate is equal evidence against the other response options. For example, v_A is winning, so the drift rates for the other three response options equal $(1-v_A)/3$.

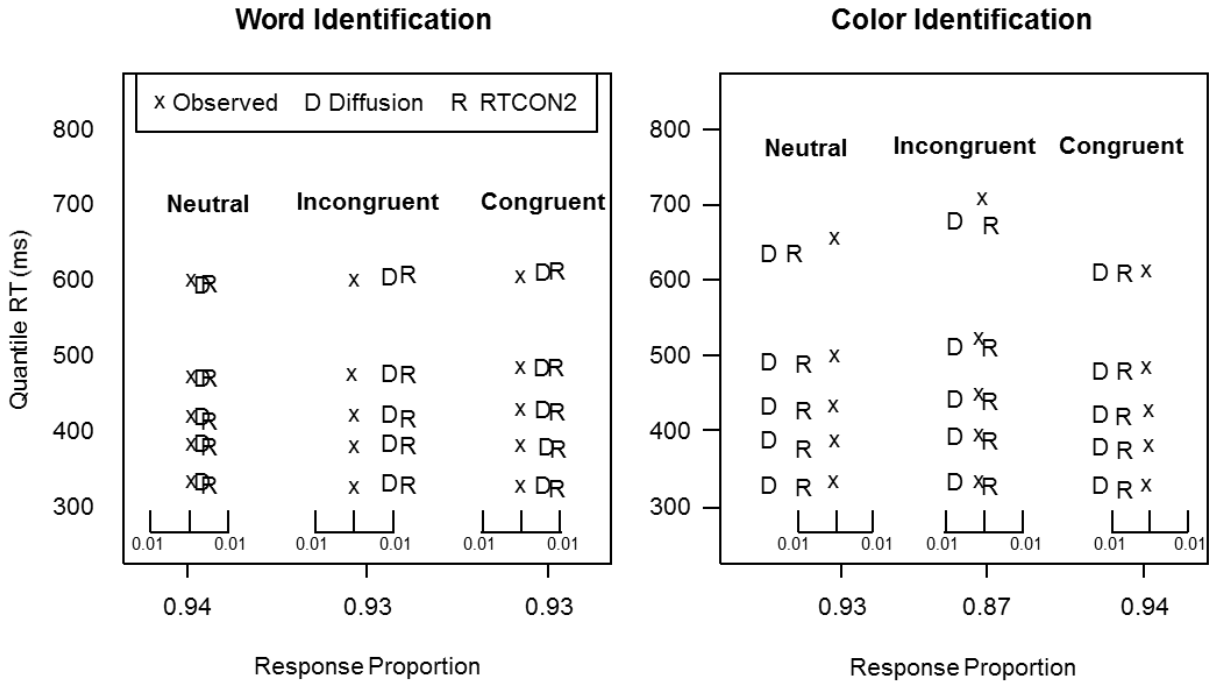

Figure 2. Quantile probability plots averaged over all participants for the word identification block (left) and color identification block (right) for Experiment 1. Empirical data is represented by x's, diffusion model predictions are represented by the letter D, and RTCON2 predictions are represented by the letter R. The .1, .3, .5, .7, and .9 quantiles are plotted in vertical columns as a function of response proportion. Only correct responses are plotted. Data is divided into 3 different columns according to the different stimulus types, neutral, incongruent and congruent. A proportion scale centered on the data shows the 1 percent deviation in response proportion.

