Advancing Methods and Mathematical Models of Perceptual Decision Making

by

Gabriel Tillman

B.Psych (Hons I)

A thesis submitted in fulfillment of the requirements

for the degree of Doctor of Philosophy (Psychology - Science)

October 30, 2016

Declaration of Authorship

- The thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository**, subject to the provisions of the Copyright Act 1968. **Unless an Embargo has been approved for a determined period.

- I hereby certify that the work embodied in this thesis has been done in collaboration with other researchers, or carried out in other institutions. I have included as part of the thesis a statement clearly outlining the extent of collaboration, with whom and under what auspices.

- I hereby certify that the work embodied in this thesis contains published papers scholarly work of which I am a joint author. I have included as part of the thesis a written statement, endorsed by my supervisor, attesting to my contribution to the joint publications scholarly work.

Signed:

_____

Acknowledgements

Where would I be without Don, Andrew, and Scott who made me the researcher I am today. I thank Ami who was always there to provide solid advice. And Adam for his enthusiasm and interesting pitches for future work. It was fun working with students in the lab who were always there to answers my questions, ask their own, and devise novel ways to procrastinate from work. All my collaborators taught me how to think critically and question everything. I am truly thankful for the support from all these people. To my wife Lucy, who has been caring, supportive, patient, thoughtful and loving, I give my deepest thanks.

List of Publications

This thesis is based on the following published/submitted work. For each paper I provide the full bibliographic citations in the order they appear in the thesis:

1. Tillman, G., Strayer, D., Eidels, A., & Heathcote, A. (Under Review). Modeling Cognitive Load Effects of Conversation Between a Passenger and Driver. Attention, Perception, & Psychophysics.

2. Tillman, G., Benders, T., Brown, S. D., & van Ravenzwaaij, D. (Under Review). An Evidence Accumulation Model of Acoustic Cue Weighting in Vowel Perception. Journal of Phonetics

3. Tillman, G., Osth, A., van Ravenzwaaij, D.,& Heathcote, A. (Under Review). A Diffusion Decision Model Analysis of Evidence Variability in the Lexical Decision Task. Psychonomic bulletin & Review

4. Tillman, G., Eidels, A., & Finkbeiner, M. (2016). A Reach-To-Touch Investigation on the Nature of Reading in the Stroop Task. Attention, Perception, & Psychophysics.

Statement of Contribution

I attest that Research Higher Degree candidate Gabriel Tillman led the manuscripts included in this thesis. Gabriel Tillman made major contributions to each manuscript including, coordinating and supervising data collection, completing all data analyses and model fitting, and served as lead author for manuscript preparation.

Name: Don van Ravenzwaaij

Signed:

Name: Andrew Heathcote

Signed:

Name: Scott Brown

Signed:

Name: Ami Eidels

Signed:

Name: Adam Osth

Signed:

Name: Titia Benders

Signed:

Name: Matthew Finkbeiner

Signed:

Name: David Strayer

Signed:

Additional Work

Listed are additional publications and presentations that have relevance to the thesis, but are not included in it:

Invited Presentations

1. Tillman, G (2016, August). Sequential Sampling Models of Perceptual Decision Making. Invited Talk presented for the Department of Linguistics, Macquarie University, Australia.

2. Tillman, G (2016, July). Advancing Cognitive Models of Perceptual Decision Making. Invited Talk presented for the Department of Psychology, Vanderbilt University, United States.

3. Tillman, G (2014, December). How Do Our Past Decisions Affect Our Present Decisions? Invited Talk presented for the University of Newcastle Cognitive Research Group, University of Newcastle, Australia.

Conference Presentations

1. Tillman, G. & Osth, A. (2016, February). Diffusion modeling reveals evidence for unequal variance signal detection models of the lexical decision task. Talk presented at the annual meeting of the Australian Mathematical Psychology Conference (AMPC), University of Tasmania, Australia.

2. Freeman, E., Tillman, G. & Osth, A. (2015, November). Recognition memory for familiar and unfamiliar items: Links between encoding and retrieval differences. Poster presented at the annual meeting of the Psychonomic Society, Chicago, United States.

3. Tillman, G. & Osth, A. (2015, November). Unequal Variance of Drift Rate Distributions in the Lexical Decision Task. Poster presented at the Computational Approaches to Cognition Symposium, Chicago, United States.

4. Tillman, G., Benders, T., Brown, S. D., & van Ravenzwaaij, D. (2015, November). Determining the Role of Spectral and Duration Cues in Vowel Perception. Talk presented at the annual meeting of the Configural Processing Consortium, Chicago, United States.

5. Tillman, G., Benders, T., Brown, S. D., & van Ravenzwaaij, D. (2015, April). Determining the Role of Spectral and Duration Cues in Vowel Perception. Talk presented at the annual meeting of the Australasian Society for Experimental Psychology, University of Sydney, Australia.

6. Tillman, G. & van Ravenzwaaij, D. (2015, February). Are Conclusions from Sequential Sampling Models Reliable? Talk presented at the annual meeting of the Australian Mathematical Psychology Conference (AMPC), Newcastle, Australia.

7. Tillman, G., Benders, T., Brown, S. D., & van Ravenzwaaij, D. (2014, November). A response time model on the role of spectral and duration cues in vowel perception. Poster presented at the annual meeting of the Psychonomic Society, Long Beach California, United States.

8. Tillman, G., Eidels, A., & Finkbeiner, M. (2014, July). Is Reading Mandatory? Reaching for Evidence in the Stroop Paradigm. Poster presented at the annual conference of the Cognitive Science Society, Quebec City, Canada.

9. Tillman, G., Eidels, A., & Finkbeiner, M. (2014, April). A Reach- To-Touch Investigation on the Nature of Reading in the Stroop Task. Talk presented at the annual meeting of the Australasian Society for Experimental Psychology, Brisbane, Australia.

10. Eidels, A., Williams, P., & Tillman, G. (2014, February). Serial-parallel model mimicry: the case for bimodality. Talk presented at the annual meeting of the Australian Mathematical Psychology Conference (AMPC), Canberra, Australia.

11. Tillman, G., Eidels, A., & Finkbeiner, M. (2013, December). Is Reading Mandatory? Reaching for Evidence in the Stroop Paradigm. Poster presented at the Priority Research Centre for Translational Neuroscience and Mental Health? Sixth Annual Postgraduate and Postdoctoral Conference, Newcastle, Australia.

Contents

Abstract

In this thesis I argue that cognitive psychologists can use the combination of sequential sampling models, Bayesian estimation methods, and model comparison via predictive accuracy to investigate underlying cognitive processes of perceptual decision-making. I show that sequential sampling models of simple and choice response time allow for researchers to analyze behavioral data and translate them into the constitute components of processing, such as speed of processing, response caution, and the time needed for perceptual encoding and overt motor responses. I use these methods and models to investigate underlying mental processes related to cognitive load, speech perception, and lexical decision-making. I also show that using different sequential sampling models to analyze the same data can lead researchers to draw different conclusions about cognitive processes, which serves as a caution for carelessly using these models. I also present a novel method that researchers can use to observe cognitive processes unfold online during perceptual decision-making tasks. I then discuss a promising collaboration emerging between researchers in the field of mathematical modeling and neuroscience.

For Lucy and Willow

Chapter 1

Methods and Models of Perceptual Decision Making

There are three stages in scientific discovery. First, people deny that it is true, then they deny that it is important; finally they credit the wrong person.

Bill Bryson

Cognitive psychologists want to understand how the human brain produces behavior. But because brains are so complex, a reasonable starting point is to investigate one of the simpler and more frequent types of human behavior – perceptual decision-making. Perceptual decisions are choices made about incoming sensory information. Is that light red or green? Was that a knock at the door? Is that bird flying towards me?

To study perceptual decisions, researchers approximate real world stimuli with simpler stimuli, such as small moving dots on a computer screen. Suppose we want to understand how participants decide whether dots on a screen are coherently moving to the left or to the right – an experiment called the motion dots task (Ball & Sekuler, 1982). In this task, some of the dots are coherently moving to the left or right and some are moving in no coherent direction. Participants need to decide which direction the coherent dots are moving. Researchers can make the dots bigger, brighter, or make different proportions of the dots move coherently. They may also ask participants to respond as fast as they can or as accurately as they can. In either case, manipulating the stimuli or task demands can have observable effects on how fast and accurately participants make perceptual decisions.

The motion dots task, or similar perceptual experiments, afford researchers substantial control over the input to human minds, while the participant's overt decisions serve as observable output (i.e., data) – but a cognitive psychologist's real goal is to understand what happens between input and output. For example, Jane and John both take part in the motion dots experiment and John is faster at responding than Jane. But is he faster because he processes the motion dot stimuli more quickly? Or does John have a faster motor response allowing him to press

the response button faster? Perhaps John requires less evidence than Jane to make his decision. One way to learn about these unobserved cognitive processes is to define a set of mathematical equations that can produce the observed data. This set of equations is known as a cognitive model. If the predicted data of the model captures all the important structure of the observed data, then researchers can draw meaningful conclusions about unobserved cognitive processes from the parameters that control the model.

In this thesis I advocate that researchers can investigate the unobserved processes involved in perceptual decision-making by using a class of mathematical models known as sequential sampling models. Accurate parameter estimation is critical if researchers are to use parameters from sequential sampling models to learn about cognitive processes. I show that Bayesian parameter estimation is a principled method for applying sequential sampling models to empirical data. A longstanding issue in cognitive science is comparing competing models of a particular phenomenon. I show that researchers can choose between competing models by assessing how well each model predicts future data. As an ancillary goal of this thesis I also develop a novel method for learning about unobserved cognitive processes in perceptual decision-making tasks. For the rest of this chapter I explain what sequential sampling models are, introduce the fundamentals of Bayesian parameter estimation, and discuss choosing between competing models by using a model's predictive accuracy. The last section of this chapter outlines the specific contributions of this thesis.

