The Stroop Effect From a Mixture of Reading Processes: A Fixed-Point Analysis

Gabriel Tillman (gabriel.tillman@newcastle.edu.au)

Zachary Howard (zachary.howard@newcastle.edu.au)

Paul Garret (paul.garret@newcastle.edu.au)

Ami Eidels (ami.eidels@newcastle.edu.au)

School of Psychology, University of Newcastle
NSW, 2308, Australia

Abstract

For the last 80 years, the Stroop task has been used to test theories of attention and cognitive control and it has been applied in many clinical settings. Most theories posit that the overwhelming power of written words overcomes strict instructions to focus on print color and ignore the word. Recent evidence suggests that trials in the Stroop task could in fact be a mixture of reading trials and non-reading trials. Here we conduct a critical test of this mixture hypothesis, where a mixture of processes should satisfy the fixed-point property (Falmagne, 1968).

Keywords: Stroop Effect; Mixture Model; Fixed-Point Analysis

The Stroop effect is one of the most replicated experimental effects in cognitive psychology (see MacLeod, 1991, for a review). The effect has been used to investigate cognitive control, and has also been applied in many clinical settings (Strauss, Sherman, & Spreen, 2006, p. 477). The task involves naming the print color of a word, where the word itself is typically the name of a color (e.g., the word GREEN printed in red print requires a response of 'red'; Stroop, 1935). People are faster at naming the print color when it matches the word (congruent stimuli, e.g., RED in red) compared to when the word and print color do not match (incongruent stimuli, GREEN in red).

A common measure of the Stroop effect is the difference in mean response time (RT) between congruent and incongruent trials. In the Stroop task, participants are instructed to name the color and ignore the word, yet it seems people cannot help but read the word (e.g., Cohen, Dunbar, & McClelland, 1990; Melara & Algom, 2003), which gives rise to faster responses on congruent trials than incongruent trials on average. Although reading must happen on some trials for an effect to be observed, it is not clear whether reading occurs on every trial, or to the same extent across trials.

Eidels, Ryan, Williams, and Algom (2014) compared the Stroop effect obtained from a standard Stroop task to the effect obtained from a novel forced-reading task. In the standard task, participants were asked to classify the print color of color-words irrespective of the content of the word. In the forced-reading task participants were asked to classify the print color of color-words (e.g., RED, GREEN), but withhold their response when presented with non-color-words (BED, GREED). To conform with the instructions, participants were forced to read every word presented. Consequently, the forced-reading Stroop task yielded a Stroop effect derived from fully processed words on every trial. The researchers found that the magnitude of the Stroop effect in the forced-reading task was larger than in the standard task, suggesting that the standard Stroop effect results from reading on only a portion of trials (see also Tillman, Eidels, & Finkbeiner, 2016).

One possible account for these results is that on any particular trial of the standard task a participant might only be processing the word to a limited extent, or not processing the word at all. A simple, formal way of explaining how different processes are mixed to yield some observed distribution of RTs is a probability-mixture model (Eidels et al., 2014; Tillman et al., 2016). Under this model, the empirical RT distributions observed in either the congruent or incongruent conditions of the standard task are a binary mixture of two unobserved distributions: one distribution of reading trials and one distribution of non-reading trials. A given trial is a sample drawn from the reading distribution (with probability p) or the non-reading distribution (with probability 1-p). The forced-reading task increases the probability of reading to (p=1).

This mixture-of-reading-processes hypothesis can be tested in a number of ways. One method assumes that a mixture of two different RT distributions should result in a bimodal observed distribution, and applies Hartigan's dip test to assess the bimodality. The test assumes the null hypothesis of unimodality over the alternative hypothesis of multimodality. If the dip statistic is greater than the 95th percentile of the reference distribution, then the null hypothesis is rejected and the observed distribution is considered bimodal (Hartigan & Hartigan, 1985). Another method is to fit both a one-component

and a two-component Gaussian mixture model to the observed data and compare both models using model selection techniques, such as AIC (Akaike, 1974).