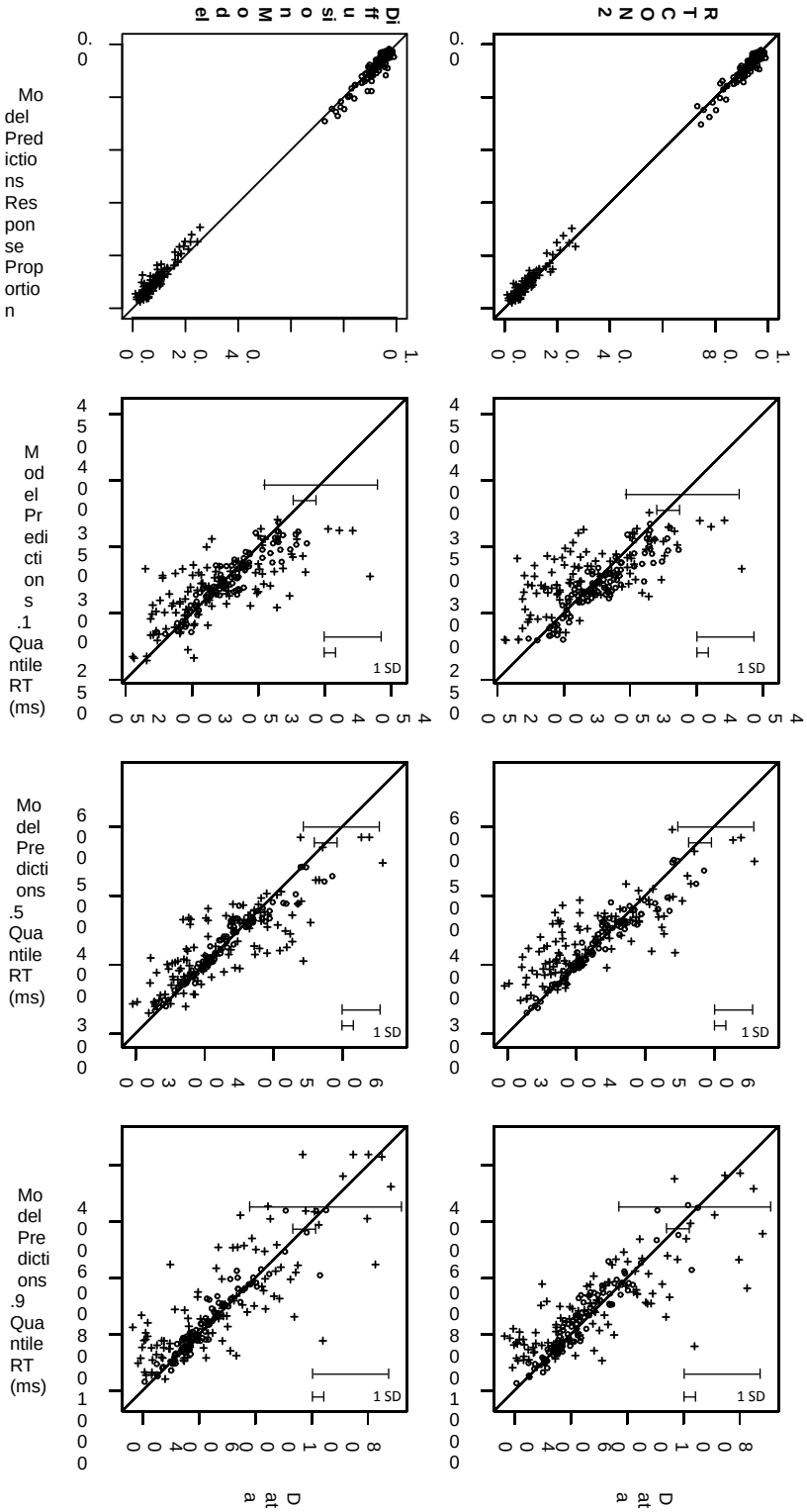

Figure 3. Empirical response proportions and quantile RTs for all individuals in all conditions in Experiment 1 plotted against the diffusion model predictions (left) and RTCON2 model predictions (right), with a reference line with a slope of 1 and intercept of 0. Error responses are crosses and open circles are correct responses. An error bar depicting one SD is presented in the lower right corner for incorrect (top)

and correct (bottom) responses. An error bar depicting two standard deviations is shown intersecting the reference line for error (top) and correct (bottom) responses. Each plot contains an average of 90 points.