1.1 Sequential Sampling Models

Since their advent (Stone, 1960), sequential sampling models have been used by researchers to investigate the unobserved mental processes involved in a range of perceptual decision-making tasks (e.g., e.g., Ratcliff, 1978; Ratcliff & Rouder, 1998; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008; van Ravenzwaaij, Dutilh, & Wagenmakers, 2012; Eidels, Donkin, Brown, & Heathcote, 2010; Forstmann et al., 2008). These models assume a simple cognitive architecture consisting of stimulus encoding, response selection, and overt response execution. To make a perceptual decision, people begin with an initial amount of evidence for all response options, the *starting point* of evidence accumulation (Figure 1.1). From the starting point, more evidence is continually sampled from the stimulus, which accumulates at a rate of *drift rate* towards the *response threshold*. When the accumulated evidence crosses a response threshold this triggers the corresponding overt response.

The quality of evidence sampled from the stimulus governs the drift rate, which can be interpreted as the speed of information processing. Higher response thresholds mean that a person needs more evidence to trigger a response, and so, threshold settings represent how cautious a person is. Starting points and response thresholds can vary across response options capturing any inherit biases people have. The time necessary for processes outside of evidence accumulation is the non-decision time, which includes the time needed for perceptual encoding and overtly executing a motor response.

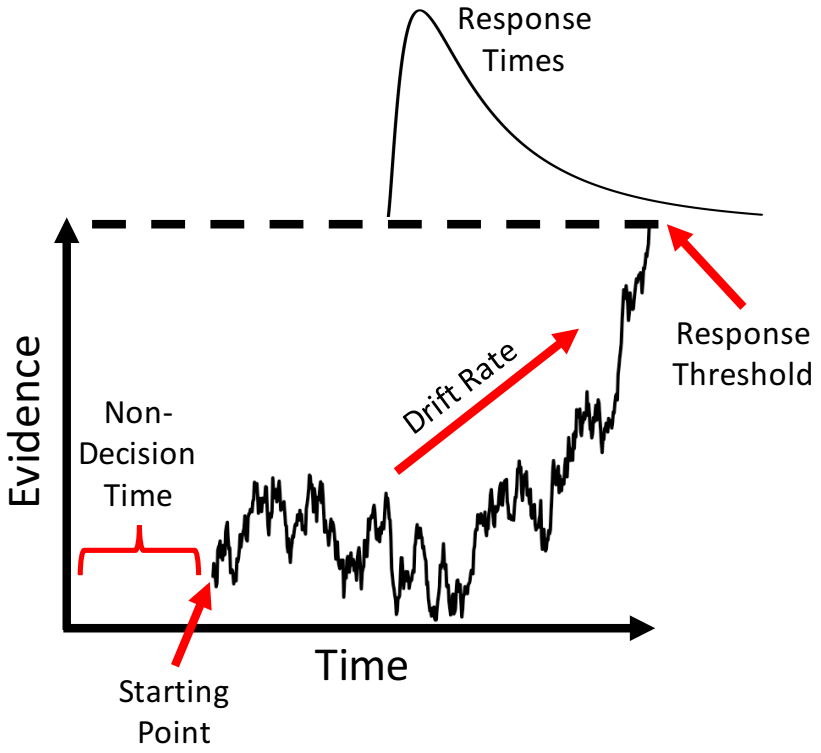


FIGURE 1.1: A typical sequential sampling model process. Only one accumulator is illustrated, but multiple alternative decisions are modeled by multiple racing accumulators (Usher & McClelland, 2001; S. D. Brown & Heathcote, 2008; S. D. Brown & Heathcote, 2005) or single accumulators with two boundaries (Ratcliff, 1978; Link & Heath, 1975).

Researchers developed sequential sampling models to account for the complex relationship between accuracy and response times (RTs), which are the ubiquitous dependent measures of perceptual decision-making tasks. Previous models only accounted for either RTs (Sternberg, 1969) or accuracy (Green & Swets, 1966), perhaps because of the difficulties of accounting for both measures simultaneously.

Firstly, there is a well-known relationship between accuracy and speed – *the speed-accuracy trade-off* – where fast decisions are more likely to be incorrect than slower decisions (e.g., Wickelgren, 1977; Luce, 1986; Heitz, 2014). Sequential sampling models offer an intuitive account of how participants trade accuracy for speed. When response thresholds are high, RTs will be longer and will more likely

be correct. When thresholds are low, the decision process terminates earlier, which speeds up RT, but increases the likelihood of responding incorrectly because the decision is made with less evidence.

Secondly, in experiments with high accuracy, researchers find that error RTs are slower than correct RTs, but when accuracy is low, error RTs are typically faster than correct RTs (Luce, 1986). Sequential sampling models account for both slow errors and fast errors by assuming trial to trial variation in drift rate (Ratcliff, 1978) and starting point (Laming, 1968), respectively. Emphasizing speed in one condition of an experiment, causing faster error RTs than correct, and emphasizing accuracy in another condition, causing slower error RTs than correct, results in a cross-over of the relative speed of error RT. Assuming variability in both drift rate and starting point (Ratcliff & Rouder, 1998) accounts for this cross-over effect.

Finally, accuracy and RTs are on radically different scales (Ratcliff & McKoon, 2008), where accuracy rates are bound between 0 and 1 and RTs are bound between 0 and infinity. As accuracy increases its variance decreases and as RT increases its variance increases. Despite these differences in scale, sequential sampling models can adequately predict the accuracy rates and RTs observed in perceptual decision-making tasks across a range of paradigms. Given the reliable predictions of sequential sampling models, researchers can inspect the model parameters, which represent the underlying constituent components of processing involved in perceptual decision-making. Processes such as information processing speed, response caution, perceptual encoding time, or motor response time.

In fact, sequential sampling models have a track record for investigating

unobserved cognitive processes. For instance, it is typically found that as people age their RTs increase in cognitive tasks. For almost 20 years the dominant theory of why performance declined with age was that aging resulted in a general slow-down (Salthouse, 1996). However, when researchers analyzed the same data with a sequential sampling model, they found that the locus of the slow-down in elderly people was higher response caution, not a lower processing speed (Ratcliff, Thapar, & McKoon, 2001, 2004; Thapar, Ratcliff, & McKoon, 2003; Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & Mckoon, 2003). Researchers later found that higher response threshold settings in the elderly population correlates with reduced white matter integrity in tracts connecting the pre-SMA to the striatum (Forstmann et al., 2011), which are brain regions associated with adjusting response caution (Forstmann et al., 2008).

1.2 Estimating Parameters: A Bayesian Approach

Drawing conclusions from sequential sampling models about cognitive processes is dependent on accurately estimating parameters. Kolmogorov-Smirnov (e.g., Voss, Rothermund, & Voss, 2004), χ^2 (e.g., Ratcliff, 2002), D*M (Verdonck & Tuerlinckx, 2015), and maximum-likelihood (e.g., Myung, 2003; Heathcote, Brown, & Mewhort, 2002) are some of the parameter estimation techniques used by researchers to fit sequential sampling models to data. But recently, there has been a surge in the usage of Bayesian parameter estimation techniques.

Bayesian parameter estimation allows researchers to justify their beliefs in certain parameter values by using probability theory. For example, if we assume

that a sequential sampling model is the generating process underlying RTs from the motion dots task, then what is the probability of a particular drift rate or response threshold, given the data we have observed. Bayesian estimation is simply the process of calculating this probability. The calculation involves combining what we believe about the parameters before having seen the data, the prior, with what the data tell us we should believe about the parameters, the likelihood, to get a more refined belief about parameters, the posterior. Bayes Rule formally describes this relationship as:

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)} \tag{1.1}$$

where θ denotes a model's parameter and D denotes the data. $P(\theta|D)$ refers to the posterior, $P(D|\theta)$ refers to the likelihood, and $P(\theta)$ refers to the prior. $P(D)$ is the probability of each possible data point across all possible parameter values – in other words, it is the evidence for the entire set of possible parameter values. Bayes rule states that if we multiply the likelihood by the prior, and then normalize the result by dividing by the evidence, we get the posterior probability – or the probability of a parameter value given the data.

In Bayesian estimation we use probability distributions to quantify the prior, the likelihood, the evidence, and the posterior. The posterior distribution is the central feature of Bayesian estimation and represents a set of possible parameter values with corresponding probabilities. The posterior distribution informs researchers about what parameters they should believe in more and what parameters they should believe in less.

In the past, Bayesian estimation was not feasible because we did not have the power of modern-day computers. Computing power is necessary because we cannot derive posteriors analytically for complex models, such as sequential sampling models. Instead, we need to numerically approximate the posterior distribution by generating tens of thousands of random samples using a class of sampling methods known as Markov chain Monte Carlo (MCMC; see van Ravenzwaaij, Cassey, & Brown, 2015, for a tutorial). The crux of MCMC Bayesian estimation is to randomly generate samples from a posterior distribution that is not normalized, which is proportional to the product between the likelihood and the prior, where

$$P(\theta|D) \propto P(D|\theta)P(\theta). \tag{1.2}$$

In summary, Bayesian estimation is a principled way for researchers quantify plausibility and uncertainty in model parameters. Bayesian methods have become available to researchers due to the production of fast personal computers, which can run MCMC numerical approximations to posterior distributions.