In simulation studies, researchers have found that the Hartigan's dip test correctly identifies bi-model distributions only 65% of the time and the AIC model selection method falsely identifies bimodality 80% of the time (Freeman & Dale, 2013). In general, bimodality is difficult to detect in empirical data and requires the underlying distributions to be well separated and variability to be low (Williams, Eidels, & Townsend, 2014). However, recent software and computational advances may facilitate a more robust approach to this problem. In this paper, we test the hypothesis that the standard Stroop effect results from a mixture of reading processes by using a mathematical property of probability-mixture distributions, the fixed-point property (Falmagne, 1968).

The Fixed-Point Property

A set of mixture distributions, which are all based on the combination of two base distributions, will all intersect at a common coordinate – the fixed-point property (Falmagne, 1968, see Figure 1). Although this mathematical property could be a powerful means of identifying mixture models, researchers in the past have not commonly employed the fixed-point property test for two reasons (van Maanen, de Jong, & van Rijn, 2014). Firstly, estimating the probability density function (PDF) of the observed RT distribution from noisy data is not trivial. Secondly, it has been difficult to provide statistical evidence for the presence of the fixed-point property, which requires providing evidence for the null hypothesis.

We address the first issue by using the Epanechnikov kernel density function (Epanechnikov, 1969), which has been shown to approximate the PDF of RT distributions well (Silverman, 1986; Turner & Sederberg, 2014). To select a bandwidth for the kernel we use Silverman's "rule of thumb" (Silverman, 1986, p. 48, eq (3.31)). The default software libraries in R (R Development Core Team, 2016) allow for easy use of both the Epanechnikov kernel and Silverman's "rule of thumb". We address the second issue by using Bayesian methods to assess the degree to which there is no difference between a particular crossing point in all mixture distributions. Bayesian hypothesis testing, or Bayes factors, quantify evidence in favor of either the null hypothesis or the alternative hypothesis as a ratio. For example, when $BF_{10} = 5$ the observed data are 5 times more likely under the alternative hypothesis than under the null hypothesis. When $BF_{10} = .2$ the observed data are 5 times more likely under the null hypothesis than under the alternative hypothesis.

There is some precedent for using a fixed-point analysis to test mixture models in RT data (Brown, Lehmann, & Poboka, 2006; van Maanen et al., 2014; van Maanen, 2016). For RT distributions, when the observed RTs are made up of a mixture of unobserved distributions, there will be one RT for which the probability of providing a response at that particular time is equal for all mixtures (see Figure 1 again).

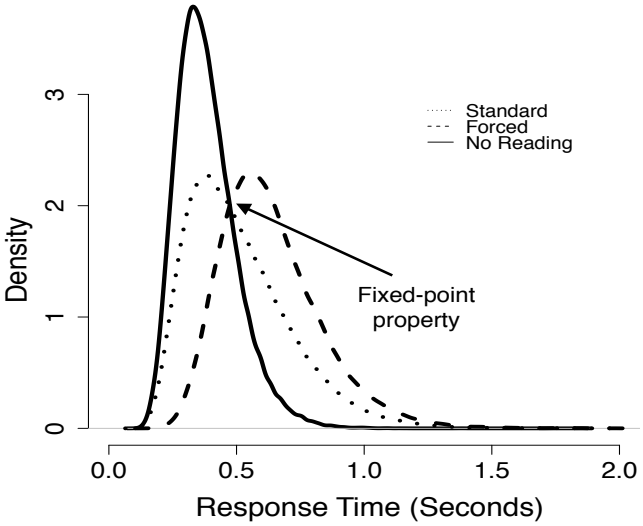

Figure 1: Illustration of the fixed-point property in Stroop distributions. The 'No Reading' distribution consists of 0% reading trials and the Forced distribution consists of 100% reading trials. The Standard distribution is a mixture of trials from both distributions. All three distributions will intersect at a common point, which is labeled "Fixed-point property" in the figure.

Here we test whether RT distributions in the Stroop task satisfy this fixed-point property. In the Stroop task participants are requested to classify the print color of color words and ignore the words' meaning. The ubiquitous Stroop effect implies they fail to exclusively focus on color and succumb to the overwhelming (perhaps automatic) attraction of reading. Previous evidence suggests that participants may not always process word meaning to the same extent (Eidels et al., 2014; Tillman et al., 2016). They may process words on some trials and not on others, in a way commensurate with a binary mixture model.