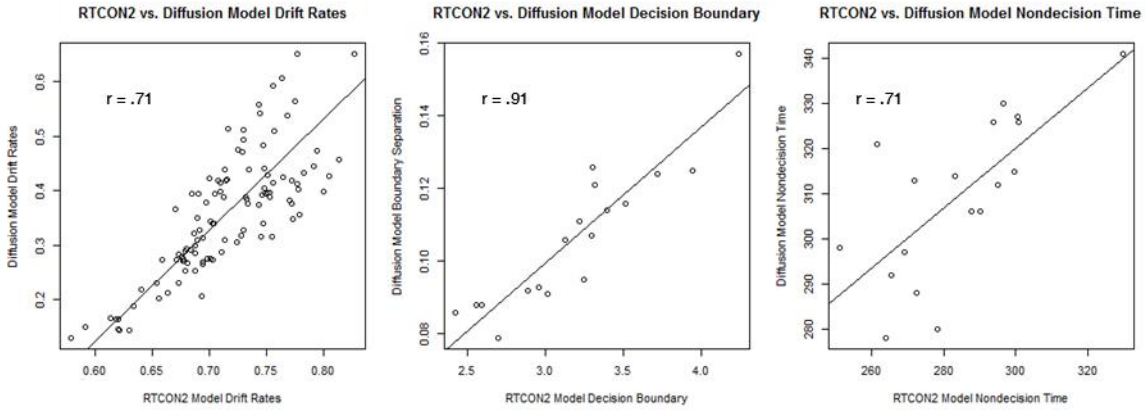

Figure 4. Comparison of drift rates (left), decision boundaries (middle), and non-decision time (right) for RTCON2 and the diffusion model for individual participants for Experiment 1. The diagonal line is a best-fitting linear regression line. Decision boundaries for RTCON2 were summed together to generate a single value. Pearson's correlation is presented in the upper left of each plot.

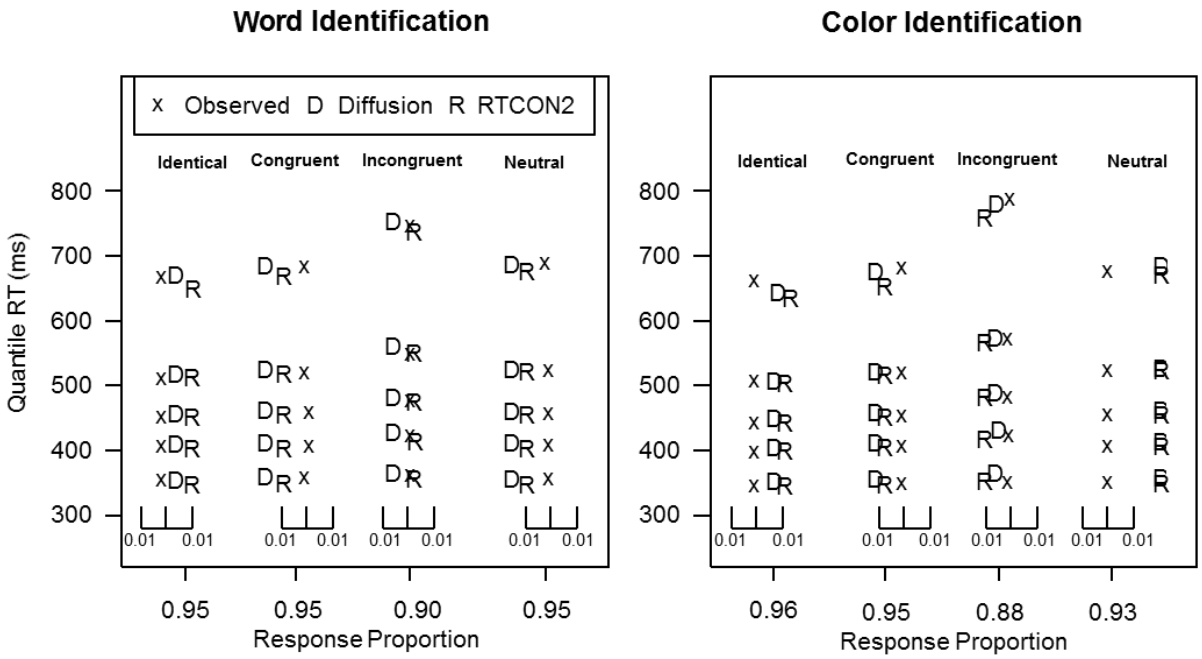

Figure 5. Quantile probability plots averaged over all participants for the word identification block (left) and color identification block (right) for Experiment 1. Empirical data is represented by x's, diffusion model predictions are represented by the letter D, and RTCON2 predictions are represented by the letter R. The .1, .3, .5, .7, and .9 quantiles are plotted in vertical columns as a function of response proportion. Only correct responses are plotted. Data is divided into 4 different columns according to the different stimulus types, identical, congruent, incongruent, and neutral. A proportion scale is centered on the data and shows the 1 percent deviation in response proportion.