1.3 Model Comparison

Recall the motion dots task that John and Jane have completed in which John was faster at responding than Jane. If a sequential sampling model was the generating process then John's faster RTs may be the result of a higher drift rate, suggesting that he processes information faster. Another possibility is that John has a lower response threshold, meaning that he is less cautious than Jane. Both

models offer competing accounts of the data. Choosing between plausible models is a fundamental problem that besets the field of psychology. The problem arises because the models we work with are not the true generating processes (Box & Draper, 1987), instead our models are approximations to the truth that are useful for inferring about underlying cognition. One way to choose between models is to assess their "usefulness".

One metric of usefulness is how well a model predicts future data. However, data typically has both important structure and random non-important structure – which we call signal and noise, respectively. A model that is too simple will poorly predict future data because it does not capture all the signal, and therefore, does not give a comprehensive explanation of the behavior of interest. On the other hand, a model that is too complex will poorly predict future data because it captures both the signal and noise, but the noise present in current data will not likely be in the future data. The best candidate model will capture all the signal and none of the noise and is considered to have the best out-of-sample predictive accuracy.

The gold standard for estimating the out-of-sample predictive accuracy of a model is cross-validation (Geisser & Eddy, 1979). This method involves partitioning your data into training data and validation data. The aim is to fit your model to training data and then assess the model's capacity to predict the validation data that were held out. Cross-validation is computationally expensive and therefore researchers have sought computationally cheap approximations to cross-validation.

Information criteria are a popular set of methods that approximate the out-of-sample predictive accuracy of a model. In short, these methods involve calculating

the goodness-of-fit of a model and subtracting a value that represents the complexity of the model from the goodness-of-fit value.

For non-Bayesian models, we can use methods such as Akaike's information criterion (AIC Akaike, 1974) or the Bayesian information criterion (BIC G. Schwarz, 1978). However, until recently there has been few methods to quantify out-of-sample accuracy for Bayesian models. The deviance information criterion has been the method of choice for over 10 years (DIC Spiegelhalter, Best, Carlin, & van der Linde, 2002), with no popular alternatives. DIC is based on the assumption that the posterior is a multivariate normal distribution and violations of this assumption can cause accuracy issues. Recently Gelman, Hwang, and Vehtari (2014) have advocated using the widely applicable information criterion (WAIC Watanabe, 2010) as an improvement on the DIC method because WAIC requires no assumptions about the posterior distribution and it is calculated from each data point, which improves accuracy. Researchers have also developed efficient implementations of the WAIC method (Vehtari, Gelman, & Gabry, 2016).

1.4 The Key Contributions

The main goal of this thesis is to investigate a number of perceptual decision-making phenomena by applying sequential sampling models to behavioral data using Bayesian estimation methods. In chapter 2, I develop a simple-response time model of cognitive load effects on drivers and passengers of motor vehicles. In chapter 3, I investigate how listeners are able to perceive phonemes in speech and address several long standing questions in the speech perception field. In chapter 4, I extend

a current sequential sampling model of lexical decision to account for effects related to different stimulus types. In these studies I select between competing models by evaluating the predictive performance of each model.

Chapter 5 and 6 address two secondary goals. In chapter 5, I highlight two conceptual differences between two prominent sequential sampling models. These differences have practical implications when it comes to drawing psychological conclusions from the models. In chapter 6, I present a novel and promising method that can be used to investigate cognitive processes in perceptual decision-making. The method involves state-of-the-art motion tracking technology that maps the arm-movements of participants, where the arm-movements serve as a window into cognition.

In chapter 7, I summarize the current work and discuss future directions of sequential sampling models. Specifically, I discuss how mathematical models combined with advanced neuroimaging technology offers a powerful tool that can be used to get a complete understanding of human cognition and the underlying neurophysiology.

Chapter 2

A Simple Response Time Model of Cognitive Load Effects

Reference

Tillman, G., Strayer, D., Eidels, A., & Heathcote, A. (Under Review). Modeling Cognitive Load Effects of Conversation Between a Passenger and Driver. Attention, Perception, & Psychophysics.

2.1 Introduction

Cognitive psychologists use the term capacity to refer to the human ability to cope with the cognitive load associated with increasing amounts of perceptual information (e.g., Eidels, Donkin, et al., 2010; Townsend & Eidels, 2011). Human capacity is often limited (Kahneman, 1973), yet many situations in modern life require simultaneous processing of information from multiple signals. Given the limited capacity for processing, it is important for researchers to understand the consequences of such limitations in safety critical activities, such as driving a car.

Cognitive load from secondary tasks, such as talking on a cell phone, is one of the main sources of distraction while driving (Strayer et al., 2013, 2015). Distraction while driving is a significant cause of injuries and fatalities for drivers and passengers on the roadway (Ranney, Mazzae, Garrott, & Goodman, 2000; Wang, Knipling, & Goodman, 1996; Sussman, Bishop, Madnick, & Walter, 1985; Dingus et al., 2006). Strayer and Johnston (2001) studied the effects of cell phone conversations on performance in a simulated driving task. They found that conversations with either a hand-held or a hands-free cell phone while driving resulted in a failure to detect traffic signals, as well as slower reactions when the traffic signals were detected (cf. Strayer, Drews, & Johnston, 2003). Surprisingly, no such decrements are observed when a similar conversation is held between the driver and a passenger in the car (Drews, Pasupathi, & Strayer, 2008). In fact, data on crash risk reveals lower accident rates when an adult passenger is in the car than when the driver is alone (Rueda-Domingo et al., 2004; Vollrath, Meilinger, & Krüger, 2002).

The Detection Response Task (DRT) is an international standard for assessing cognitive load on drivers' attention (International Organization for Standardization, 2015) that can safely be deployed with no appreciable effect on driving performance (Strayer, Turrill, Coleman, Ortiz, & Cooper, 2014). The DRT measures cognitive load by asking participants in a driving simulator to respond when they detect a small light in their peripheral vision. Increases in response times (RT) in the DRT measure the effect of increased cognitive load. Although the DRT is a valid measure of the effects of cognitive load during driving (Strayer et al., 2013, 2015), there is little research on what components of DRT processing are affected by increased cognitive load.

For instance, when using a hands-free cell phone, drivers are slower to respond in the DRT compared to when they are not using the device (Strayer et al., 2013). The increased RT is believed to result from a lower rate of information processing, perhaps because the DRT and cell phone share a limited pool of processing resources (Strayer, Watson, & Drews, 2011; Strayer et al., 2013). However, other causes are also possible. People could be more cautious in the DRT with increased cognitive load by setting a higher threshold for the amount of evidence needed to decide the light is present. Or people may require more time for non-decision processes such as stimulus encoding or response production. We address the role of processing-rate, threshold, non-decision time, or some combination of these three, by fitting a cognitive model of the DRT task under conditions that vary in the load imposed by conversation. In the next section we outline the modeling framework applied to the DRT data. The data was collected from both drivers and passengers performing a simulated driving task. Cognitive load was manipulated by having the

driver converse with a passenger in person or over a hands-free cell phone. These conditions were compared to a baseline where the driver took part in the simulator and DRT without any conversation.

2.1.1 Modeling the Detection Response Task

Sequential sampling models characterize responding as the result of a noisy process of accumulating evidence towards a response threshold. They have been extensively used to understand choice RT in terms of effects on evidence accumulation rate, response threshold, and non-decision time (S. D. Brown & Heathcote, 2008; Ratcliff & McKoon, 2008). Recently, sequential sampling models – and in particular the single-bound diffusion model (W. Schwarz, 2001; Heathcote, 2004) – have been applied to simple RT data (i.e., data where participants make only one type of response) from a range of paradigms. Paradigms such as the psychomotor vigilance test and brightness detection tasks (Ratcliff & Van Dongen, 2011), simulated driving tasks (Ratcliff & Strayer, 2014; Ratcliff, 2015), go/no-go tasks (W. Schwarz, 2001; Heathcote, 2004), as well as pointing, picture naming and eye-movement tasks (Anders, Alario, & van Maanen, 2016). We collected simple RT data from the DRT ('press a key if you detect light') and fit the single-bound diffusion model in order to investigate the causes underlying slowing due to increased cognitive load.