To test the mixture model we presented participants with three experimental conditions, each intended to induce a different level of reading (gauged by the probability p): a color naming task involving rectangles (probability of reading, $p = 0$), forced reading Stroop task in which each word must be read to its full extent, on each and every trial ($p = 1$), and a standard Stroop task, where participants may involuntarily read the words on some proportion of the trials ($0 < p < 1$). A probability-mixture account of reading in the Stroop task predicts that RT distributions of the three conditions will cross each other at a common point (the 'fixed-point property').

Method

Participants

Twenty two students (19 females and 3 males) from the University of Newcastle (mean age = 22.41 and SD age = 4.74) participated in the study. Participants had a proficiency in English and normal or corrected to normal vision with intact color vision. Each participant completed the standard, forced, and color naming Stroop tasks and participants were reimbursed $15 per session.

Apparatus

Each task was carried out on Dell computers running Windows XP with 17" Diamond View color monitors. Contrast and brightness were set to 80 and 50, respectively. We used the Tektronix J17 Lumacolor digital photometer and J1800 series sensor heads to calibrate color clarity across all testing stations. The software 'Presentation' was used to run the experiment and record data. Participants responded using a Cedrus response pad, model RB-830. The response keys on the response pad were marked with color stickers corresponding to the red, green or blue response.

Stimuli

For the color naming task, the stimuli were color filled rectangles in the center of the screen. For the standard and forced task, the stimuli were were 12 words that were printed in either the color red, green, or blue. The 12 words were RED, GREEN, and BLUE and three variants for each of these words. The variants differed from the color words by one letter and if substituting one letter resulted in a non-word, two letters were changed instead. The variants were GREED, GRAIN, QUEEN, RENT, ROD, BED, BASE, BLUR, and GLUE.

The variants made up the neutral stimuli for the standard and forced-reading task. The neutral stimuli were matched to the color stimuli on length, neighborhood frequency, and phonetics using the software N-Watch (Davis, 2005) and based on the CELEX word frequency database. All words were written in uppercase bold Arial font, with no words exceeding 2.55cm, or 4 visual degrees when the participant was seated 60cm from the screen.

Red, green and blue print colors of the words and rectangles had RGB values of R= 220, G=0, B=0 for red, R=0, G=0, B=240 for blue, and R=0, G=170, B=0 for green. The stimuli made up three conditions in the standard and forced-reading task. The congruent condition consisted of stimuli that had the print color and word match (RED in red, GREEN in green). The incongruent condition consisted of stimuli that had the print color and word mismatch (RED in green, GREEN in red). All non-color words were classified as neutral trials.

Procedure

Each participant completed three sessions on separate days. Each session involved the standard, forced-reading, or color naming task. The former two took about an hour to complete and consisted of 10 experimental blocks with 1 minute breaks between each. The color task took 20mins and consisted of one experimental block. The order of task presentation and position of the response buttons was counterbalanced across participants.

Each task was completed in a dark room with a desk lamp as the light source. At the beginning of each session, participants were shown 9 example trials that demonstrated the correct response. They also completed two practice blocks that consisted of 24 trials, with feedback for correct and incorrect responses in the first block.

In the color naming task, participants were instructed to respond to the print color of the rectangles by pressing the corresponding button on the response pad. In the standard task, participants were instructed to ignore the word and respond to the print color of the word. In the forced-reading task, participants were instructed to respond to the print color of words, but withhold responses to neutral words (e.g., BED, GREED, RENT).

On each trial, a fixation cross appeared in the center of the screen for 500ms, followed by a blank screen for 500ms. Following this, either a rectangle printed in color (color naming task) or a word printed in color was presented for 500ms in a random position within 40 pixels distance from the center. The spatial uncertainty prevented participants from using spatial cues to respond. Participants were required to respond within 2500ms after stimulus presentation before the trial timed out.

The color naming task involved 50 trials of blue, red, and green rectangle presentations, making for 150 trials in total per participant. For the standard and forced task, each of the ten experimental blocks consisted of 18 congruent trials, 36 incongruent trials, and 54 neutral trials. In the forced task, this allowed for half the trials to contain no response, which controls for participants predicting a non-response trial. Each combination of congruent and incongruent stimuli were presented 6 times per block. The order of stimulus presentation was randomized within each block. The RT was recorded in milliseconds.