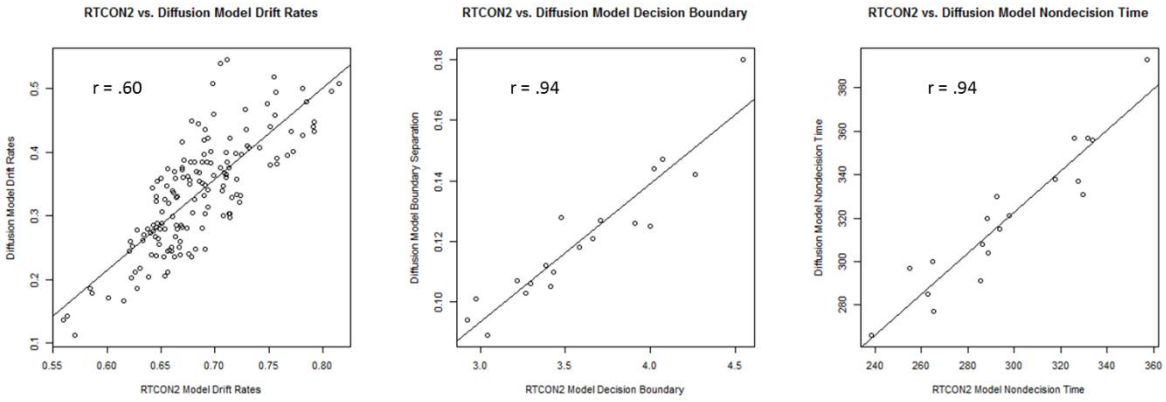

Figure 6. Comparison of drift rates (left), decision boundaries (middle), and non-decision time (right) for RTCON2 and the diffusion model (with a best fitting regression line) for Experiment 2. Decision boundaries for RTCON2 were summed together to generate a single value. Pearson's correlation is presented in the upper left of each plot.