Figure 2.1 is a schematic of the single-bound diffusion model. The response threshold, 'a', quantifies the amount of evidence needed to make a response. On each trial, noisy evidence accumulates towards the response threshold at some rate – the drift rate. Within-trial (moment-to-moment) noise causes accumulation of evidence

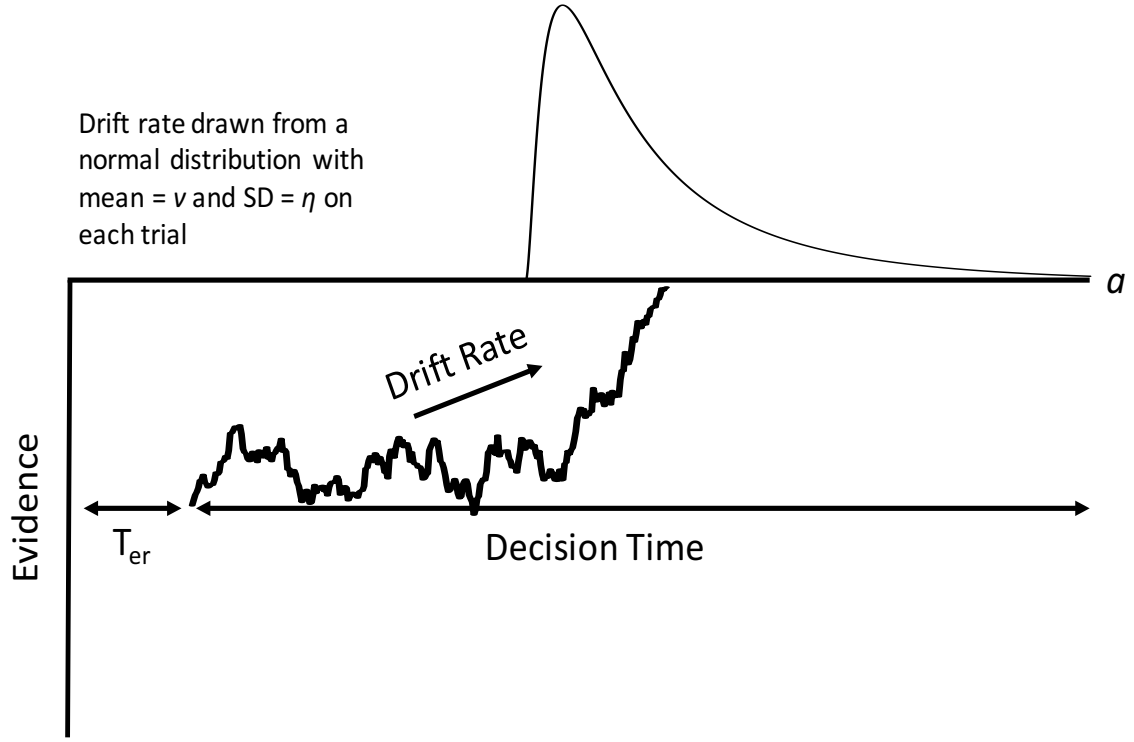


FIGURE 2.1: The single-bound diffusion model and its parameter values: response boundary (a), mean drift rate (v), between-trial variability in drift rate (η), and non-decision time (T_{er}).

towards the threshold according to a Brownian motion. The Wald distribution (Wald, 1947) describes the first passage times for Brownian motion with positive drift rate toward a positive response threshold. When the threshold is crossed response production is triggered. The time it takes to reach the response threshold is the decision time. Non-decision time, T_{er}, is added to the decision time to make up the total observed RT, so simple RT is described by a shifted-Wald distribution, with a shift equal to the non-decision time.

Ratcliff and Van Dongen (2011; see also Ratcliff & Strayer, 2014; Ratcliff, 2015) fit an elaborated version of the single-bound diffusion model, where on each trial the drift rate is sampled from a normal distribution with mean v and standard deviation η. When the sampled drift rates are strictly positive, the resulting mixture

of Wald distributions has an easily computed likelihood (see Equation 3 in Desmond & Yang, 2011). However, when the sampled drift rates can be negative the likelihood cannot be directly computed, and so Ratcliff and Van Dongen resorted to simulation methods. They were interested in negative rates because they can result in the threshold never being crossed, and so can account for failures to respond, which were common in their application; simple RT data from sleep-deprived participants.

The trial-to-trial rate variability discussed above gives the single-bound diffusion model more flexibility (Ratcliff & Van Dongen, 2011), yet also has a down side; it is only possible to identify two of the three parameters associated with evidence accumulation (i.e., the response threshold and the drift rate mean and standard deviation; see Ratcliff & Van Dongen, 2011). For completeness, we fit models both with and without trial-to-trial rate variability. Because accumulation rates in the diffusion model are sampled from a normal distribution they could be negative and accumulated evidence will not cross the response threshold, resulting in a failure to respond. However, as failures to respond were relatively rare in our data (3% of all trials), we assumed that drift rate variability followed a normal distribution truncated below zero. This truncation enabled the easy calculation of likelihoods, and consequently allowed us to use hierarchical Bayesian methods of estimation. This in turn allowed us to fit data sets with a relatively small number of observations per participant (114 per condition) based on the extra constraint afforded by hierarchical shrinkage effects (Shiffrin, Lee, Kim, & Wagenmakers, 2008).

2.1.2 The Cognitive Load Effects of Conversation

There is a correspondence between established measures of cognitive capacity (Townsend & Nozawa, 1995; Townsend & Eidels, 2011) and drift rates in choice RT tasks (Eidels, Donkin, et al., 2010). Increased cognitive load has been shown to have large effects on the tail of RT distributions in both choice (Shahar, Teodorescu, Usher, Pereg, & Meiran, 2014) and simple (Ratcliff & Strayer, 2014) RT tasks. Smaller drift rates and larger response thresholds are also known to lengthen the tail of RT distributions (Usher & McClelland, 2001; Ratcliff & McKoon, 2008; S. D. Brown & Heathcote, 2008; Matzke & Wagenmakers, 2009). Thus, it is tempting to conclude that increased cognitive load is related to and possibly even causes changes in drift rates and/or thresholds.

However, when researchers have used sequential sampling models to investigate cognitive load manipulations they have found that increased load affects a range of cognitive processes. Specifically, these studies found that increases in load either increase response thresholds (e.g., Heathcote, Loft, & Remington, 2015), trial-to-trial drift rate variability (McVay & Kane, 2012), or non-decision times (Shahar et al., 2014), or decrease drift rates (e.g., Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007; Sewell, Lilburn, & Smith, 2016). When the single-bound diffusion model with trial-to-trial rate variability was fit to data from a simulated driving task – where participants needed to press the brake to prevent a collision with a car in front – talking on a cell phone affected the drift rate and/or response threshold of drivers, but the effects could not be disentangled because of the aforementioned parameter identifiability issues (Ratcliff & Strayer, 2014).

To date, it is not clear how cognitive load imposed by passenger and cell phone conversation impacts the cognitive processes underpinning DRT performance. Our experiment investigated this issue by assigning pairs of participants to roles of a passenger or a driver in a high-fidelity driving simulator. Passengers were either seated next to the driver or in a separate room. In both cases they were instructed to converse casually with the driver, but to refrain from comments concerning the road. The latter stipulation aimed to remove a likely cause of the lack of DRT decrements noted by Drews et al. (2008) when the passenger was in the car; facilitation due to passenger-supplied warnings. However, other causes, such as timing of conversation to avoid conflict with safety critical events, may remain. Both driver and passenger were fitted with a DRT device (Strayer et al., 2013), as illustrated in Figure 2.2. The driver was requested to drive as normal but also to respond quickly and accurately to the DRT signal when they detected the red light in their visual field.

FIGURE 2.2: The DRT device used in the current study.

Cognitive load was manipulated across three conditions for the driver: a baseline where they were driving alone with no conversation, driving while talking with a passenger sitting next to them in the simulator, or driving while talking over a hands-free cell phone to a person in another room. We used a one-way Bayesian ANOVA (Morey, Rouder, & Jamil, 2014; Rouder, Morey, Speckman, & Province, 2012) to examine directly observed DRT performance. We hypothesized that drivers in the no conversation condition would respond more quickly to the DRT signal relative to driving while conversing over a cell phone. We also hypothesized that the decrements due to conversation could be larger with the hands-free cell phone relative to in-car conversation, but that this difference may be minimal due to our instruction to avoid comments about the driving task.

We then fit a set of single-bound diffusion models with different parameter settings instantiating different explanations of the effects of experimental manipulations in terms of drift rates, response thresholds, non-decision times, or any combination thereof. These competing explanations were compared based on the WAIC measure of out-of-sample prediction error (Watanabe, 2010; Gelman et al., 2014). In addition, we compared model fits with and without between-trial drift rate variability by comparing the predictive performance of models with the standard deviation of drift rates either fixed at zero or freely estimated.

2.2 Method

2.2.1 Participants

Forty undergraduate students at the University of Utah participated in this study in return for course credit (mean age = 23, 22 males). They all had normal or corrected to normal vision, and a valid drivers license.

2.2.2 Stimuli and Design.

The DriveSafety^{TM}DS-600 simulator was used in this experiment. The DS-600 consists of a Ford Focus cab surrounded by three large screens encompassing a 270° view. The simulated vehicle is based on the vehicular dynamics of a compact passenger sedan with automatic transmission. The driving scenario was designed using DriveSafety HyperDrive Authoring Suite. A two-way, four-lane interstate highway scenario was designed for this experiment. The roadway has four straight sections (10 miles each) connected by two wide-radius curves (1 mile each).

Both the driver and passenger were fitted with the DRT. The light diode was positioned an average 15° to the left and 7.5° above the participant's left eye and was held in a fixed position on the head with a headband (see Figure 2.2 again). RT to the DRT signal was recorded with millisecond accuracy via a button attached to participant's left thumb and encompassed the time between stimulus presentation and response.

There were a total of five within-subject conditions for each pair of participants, which were counterbalanced using a quasi Latin Square design. For drivers these were: (1) single (driving only) task, where drivers drove the simulated car and responded to the red light, but were not engaged in any type of conversation; (2) Dual task 'passenger' – driving and conversing with a passenger seated next to the driver; (3) Dual task 'mobile' – driving and conversing, through a mobile phone, with another person seated in a separate room. There were an additional two conditions for passengers; in the drivers' conditions (2) and (3) the passengers were also fitted with a DRT device, and asked to detect the red light.

2.2.3 Procedure

Participants drove on a simulated multi-lane freeway with moderate traffic, which had approximately 1500 vehicles in each lane per hour. Participants were given a five-minute practice session to familiarize themselves with the driving simulator. In each of the conditions, except for condition 1, the drivers and passengers were asked to speak and listen in equal proportions (i.e., 50% speaking and 50% listening).