Results

The probability mixture account makes two testable predictions. First the fixed cross point, where all three Stroop distributions will have a single RT with equal probability of providing a response at that time – we test this in the following section. The mixture account also predicts that the observed (mixture) distribution will be bound between the faster non-reading distribution and the slower forced-reading distribution. This is exactly

what we observed in our data. The color naming task had a mean RT of 436ms (SD = 132ms). In the congruent condition, the standard and forced-reading tasks had mean RTs of 470ms (SD = 150ms) and 700ms (SD = 208ms), respectively. While in the incongruent condition, these tasks had mean RTs of 495ms (SD = 172ms) and 832ms (SD = 233ms), respectively.

We also used Bayesian paired samples t-tests to evaluate the evidence for differences between the color naming, standard, and forced-reading mean RTs in the congruent and incongruent conditions. In the congruent condition, participants were slower in the standard task than the color naming task ($BF_{10} = 8.8 \times 10^{660}$) and were slower in the forced-reading compared to the standard task ($BF_{10} = 5 \times 10^{559}$). In the incongruent condition, participants were slower in the standard task than the color naming task ($BF_{10} = 1.2 \times 10^{1295}$) and were slower in the forced-reading task compared to the standard ($BF_{10} = 6.9 \times 10^{2770}$).

Fixed-Point Analysis

The analysis was carried out using the 'fp' package (van Maanen et al., 2014) in R (R Development Core Team, 2016) – but we used the Epanechnikov kernel instead of the default Gaussian kernel as recommended by Silverman (1986, p. 43).

The analysis involved calculating the probability density of each RT distribution in each task. For example, focusing only on the congruent condition (and later similarly focusing on the incongruent condition) we estimated the RT distribution for the color naming task, the standard task, and the forced reading task, which by design have a mixture proportion of $p = 0$, $0 < p < 1$, and $p = 1$, respectively. We then found the crossing point of each pair of distributions (i.e., forced-standard, forced-color naming, standard-color naming). The fixed-point property holds if all pairs cross at the same point along the x and y axis (see Figure 1).

We tested whether the fixed-point property holds for the sample of participants in our study. We calculated the crossing points per pair of mixture proportion tasks for each of the participants for both the congruent and incongruent distributions, but the color naming distribution was the same for both the congruent and incongruent comparison. We then subjected these crossing points to Bayesian analysis of variance (ANOVA). The Bayesian analysis was conducted using the Bayes Factor package (Morey, Rouder, & Jamil, 2014; Rouder, Morey, Speckman, & Province, 2012) in R. The Bayes factor from the ANOVA provides evidence for or against the fixed-point property.

We calculated the Bayes factor as the ratio of the evidence for the null hypothesis over the alternative. The null hypothesis posits that there is no difference in crossing points between all distributions in question, and thus suggests that the fixed-point property is satisfied. In line with Kass and Raftery (1995) we consider a Bayes factor greater than 3 as positive evidence in favor of the null (fixed cross point) and against the alternative hypothesis that there is a difference between crossing points.

The RT distributions for the congruent and incongruent trials are presented in Figure 2. For the congruent condition, the Bayes factor ANOVA revealed that the null model was preferred to the alternative model by a Bayes factor of 1.25. The data provide equivocal evidence in favor of both the null and alternative hypothesis for the congruent Stroop distributions. For the incongruent condition, the Bayes factor ANOVA revealed that the null model was preferred to the alternative model by a Bayes factor of 2.27. The data provides evidence in favor of the hypothesis that there is no differences between crossing points, but the evidence is inconclusive.

General Discussion

In the Stroop task, slower responses on incongruent trials relative to congruent or even neutral trials implies participants read the words despite instructions to focus on color and ignore the words' meaning. Recent evidence suggests participants may read on some proportion of the trials and not on others (Eidels et al., 2014; Tillman et al., 2016). When the observed RT on a single trial is sampled from a non-reading distribution, color naming will *not* be slowed down by the incongruent word. When the observed RT on a single trial is sampled from a reading distribution, the speed of color naming will be slowed down by an incongruent word, therefore, contributing to a Stroop effect. The magnitude of an observed Stroop effect reflects the proportion of trials in which the participant has read on - the greater the proportion, the larger the effect. To statistically test for this mixture of reading processes in the Stroop task, we ran a fixed-point property analysis on Stroop RT distributions with different reading proportions. We found some evidence for a mixture of distributions in the incongruent condition, but the results of the analysis were not conclusive.