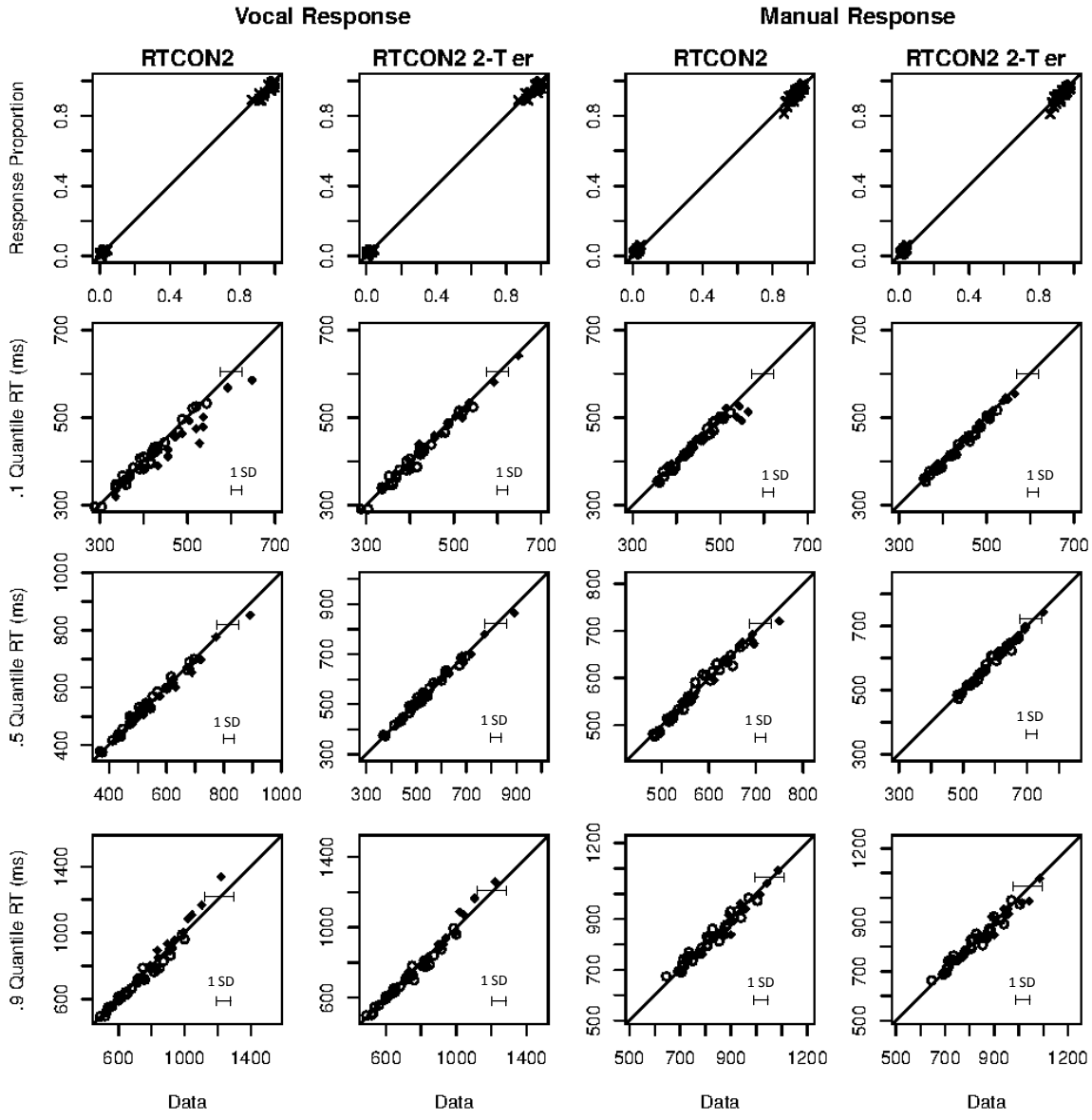

Figure 7. Empirical response proportions, .1, .5, and .9 quantile RTs for all individuals in all conditions in the vocal response (left) portion of Experiment 3 plotted against RTCON2 (far left) as well as the 2-Ter variant (middle left) model predictions, and the manual response (right) portion of Experiment 3 for RTCON2 (middle right) and the 2-Ter(far right) variant. Data presented is for correct responses only, with a reference line with a slope of 1 and intercept of 0. An error bar of one standard deviation for each RT quantile is presented in the bottom left of each plot, and an error bar of 2 standard deviations is shown intersecting the reference lines. Congruent and neutral trials are open circles and incongruent trials are filled diamonds.

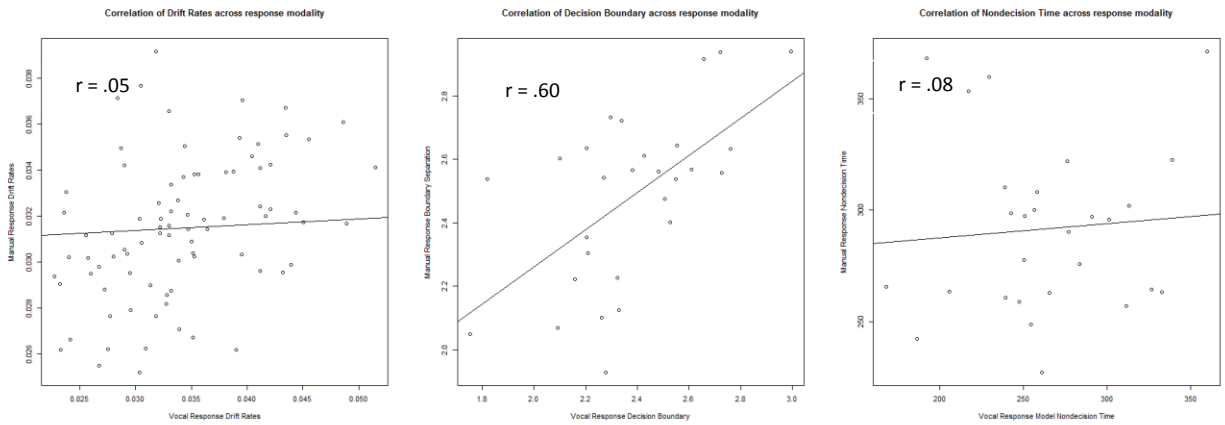

Figure 8. Scatter plots of drift rate for each participant in each condition (left), boundary separation for each participant (middle), and nondecision time for each participant (right) for the 1-Ter RTCON2 model across four-choice manual response and vocal response Stroop tasks, with a best-fitting regression line. Pearson's correlation is presented in the upper left of each plot.