In both the passenger and cell phone conversation conditions, participants were instructed to have a natural conversation as they would in real life; no restrictions about the topics covered in the conversation were provided to them. In conditions 3 and 5 the driver and passenger initiated a call via a hands-free Bluetooth earpiece. The volume was adjusted to a comfortable level before the experiment began.

The DRT task presented red lights every three to five seconds via the head mounted device. The lights were presented at exactly the same time for the driver and passenger. Both participants were required to respond to the light.

2.3 Mean RT Analysis

All RTs below 250ms (0.02%) were discarded, as were all RTs slower than 1000ms (0.2%). Using the R programming language (R Development Core Team, 2016), RTs were analyzed with a one-way JZS Bayesian ANOVA (Morey et al., 2014; Rouder et al., 2012), with a default setting for Cauchy priors and with subjects included as a random effect. The cognitive load manipulation was included as a main effect with levels – driver responding only (D), driver responding while talking with a passenger in the car (DP), driver responding while talking to the passenger on a cell phone (DC), passenger responding while with the driver in the car (PD), and passenger responding while talking to the driver on the cell phone (PC). The mean RTs (in seconds) in each condition were D = .466, DP = .502, DC = .505, PD = .461, and PC = .452.

We tested the main effect model against a null model which suggests no main effect on RT, reporting Bayes factors (BF_{10}), which quantify evidence in favor of the main effect model over the no main effect model as a ratio. For example, when $BF_{10} = 5$ the observed data are 5 times more likely under the main effect model than under the no main effect model. When $BF_{10} = 1/5 = .2$ the observed data are 5 times more likely under the no main effect model than under the main effect model. The Bayes factor ANOVA revealed that the cognitive load main effect model was

preferred to the null model by a Bayes factor of 4.8. Thus the data provide positive evidence (Kass & Raftery, 1995) against the hypothesis of no main effect on RT.

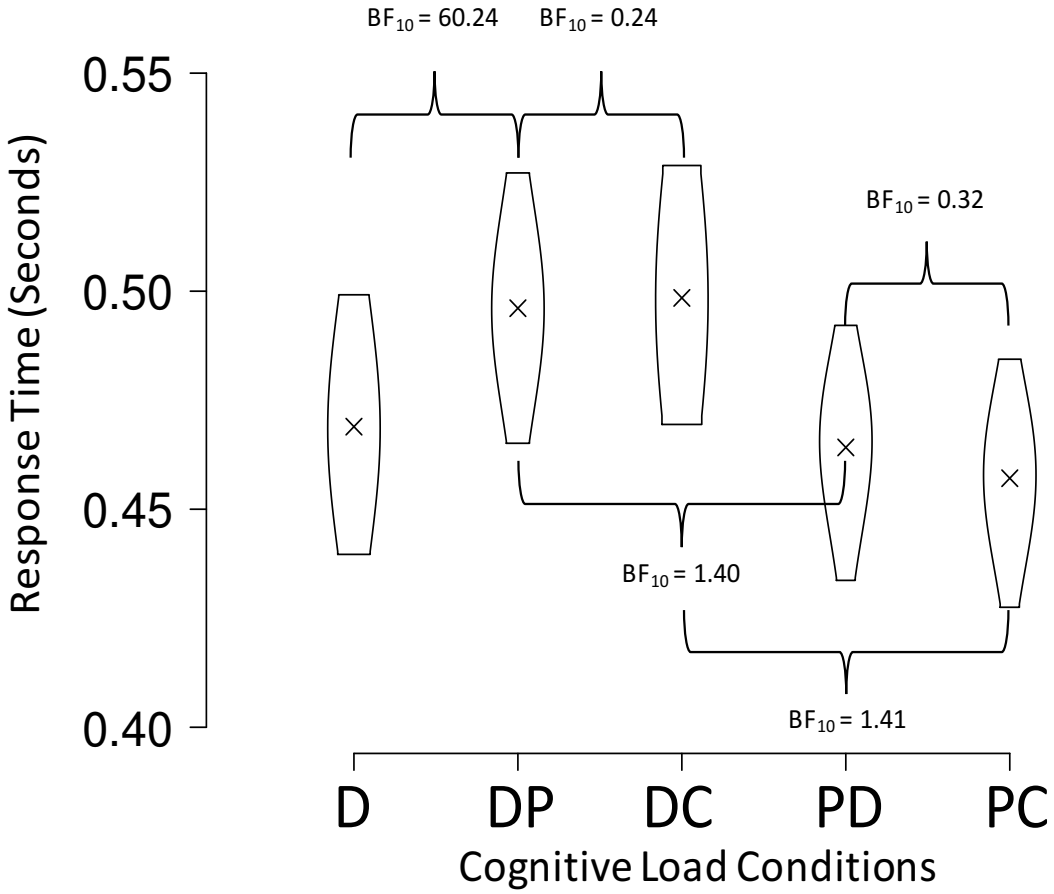


FIGURE 2.3: Violin plots of predicted response time data from the one-way JZS Bayesian ANOVA. Violin plots include an ×, which marks the median RT, and mirrored on either side are rotated kernel density plots of the 95% highest density interval of each posterior distribution. Superimposed are Bayes factors from post-hoc paired-samples t-tests. The 5 cognitive load conditions were driver responding only (D), driver responding with a passenger in the car (DP), driver responding while talking to the passenger on a cell phone (DC), passenger responding with the driver in the car (PD), and passenger responding while talking to the driver on the cell phone (PC).

We conducted post-hoc Bayesian paired-samples t-tests to see which conditions differed from each other, with detailed results reported in Figure 2.3. There was strong evidence that the driver's responses to the DRT signal were slower when conversing (DP vs. D), indicating the DRT was sensitive to the additional cognitive load. There was positive evidence for no difference in the driver's RT as a function

of the passenger's location (DP vs. DC). There was also positive evidence for no difference in the passenger's RT between locations (PD vs. PC). To test if driving affected DRT performance we compared the corresponding driver and passenger conditions; Figure 2 shows a clear trend whereby driver responses (DP and DC conditions) were slower than the corresponding passenger conditions (PD, PC), but the statistical evidence was equivocal (BF of 1.40 and 1.41), likely indicating substantial individual differences.

2.4 Model-Based Analysis

Details about the hierarchical Bayesian model fitting routine are presented in the Appendix A. To select between competing models we measured how well each model could predict future data using WAIC. WAIC includes a goodness of fit value and a measure of the model's complexity. Model complexity is subtracted from the goodness-of-fit measure to approximate an unbiased estimate of the model's out-of-sample prediction error. When comparing models, the model with the lower WAIC value is better able to predict future data.

We fit 10 separate single bound diffusion models to the DRT data, five models with trial-to-trial rate variability (η) fixed at zero and five models where we estimated η. Plots presented in Appendix A show that all models provide an accurate account of the data. However, WAIC (Table 2.1) did indicate a clear preference for allowing both response threshold and non-decision time to vary over cognitive load conditions. Most importantly, models in which the mean rate explained the effect of the experimental factor were strongly rejected.

There was very little difference between the versions of the response threshold and non-decision time model with and without trial-to-trial rate variability, with the model with η fixed at zero slightly preferred. We will refer to the latter model as the winning or $a + T_{er}$ model, and focus our further analysis on it given it is both more parsimonious and less effected by parameter identifiability issues than the alternative.

Table 2.1 also provides a measure of the best fit for each model, the deviance of the mean of all posterior parameter estimates. The two $a + T_{er}$ models have the best fit among the five model that share their assumption about η, showing that their advantage in WAIC is not purely due to being more parsimonious. The addition of trial-to-trial rate variability hardly improves the fit of the $a + T_{er}$ model variants, consistent with generally small estimates of the η parameter for that model (mean of group level mean posterior of $\eta = 0.177$). In Appendix A we present hazard functions, which provide evidence against the inclusion of trial-to-trial rate variability, further confirming our selection of the simpler model with $\eta = 0$.

TABLE 2.1: WAIC results, number of effective parameters and deviance of the posterior mean for all models.

Model	WAIC (η)	Effective Parameters	Deviance	WAIC ($\eta = 0$)	Effective Parameters	Deviance
a $\sim$ F & v $\sim$ 1 & $\eta \sim 1$ & $t_{er} \sim 1$	-19295	130.8	-19428	-19310	128.5	-19441
a $\sim$ F & v $\sim$ 1 & $\eta \sim 1$ & $t_{er} \sim$ F	-19476	180.6	-19665	**-19477**	178.3	-19664
a $\sim$ 1 & v $\sim$ F & $\eta \sim 1$ & $t_{er} \sim 1$	-18897	131.2	-19030	-18873	129.8	-19005
a $\sim$ 1 & v $\sim$ F & $\eta \sim 1$ & $t_{er} \sim$ F	-19329	197.7	-19536	-19321	196.5	-19526
a $\sim$ 1 & v $\sim$ 1 & $\eta \sim 1$ & $t_{er} \sim$ F	-18958	137.0	-19102	-18943	133.0	-19083

Note. Bold WAIC value indicates the preferred model.

$\sim$F indicates different parameter estimates were allowed for each level of the experimental factor.

$\sim$1 that the same value was estimated for all levels.

Table 2.2 shows median values of the group-level mean posterior distributions for response threshold and non-decision time from the $a+T_{er}$ model with $\eta = 0$.

We used Bayesian predictive p-values to statistically test for differences between the posterior distributions (Meng, 1994). We calculated the difference between subject level posterior distributions, or plausible values (Marsman, Maris, Bechger, & Glas, 2016), and then averaged the differences over subjects.

TABLE 2.2: Median values of the group-level mean posterior distributions from the winning $a + T_{er}$ model

	D	DP	DC	PD	PC
a	1.298	1.444	1.450	1.287	1.282
T_{er}	.177	.183	.182	.181	.166
v	4.589	-	-	-	-

Note. T_{er} values are in seconds, and the same v applies for all conditions.