The fixed-point property analysis is one method for testing for a mixture of processes, but it requires the strong assumption that there is a pure mixture of reading and non-reading processes. That is, the approach assumes that the *only* difference between the three tasks is the proportion of reading trials. This assumption may be compromised by other contaminant processes across the tasks. For example, the Stroop effect can be effected by attentional resources (Kahneman & Chajczyk, 1983), practice (MacLeod & Dunbar, 1988), dimensional discriminability and experimental correlation (Dishon-Berkovits & Algom, 2000), target set size (La Heij & Vermeij, 1987), and the number of colored letters in the stimulus word (Besner, Stolz, & Boutilier, 1997). Further, there are differences in stimuli (words vs rectangles) across tasks. Whilst our results are inconclusive with re-

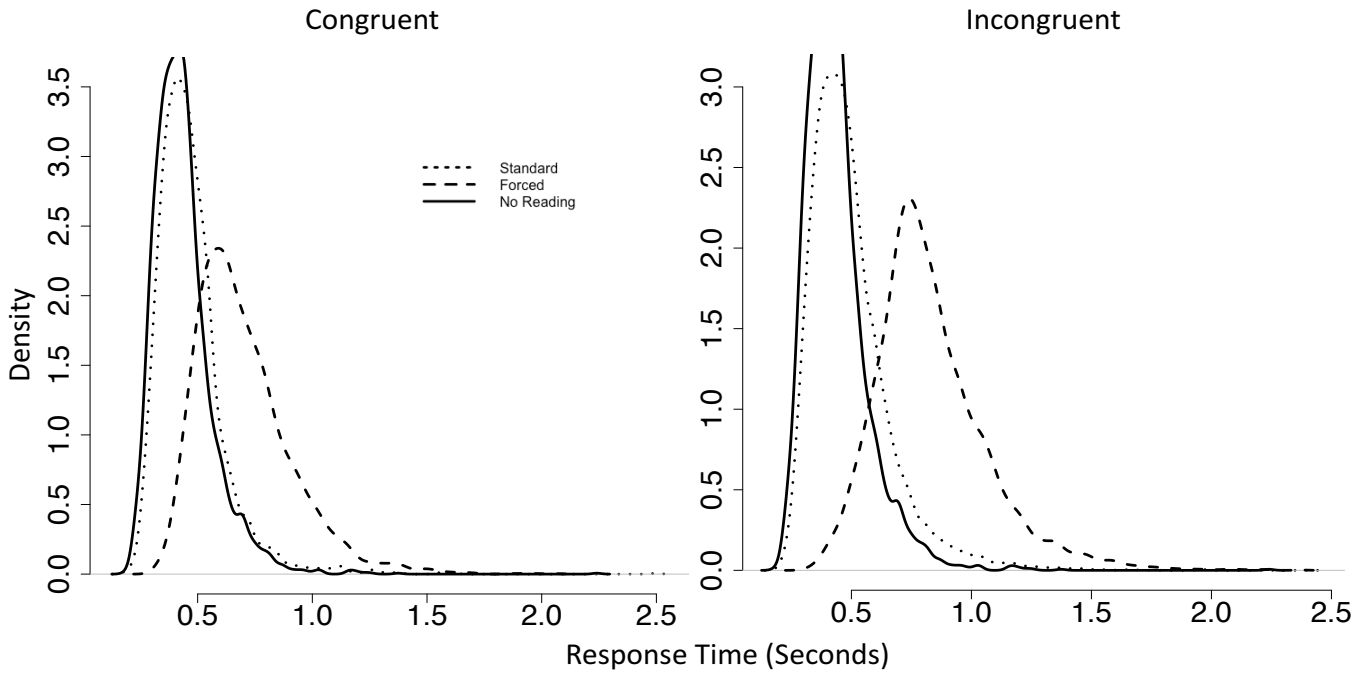

Figure 2: Overall RT density for congruent and incongruent Stroop distributions.

gards to identifying a mixture process, they certainly do not preclude the mixture hypothesis as being a viable explanation for the Stroop effect.