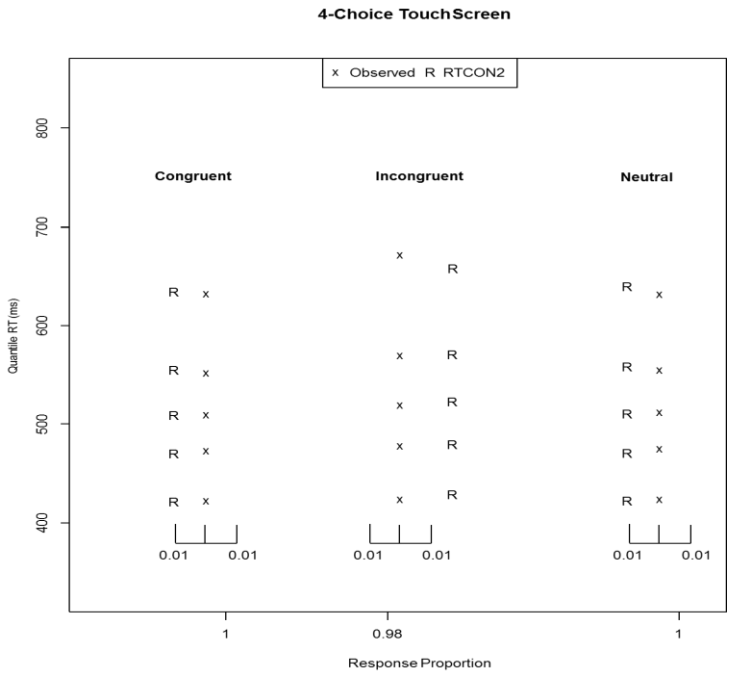

Figure 9. Quantile probability plots averaged over all participants for Experiment 4. Empirical data is represented by x's, and RTCON2 predictions are represented by the letter R. The .1, .3, .5, .7, and .9 quantiles are plotted in vertical columns as a function of response proportion. Only correct responses are plotted. Data is divided into 3 different columns according to the different stimulus types, neutral, incongruent and congruent. A proportion scale is centered on the data shows the 1 percent deviation in response proportion.

Table 1

Mean RT, (ms.) SE in RT across participants, (ms.) and accuracy for Experiment 1, 2, 3, and 4

Experiment	***Task***	***Stroop Condition***	**Mean RT (ms.)**	**SE in RT (ms.)**	**ACC**
Experiment 1	*Color Identification*	Congruent	460	51	0.94
		Incongruent	494	75	0.87
		Neutral	471	55	0.93
	Word Identification	Congruent	454	54	0.92
		Incongruent	452	54	0.92
		Neutral	449	58	0.94
Experiment 2	*Color Identification*	Identical	478	47	0.96
		Congruent	491	57	0.95
		Incongruent	519	52	0.88
		Neutral	490	53	0.93
	Word Identification	Identical	486	42	0.95
		Congruent	494	42	0.95
		Incongruent	519	82	0.90
		Neutral	495	44	0.95
Experiment 3	*Vocal Response*	Congruent	533	82	0.99
		Incongruent	669	112	0.95
		Neutral	555	91	0.99
	Manual Response	Congruent	594	57	0.95
		Incongruent	657	75	0.93
		Neutral	611	63	0.94
Experiment 4	*Color Identification*	Congruent	520	65	0.99
		Incongruent	537	76	0.98
		Neutral	522	63	0.99

Table 2

		T_{er}	a_s	σ	s_b	s_t	s_v	$b_{1\text{-}2}$	v_{idC}	v_{cC}	v_{inC}	v_{neC}	v_{idW}	v_{cW}	v_{inW}	v_{neW}	χ^2
Experiment 1	*Mean*	*284*	*.040*	*.1*	*.58*	*132*	*.064*	*1.59*		*.73*	*.66*	*.70*		*.72*	*.73*	*.74*	*119*
	SD	*19*	*.010*		*.11*	*33*	*.021*	*.24*		*.04*	*.04*	*.04*		*.04*	*.05*	*.05*	*26*
Experiment 2	*Mean*	*297*	*.049*	*.1*	*.60*	*137*	*.059*	*1.80*	*.71*	*.70*	*.64*	*.69*	*.70*	*.69*	*.65*	*.69*	*132*
	SD	*32*	*.004*		*.08*	*40*	*.012*	*.23*	*.05*	*.05*	*.04*	*.05*	*.03*	*.04*	*.03*	*.04*	*26*

Experiment 1 & 2: RTCON2 mean parameter values and SDs in the parameter values across participants. T_{er} is non-decision time in ms. a_s is the scaling parameter, multiplying drift rate, σ is SD in within-trial variability in drift rate, s_b is between-trial variability in decision criterion, s_t is between-trial variability in non-decision time, s_v is between-trial variability in drift rate, $b_{1\text{-}2}$ are the decision criterion, v is drift rate for the various conditions (C is for color identification blocks, W is for word identification blocks, id is identical, c is congruent, in is incongruent, and ne is neutral)