We calculated the probability (p-value) that the difference distributions were equal to or less than 0. Similar to the traditional p-value, a low predictive p-value indicates a low probability of observing this or more extreme data if the null hypothesis was true. The response threshold increased from the D to the DP condition ($p < .001$) and from the D to the DC condition ($p < .001$), suggesting that drivers were more cautious when cognitive load increases. Thresholds are comparable between the DP and DC conditions ($p = .46$), indicating the same level of elevated caution for cell phone and in-car conversations. Drivers had higher thresholds than passengers when the passenger was in the car ($p < .001$) and when they were talking over the cell phone ($p < .001$), which further suggests that increased cognitive load increases response thresholds in the DRT.

Non-decision time for passengers on the phone outside the simulator was 15ms faster compared to passengers talking in person in the simulator ($p = .06$), but no other non-decision time parameters were different from each other (all $p > .23$).[1] This could be due to the added information that the passenger could see in a dynamic simulator environment compared to the static room from which they conversed over the phone, and so is not relevant to our focus here on cognitive load effects.

2.5 General Discussion

In our experiment, pairs of participants, assigned as either driver or passenger, took part in a driving simulator and detection response task (DRT) simultaneously. In the DRT, participants were required to respond when a small light appeared in their peripheral vision. The driver completed the DRT in three different conditions: by themselves in the simulator, talking with a passenger in the simulator, and talking to a passenger (who was outside of the simulator) on a cell phone. We recorded the response times (RT) in the DRT from both the driver and passenger.

RTs in the DRT are a validated measure of cognitive load (International Organization for Standardization, 2015), where slower RTs represent increased cognitive load. We found that drivers had slower RTs when they were conversing with a passenger in person or over the phone compared to when they were by themselves

[1]We confirmed that the T_{er} effect was genuine by fitting another model of simple RT (see Appendix A), the log-normal race (Heathcote & Love, 2012).

– suggesting that both types of conversation increased cognitive load. We modeled the DRT behavioral data with the single-bound diffusion model to determine if longer RTs, which reflect increased cognitive load, are due to differences in drift rates, response thresholds, or non-decision times.

We found that the cognitive load effect on DRT performance was due to an increase in the participant's response thresholds, but no evidence of an effect of cognitive load on the time to encode stimuli and to produce responses or on the rate of evidence accumulation. In contrast to Ratcliff and Van Dongen (2011), but consistent with Anders et al. (2016), we did not find it necessary to allow for variability in the rate of evidence accumulation from trial to trial to provide a good account of our DRT data.

Our findings may at first seem surprising because they are not in line with the capacity sharing account of DRT and driving performance (e.g., Strayer et al., 2011, 2013). However, separate pools of capacity for DRT and driving is consistent with the finding that having to preform the DRT does not adversely impact driving (Strayer et al., 2014). Why then is the DRT a sensitive measure of cognitive load? We suggested that it may be because of a general tendency for people to be more cautious when under increased cognitive load, but further work is required to better understand the processes underlying threshold adjustments, and why they occur.

One possibility is that the process is consciously mediated, with participants slowing in both the DRT and driving task because they deliberately set higher threshold for the secondary DRT task when they perceive they are subject to a higher workload in the primary driving task. This possibility is consistent with

the strong correlation found between DRT decrements and self-report measures of subjective workload (Strayer et al., 2013), such as the NASA Task Load Index (Hart & Staveland, 1988).

Alternately, threshold increases may occur to reduce the chance of response conflicts (i.e., one response preempting another), as suggested by *delay theory* (Loft & Remington, 2013) of dual task costs in prospective memory tasks (Heathcote, Loft, & Remington, 2015). Such conflicts may not necessarily be peripheral in nature; they could also be mediated by response gating for both tasks being handled in the same brain area, such as the basal ganglia (Forstmann et al., 2008), again with priority given to the primary driving task.

Many cognitive tasks require a decision between two or more alternatives and record both the choice and the time to make that choice. In such choice RT data, threshold and rate effects are relatively easy to disambiguate as they have opposite effects on accuracy and RT (i.e., a higher threshold increases accuracy and RT, whereas a higher drift rate increases accuracy but decreases RT). In simple RT, in contrast, these effects are differentiated only by relatively subtle effects on the distribution of RT. Although tests based on out-of-sample predictive accuracy clearly favored a threshold account of cognitive load effects, and the corresponding model produced clear and sensible effects on threshold estimates, it would be prudent in future work to seek converging evidence about our somewhat surprising findings.

One potential way forward is to examine the effects of speed vs. accuracy instructions, which are usually assumed to selectively affect response thresholds (but see Rae, Heathcote, Donkin, Averell, & Brown, 2014). Another possibility is

to compare cognitive load effects on the traditional (simple RT) DRT and a version requiring a choice response, which should allow for a stronger comparison of rate vs. threshold models. A choice version of the DRT (e.g., respond 'A' for a green light, 'B' for a red light) could offer more nuanced tests of cognitive load effects as long as it does not have a detrimental impact on driving performance.

Chapter 3

A Linear Ballistic Model of Vowel Perception

Reference

Tillman, G., Benders, T., Brown, S. D., & van Ravenzwaaij, D. (Under Review). An Evidence Accumulation Model of Acoustic Cue Weighting in Vowel Perception. Journal of Phonetics

3.1 Introduction

Phonemes are linguistic representations with an acoustic counterpart that can be characterized in a multidimensional acoustic space. Values along each acoustic dimension can serve as cues for listeners to recognize a speech sound as a particular phoneme. The cues, such as first (F1) and second (F2) formant frequency, duration, and fundamental frequency, are acoustic and continuous. Yet, these cues map onto phonological representations that may not be continuous, i.e., the phonemes. Phonemes can be viewed as clusters of exemplars in a multidimensional phonetic space (Pierrehumbert, 2001), or as abstract representations that are connected to a range of values along multiple phonetic dimensions (Boersma, 2007). Speech perception is the process of mapping the continuous acoustic information onto the phonological categories (Holt & Lotto, 2010).

Each phoneme correlates with multiple acoustic dimensions (Lisker, 1986) and multiple acoustic cues influence each phoneme categorization (Holt & Lotto, 2006). Some cues contribute strongly to a listener's decision and some cues contribute weakly to the decision – a phenomenon called *cue weighting*. Cue weighting in speech perception often reflects the reliability of the cues for the recognition of phonological categories in the ambient language (Holt & Lotto, 2010).

Researchers investigate cue weighting using a range of methods: computational statistical modeling (Toscano & McMurray, 2010; McMurray, Aslin, & Toscano, 2009), eye-tracking (Reinisch & Sjerps, 2013), neuro-physiological measurements (Lipski, Escudero, & Benders, 2012), in normal-hearing and hearing-impaired

populations (Winn, Chatterjee, & Idsardi, 2012; Winn, Rhone, Chatterjee, & Idsardi, 2013), and most commonly, with behavioral data from phoneme categorization tasks (Repp, 1982). In the latter, researchers systematically vary the acoustic cue values of sounds that are played to participants and observe the effects on phoneme categorization. Cue weighting is measured by how much each cue contributes to the categorization response and is therefore based on a measure at the end of processing and decision-making. To use categorization data to learn how acoustic cues are connected with phonological categories, we have to make the assumption that categorization data directly reflects the mapping of the experimentally manipulated cues onto the phonological categories. However, there are two fundamental issues with this assumption.

The first problem is that cue weighting is measured for a phoneme contrast and does not give us the association between cues and each category separately (i.e., the cue-to-*one*-phoneme mapping). For example, a cue that is strongly associated with one phoneme in the contrast and only loosely associated with the other phoneme can appear to be indiscriminately 'heavily weighted', because the cue contributes relatively strongly to the decision between these two phonemes. Given this confound, it is difficult to infer how much each acoustic cue contributes to each individual phoneme in the contrast.[1]

The second problem is that researchers only observe the association between experimentally manipulated cues and overt behavioral responses (i.e., the

[1]We are interested in how much each cue contributes to each phoneme in the contrast, which is not the same thing as investigating how much an acoustic cue contributes to a particular phoneme outside the context of the contrast.

cue-to-response association), which means they need to assume that this association directly reflects the cue-to-phoneme mapping. Yet, a strong cue-to-phoneme mapping may not manifest as a strong cue-to-response association. One reason for a weak association between cues and responses despite a strong mapping could be that listeners do not have good access to the cue. Perhaps the cue is not always loud enough to be perceived or perhaps the cue appears late in the speech signal. Cues that appear later in the signal might be strongly associated with a phoneme, but may not appear as such in a categorization task because earlier appearing cues have already been processed and potentially determined the response (cf. McMurray, Clayards, Tanenhaus, & Aslin, 2008; Reinisch & Sjerps, 2013). In order to address this issue, it is necessary to learn more about how listeners process the acoustic cues. For instance, is cue weighting as inferred from categorization data driven by differences in when cues are available in time, or by listeners processing one acoustic cue faster than another? In any case, researchers need a way to investigate such latent processes in order to derive more accurate conclusions about acoustic cue weighting in terms of cue-to-phoneme mapping.

Both problems limit our ability to use categorization data to learn about how listeners map acoustic information onto phonological categories. Therefore, we need a method to account for how acoustic cues are cognitively processed for each phoneme in the contrast. Below we discuss response times (RT) and eye-tracking, which are alternative measures to categorization data that give insight into the processing of acoustic information, but neither of these measures address both issues.