Our study also reflects the difficulties in distinguishing between single-process and dual-process mental phenomena, which is an issue that besets cognitive psychology (e.g., Yap, Balota, Cortese, & Watson, 2006; Wixted, 2007; Freeman & Dale, 2013). Nonetheless, the mixture model of Stroop has clinical, empirical, and theoretical implications. If the Stroop effect distribution is derived from a reading distribution and a non-reading distribution, and the combination of these distributions makes up the observed distribution, then clinical applications of the Stroop task need to consider this mixture of reading processes. For instance, differences in Stroop effect magnitude may not only reflect differences in attentional control, but could simply reflect a difference in the proportion of reading across trials. Empirically, future work could account for the proportion of reading trials by employing the benchmark forced-reading task along with the standard task. Finally, theories of Stroop (e.g., Cohen et al., 1990; Melara & Algom, 2003) will need to consider what mechanism allows for a Stroop effect to only arise on some proportion of trials but not others. Given these implications, we hope to see more robust testing of the mixture-of-reading-processes hypothesis outlined here.

Acknowledgments

We would like to thank Katie Berka for data collection and Scott Brown for helpful comments related to the fixed-point analysis.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

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

Brown, S. D., Lehmann, C., & Poboka, D. (2006). A critical test of the failure-to-engage theory of task switching. *Psychonomic bulletin & review*, *13*(1), 152–159.

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), 332-361.

Davis, C. J. (2005). N-watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior research methods*, *37*(1), 65–70.

Dishon-Berkovits, M., & Algom, D. (2000). The stroop effect: It is not the robust phenomenon that you have thought it to be. *Memory & Cognition*, *28*(8), 1437–1449.

Eidels, A., Ryan, K., Williams, P., & Algom, D. (2014). Depth of processing in the stroop task: Evidence from

a novel forced-reading condition. *Experimental Psychology*, *61*(5), 385-393.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), 153–158.

Falmagne, J. C. (1968). Note on a simple fixed-point property of binary mixtures. *British Journal of Mathematical and Statistical Psychology*.

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior research methods*, *45*(1), 83–97.

Hartigan, J. A., & Hartigan, P. (1985). The dip test of unimodality. *The Annals of Statistics*, 70–84.

Kahneman, D., & Chajczyk, D. (1983). Tests of the automaticity of reading: dilution of stroop effects by color-irrelevant stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(4), 497-509.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

La Heij, W., & Vermeij, M. (1987). Reading versus naming: The effect of target set size on contextual interference and facilitation. *Perception & psychophysics*, *41*(4), 355–366.

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

MacLeod, C., & Dunbar, K. (1988). Training and stroop-like interference: evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 126-135.

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

Morey, R., Rouder, J., & Jamil, T. (2014). Bayesfactor: Computation of bayes factors for common designs. *R package version 0.9*, *8*.

R Development Core Team. (2016). The r project for statistical computing [Computer software manual]. Vienna, Austria.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356-374.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford University Press, USA.

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

Tillman, G., Eidels, A., & Finkbeiner, M. (2016). A reach-to-touch investigation on the nature of reading in the stroop task. *Attention, Perception, & Psychophysics*, *78*(8), 1 - 11.

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic bulletin & review*, *21*(2), 227–250.

van Maanen, L. (2016). Is there evidence for a mixture of processes in speed-accuracy trade-off behavior? *Topics in cognitive science*, *8*(1), 279–290.

van Maanen, L., de Jong, R., & van Rijn, H. (2014). How to assess the existence of competing strategies in cognitive tasks: a primer on the fixed-point property. *PloS one*, *9*(8), e106113.

Williams, P., Eidels, A., & Townsend, J. T. (2014). The resurrection of tweedledum and tweedledee: Bimodality cannot distinguish serial and parallel processes. *Psychonomic bulletin & review*, *21*(5), 1165–1173.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological review*, *114*(1), 152.

Yap, M. J., Balota, D. A., Cortese, M. J., & Watson, J. M. (2006). Single- versus dual-process models of lexical decision performance: Insights from response time distributional analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1324–1344.