Table 3

		a	T_{er}	η	s_z	p_o	s_t	z	v_{idC}	v_{cC}	v_{inC}	v_{neC}	v_{idW}	v_{cW}	v_{inW}	v_{neW}	χ^2
Experiment 1	*Mean*	*.107*	*309*	*.144*	*.065*	*.02*	*136*	*.053*		*.37*	*.27*	*.33*		*.37*	*.38*	*.41*	*96*
	SD	*.020*	*18*	*.080*	*.025*	*.03*	*.03*	*.010*		*.09*	*.11*	*.11*		*.10*	*.11*	*.12*	*23*
Experiment 2	*Mean*	*.120*	*320*	*.126*	*.073*	*.01*	*130*	*.060*	*.39*	*.36*	*.24*	*.34*	*.36*	*.35*	*.27*	*.35*	*116*
	SD	*.022*	*32*	*.067*	*.024*	*.02*	*.05*	*.011*	*.10*	*.09*	*.08*	*.09*	*.06*	*.07*	*.07*	*.07*	*27*

Experiment 1 & 2: Diffusion model mean parameter values and SDs in the parameter values across participants. a is boundary separation, T_{er} is non-decision time in ms., η is between-trial variability in drift, s_z is the range in starting point, p_o is contaminant responses, s_t is between-trial variability in non-decision time, z is starting point, v is drift rate for the various conditions (C is for color identification blocks, W is for word identification blocks, id is identical, c is congruent, in is incongruent, and ne is neutral)

Table 4

		Task	T_{er}	T_{er2}	a_s	σ	s_b	s_t	s_v	b_{1-4}	v_c	v_{in}	v_{ne}	χ^2
Experiment 3	RTCON2	Vocal Mean	264		.040	.099	.47	122	.054	2.38	.95	.70	.89	73
		Manual Mean	291		.040	.119	.42	121	.049	2.48	.83	.73	.80	84
		Vocal SD	47		.001	.001	.08	30	.013	.28	.14	.09	.14	42
		Manual SD	38		.001	.001	.09	34	.014	.27	.07	.06	.06	41
	2-Ter RTCON2	Vocal Mean	260	312	.040	.098	.45	106	.061	2.26	.90	.73	.85	51
		Manual Mean	288	303	.040	.099	.46	116	.050	2.09	.73	.67	.71	75
		Vocal SD	42	48	.001	.003	.09	28	.030	.26	.11	.11	.12	33
		Manual SD	36	44	.001	.001	.08	27	.014	.22	.06	.05	.05	37
Experiment 4	RTCON2	Color identification Mean	322		.040	.120	.47	104	.051	2.01	1.04	.98	1.03	31
		Color identification SD	52		.001	.001	.15	47	.017	.27	.11	.12	.10	19

Mean parameter values and standard deviations for Experiment 3 and 4. T_{er} is non-decision time in ms. $T_{er\,2}$ is non-decision time for the incongruent condition in ms., a_s is the scaling parameter, multiplying drift rate, σ is SD in within-trial variability in drift rate, s_b is between-trial variability in decision criterion, s_t is between-trial variability in non-decision time, s_v is between-trial variability in drift rate, b_{1-4} are the decision criterion, which are set equal across response options so only one is shown, v is drift rate for the various conditions (c is congruent, in is incongruent, and ne is neutral).

Table 5

	T_{er}	T_{er2}	σ	s_b	s_t	s_v	b_{1-4}	v_c	v_{in}	v_{ne}
RTCON2	.08		.26	.02	.35	.15	.60***	.06	-.02	.16
2-Ter RTCON2	.17	.20	-.10	-.23	.37	.58**	.57**	.05	-.04	.23

Correlations of model parameters between manual and vocal response tasks. *T_{er} is non-decision time in ms. $T_{er\,2}$ is non-decision time for the incongruent condition in ms., σ is SD in within-trial variability in drift rate, s_b is between-trial variability in decision criterion, s_t is between- trial variability in non-decision time, s_v is between-trial variability in drift rate, b_{1-4} are the decision criterion, which are set equal across response options so only one is shown, v is drift rate for the various conditions (c is congruent, in is incongruent, and ne is neutral) . ** p < .01, *** p < .001*