First, researchers can use the RT associated with phonological decisions to

investigate phoneme perception. For example, researchers have investigated processing differences between non-identical and identical phonemes (Pisoni & Tash, 1974) and have determined that phoneme categorization decisions depend more on a phoneme's position in acoustic space than their perceived category goodness (Miller, 2001).

However, there are difficulties with analyzing either choice data or RT in isolation. We know that the accuracy of a decision depends on how fast the decision is made – in other words, a participant's speed-accuracy trade-off setting (e.g., Wickelgren, 1977; Luce, 1986; Heitz, 2014). Without any insight into the trade-off settings used by participants, researchers may draw incorrect conclusions from choice or RT data alone. Furthermore, to analyze RT researchers typically average over all observations for each participant in order to subject the means to a statistical test, such as ANOVA. Analyzing the RT in this manner can lead to researchers drawing incorrect conclusions (e.g., Ashby, Maddox, & Lee, 1994; Curran & Hintzman, 1995; Heathcote, Brown, & Mewhort, 2000) and does not allow researchers to learn about the latent cognitive processes involved in speech perception. For example, an RT of 700ms on a given trial suggests that 700 ms was needed to perceptually encode the sound, decide what phoneme was heard, and execute a motor response. But, we cannot know how long each of these processes takes from analyzing mean RT with linear models. Given that RT is a measure at the end of processing, analyzing RTs alone only inform researchers about the cue-to-response association but not the cue-to-phoneme mapping.

Eye tracking is a another useful measure that is frequently used to observe

how listeners process experimentally manipulated cues online (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). For example, eye-tracking can be used to infer whether the order in which acoustic cues become available to listeners affects listeners' interpretation of the speech signal (McMurray et al., 2008; Reinisch & Sjerps, 2013). In fact, McMurray et al. (2008) showed that listeners do not wait for cues that are available later in a speech signal (e.g., vowel duration) to begin using earlier available cues (e.g., voice onset time). Moreover, Reinisch and Sjerps (2013) showed that listeners use vowel spectral cues before vowel duration cues, because listeners need to wait for the vowel offset before they have full information about the duration.

Eye-tracking data, like RTs, are typically averaged over all observations for each participant, meaning that the aforementioned objections against inferences from averaged data hold for eye-tracking data as well. Furthermore, eye-tracking data are subject to the first confound of categorization data discussed in detail above. That is, they can give insight into cue-weighting, but do not give the cue-to-one-phoneme mapping for phoneme contrasts.

Categorization, RT, and eye-tracking are all useful methods in speech perception research, but none of them address both the cue-to-one-phoneme mapping and the cue-to-phoneme mapping issues discussed above. In this paper, we advocate the simultaneous analysis of phoneme categorization data with their associated RTs using a sequential sampling model (e.g., Ratcliff & McKoon, 2008; Usher & McClelland, 2001; S. D. Brown & Heathcote, 2008). The following section describes the sequential sampling model we use are and what it can add to the current speech perception literature.

3.1.1 The Linear Ballistic Accumulator

A parsimonious sequential sampling model that retains all of the explanatory power of more complex models (e.g., Ratcliff & Rouder, 1998; Usher & McClelland, 2001), while having the advantage of being tractable, is the linear ballistic accumulator (LBA; S. D. Brown & Heathcote, 2008).[2] The LBA has been applied to a number of perceptual discrimination paradigms (e.g., Ho, Brown, & Serences, 2009; Forstmann, Brown, Dutilh, Neumann, & Wagenmakers, 2010; Forstmann et al., 2008; Cassey, Heathcote, & Brown, 2014; van Ravenzwaaij, Provost, & Brown, 2016) and has been fit to tasks where the responses are categories (e.g., ?, ?; Trueblood, Brown, & Heathcote, 2014), which is the same set-up as used in a phoneme categorization task. Therefore, the LBA can be feasibly extended to model phonological decisions in categorization tasks that yield choice data and RTs.

Suppose a participant needs to identify whether they have heard /ɑ/ or /aː/ – a schematic of the LBA explanation for this task is shown in Figure 3.1. At the beginning of a trial a participant hears a sound through a pair of headphones. It takes the listener time to perceptually encode the sound. After perceptual encoding, the evidence from the stimulus serves as input for the decision process. The LBA does not commit to what evidence is sampled from stimuli, only that the evidence leads to a response. In the discussion we suggest that one possible interpretation of the sampled evidence is the strength of the mapping between the acoustic information and the phoneme category associated with the response option. The evidence drives the decision process, which involves independent evidence accumulators for each

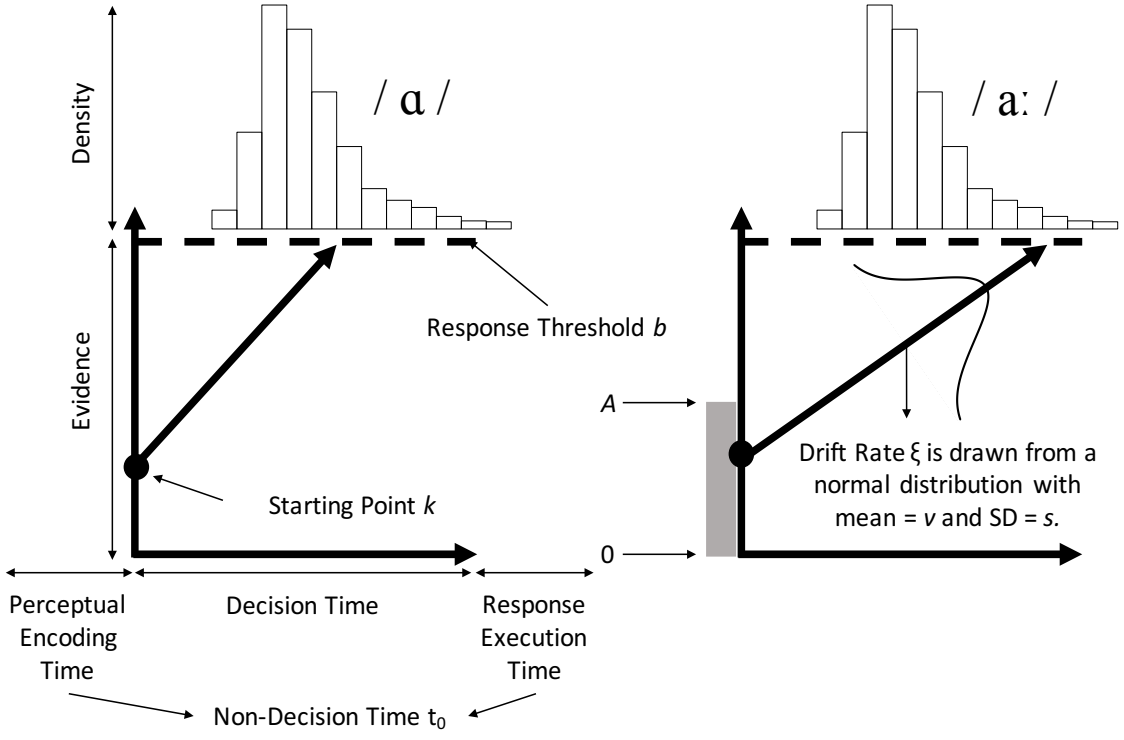
[2]There is a closed form expression for the likelihood function, which allows for relatively quick model fitting.

response option. Each accumulator has a starting point of evidence accumulation k, which is a random value between 0 and A on each trial, a drift rate ξ that specifies the rate of evidence accumulation, and a response threshold b that specifies the amount of evidence needed to make a decision. From k, both accumulators independently race towards b at their respective drift rates. On each trial, drift rates are sampled from a normal distribution with mean drift rate v and standard deviation s.[3] The first accumulator to reach the response threshold determines the decision that is made. The time it takes for the LBA process to go from the starting point to the response threshold is the decision time. After a decision is made, the participant will need to overtly execute their response with a key press – the time needed for this overt response is the response execution time. The sum of the perceptual encoding time and response execution time makes up the non-decision time t_0. The total RT on a given trial is the sum of the decision time and the non-decision time. For convenience, an overview of the LBA parameters is provided in the table at the bottom of Figure 3.1.

3.1.2 Measuring Cognitive Processes From Behavioral Data

The LBA divides response times into decision times and non-decision times. The decision time is re-expressed as the amount of evidence needed for a response divided by the speed of evidence accumulation, which can be formally expressed as $\frac{b-k}{\xi}$. With the LBA we can go beyond the temporal aspects of the decision process

[3]Human performance and the activity of neuron populations result in highly variable behavior in experiments, even when participants are presented with the same stimuli (see Usher & McClelland, 2001, for a discussion of sources of variability in sequential sampling models). In addition, these variability parameters explain key aspects of RT data, such as slow error RT (Ratcliff, 1978) or fast error RT (Laming, 1968) relative to correct RT.



Parameter labels and descriptions.

Symbol	Name	Psychological Interpretation
v	Mean Drift Rate	Speed of information processing
b	Response Threshold	Evidence required to make decision
A	Starting Point Variability	Variability in the initial evidence across trials
s	Drift Rate Variability	Variability in the speed of processing across trials
t_0	Non-Decision Time	Time for processes other than decision processing

FIGURE 3.1: The linear ballistic accumulator model and its account of choosing between /ɑ/ or /aː/. The top left panel shows the accumulator corresponding to the /ɑ/ response. The top right panel shows the accumulator corresponding to the /aː/ response. In the bottom panel we provide labels and descriptions for each LBA parameter.

by estimating drift rate. The drift rate is governed by the quality of evidence being sampled from the stimulus, with larger drift rates meaning faster speed of processing. We can also estimate non-decision time or response threshold and learn about the time needed for processes outside of the decision process or the response caution or bias associated with a participant's decision.

But how can we get unique estimates of non-decision time, drift rate, and response threshold from behavioral data? We can recover unique estimates because these parameters have unique behavioral signatures in response proportions and RT distributions (Ratcliff & McKoon, 2008). For example, the non-decision time parameter determines the smallest possible RT and also shifts the entire RT distribution, yet this parameter has no effect on the response proportions. Changes in mean drift rate cause small changes to the leading edge of the RT distribution – the fastest responses – and large changes to the tail – the slowest responses. In contrast, changes in response thresholds also shift the leading edge and cause relatively small changes to the tail. Higher mean drift rate leads to faster RTs in combination with a higher accuracy, whereas lower response thresholds lead to faster RTs with lower accuracy.

There are several benefits to using the LBA to analyze choice and RT data from a phoneme categorization task. First, the adjustment of response thresholds in the LBA is an intuitive account of the speed-accuracy trade-off, which cannot be addressed by linear models of choice or RT. Second, rather than analyzing averaged data that can lead to researchers drawing incorrect conclusions (e.g., Ashby et al., 1994; Curran & Hintzman, 1995; Heathcote et al., 2000), the LBA can be used to analyze entire RT distributions for all responses. This analysis decomposes the behavioral data into their underlying constituent components of processing – such as drift rates, response thresholds, and non-decision times. Finally, the LBA reconciles the fact that continuous acoustic information leads to categorical decisions by explicitly describing phoneme decision-making as a result of an evidence accumulation process. Taken together, the LBA allows us to address both the cue-to-one-phoneme mapping and the cue-to-phoneme mapping issues outlined in the

introduction. Specifically, we can learn about the cue-to-one-phoneme mapping by investigating the evidence accumulation dynamics in each accumulator, which represent processing for each phoneme response option. We learn about the cue-to-phoneme mapping by investigating how the parameters of the LBA model are affected by changes in acoustic information.

3.1.3 The Current Study

Spectral quality and duration are two of the acoustic cues that listeners can use to categorize vowels (e.g., Bohn & Flege, 1990; Flege, Bohn, & Jang, 1997; Adank, Van Hout, & Smits, 2004; Gerrits, 2001; Escudero, Benders, & Lipski, 2009; Reinisch & Sjerps, 2013). For example, first language (L1) English speakers weigh static spectral cues more than duration cues as they mostly use properties in the F1 and F2 to recognize the contrast between /iː/ and /ɪ/ and between /æ/ and /ɛ/ (Flege et al., 1997). On the other hand, second language (L2) speakers of English with L1 German mostly use the duration of the stimuli to recognize vowels in those same contrasts (Bohn & Flege, 1990). Similarly, L1 Dutch listeners weigh spectral cues heavier than duration cues to distinguish between /ɑ/ and /aː/, whereas Turkish and Spanish L2 learners of Dutch weigh vowel duration heavier than spectral quality (Nooteboom & Cohen, 1984; van Heuven, Van Houten, & De Vries, 1986; Escudero et al., 2009; van der Feest & Swingley, 2011).

Here, we analyze data from an experiment in which L1 Dutch listeners categorize synthetic vowels as the Dutch short and closed /ɑ/ and the long and open /aː/. /ɑ/ and /aː/ are a useful set of stimuli as these vowels are typically

realized with both spectral and duration differences (Adank et al., 2004). These two vowels are the only two low vowels in Dutch, which are each other's closest neighbors in the acoustic vowel space defined by F1 and F2 and alternate in the singular–plural pairs of some nouns (e.g., /pɑd/ - /pɑːdə/) as well as verbs (/kʋɑm/ - /kʋaːmə/).

We will subject the categorization and RT data to Bayesian logistic regression and Bayesian ANOVA (Rouder et al., 2012), respectively. These more traditional analyses will attempt to replicate the relatively heavier weighting of vowel quality compared to the duration of a vowel, which is typically observed in categorization tasks involving the Dutch contrast between /ɑ/ and /aː/ (e.g., Nooteboom & Cohen, 1984; van Heuven et al., 1986; Escudero et al., 2009; van der Feest & Swingley, 2011; Reinisch & Sjerps, 2013). Then, we will analyze both RT and accuracy simultaneously with the LBA. The LBA analysis will test if spectral quality and duration affect the drift rates of participants when they are categorizing /ɑ/ and /aː/. We will also test if the effects of the acoustic cues on drift rates as well as the participants' response thresholds are different across the two vowels. Finally, and in line with Reinisch and Sjerps (2013), we will test if longer vowel durations systematically increase the time needed for perceptual encoding, which will manifest as longer non-decision times for sounds with longer durations.

3.2 Method

3.2.1 Participants

Thirty participants were tested (5 males and 25 females) at the Radboud University Nijmegen, the Netherlands. All participants had normal or corrected to normal vision and reported no hearing impairments. They were native Dutch speakers. All participants received monetary reimbursement for their participation.

3.2.2 Materials

One-hundred different vowel stimuli were used. They were synthetic isolated vowels covering a 10×10 matrix ranging from the typical /ɑ/, with low formants and a short duration, to a typical /aː/, with high formants and a long duration. The F1 and F2 values as well as the duration of each of the 10 steps are presented in Table 3.1. The spectral quality of the typical /ɑ/ and /aː/ was based on production data from 50 male speakers (Pols, Tromp, & Plomp, 1973), while the duration values were based on 10 male speakers (Adank et al., 2004), following the stimulus creation procedure in Escudero et al. (2009). The sounds had a falling fundamental frequency from 150Hz to 100Hz, to simulate male speech.

The 100 stimuli used in the experiment were synthesized in the computer program Praat (Boersma, 2002). The difference between two consecutive steps on the duration dimension equals one-tenth of the difference between the logged duration (ms) of the typical /ɑ/ and /aː/. The difference between two consecutive steps on the spectral quality dimension equals one-tenth of the difference between the

typical /ɑ/ and /aː/ spectral quality in mel. By generating stimuli in this way we can approach equal psychoacoustic scaling across the steps within each dimension, although not necessarily across dimensions. We can compare cue weighting across dimensions as the dimension endpoints were based on typical values of the vowels produced in language.

TABLE 3.1: F1, F2, and duration values, which were crossed factorially to generate all 100 stimuli.

Step	Spectral Quality (F1/F2 Hz)	Duration (ms)
1	687/1099	96
2	699/1121	104
3	711/1143	113
4	723/1165	123
5	736/1188	134
6	749/1211	146
7	762/1235	158
8	775/1259	172
9	788/1283	187
10	801/1308	203

The experiment was run in a sound attenuated quiet booth using Dell Precision T3600 computers, with an Intel Xeon Processor E5-1620 and a Sound Blaster ZX Gamer audio card. Stimuli were played over Sennheiser HD 215 MKII DJ headphones.

3.2.3 Procedure

On each trial, participants heard a sound over headphones and saw the orthographic symbols for /ɑ/ ("a") and /aː/ ("aa") on the left and right of the computer screen. The position of the symbols corresponded to the response key side, which was fixed for each participant and counterbalanced across participants. Participants were required to respond by pressing one of the two response keys within 2500ms after stimulus onset, otherwise they were presented with 'Te Langzaam' (Too Slow), which remained on screen for 3000ms. Once a response was made the next trial started immediately. The 100 unique stimuli (10 formant steps × 10 duration steps) were played once per block in randomized order. Participants heard 5 blocks, for a total of 500 trials. Participants were allowed to take a short break after each block and could continue to the next block when ready. The experiment started with 10 practice trials that only presented the stimulus with the lowest F1 and shortest duration, a typical /ɑ/, and the stimulus with the highest F1 and longest duration, a typical /aː/. The experiment was run in the software PsychoPy (Peirce, 2007).

3.2.4 Behavioral Data Analysis

Before analyzing the response proportion and RT data together with the LBA, we ran the more traditional analyses on both dependent measures separately. The proportions of /aː/ responses were analyzed with a Bayesian logistic regression model and the RTs were analyzed with a two-way Bayesian ANOVA. Both analyses included subjects as a random effect, meaning that each subject had their own intercept, which was drawn from a normal distribution. Post-hoc Bayesian

paired-samples t-tests were conducted where appropriate. The logistic regression was carried out using the R Stan package (Stan Development Team, 2016) in R (R Development Core Team, 2016) and the ANOVA analyses was carried out using JASP (The JASP Team, 2016; Morey et al., 2014; Rouder et al., 2012).

The main reason for using Bayesian linear models instead of frequentist alternatives is to allow calculation of Bayes factors (BF), which we motivate below. In addition, Bayesian models also allow for researchers to calculate posterior distributions of parameters. Posterior distributions provide a range of plausible values as well as their corresponding probabilities. They have a similar purpose to standard errors of estimation, but do not make the assumption that estimation error is symmetrical and normally distributed. This assumption is often incorrect, especially when dealing with data that are themselves not normally distributed.

For the behavioral data analysis (with the exception of the logistic regression), Bayes factors were used in place of conventional p values, as Bayes factors are arguably more appropriate for assessing statistical evidence (see Wagenmakers, 2007). We refer the interested readers to Kruschke (2011) and Lee and Wagenmakers (2013) for accessible introductions to Bayesian statistics for social scientists. Bayes factors represent "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378). They 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. To determine the evidence for a particular effect (e.g., spectral quality), we calculated an inclusion

BF ($BF_{\text{Inclusion}}$). The $BF_{\text{Inclusion}}$ statistic represents the evidence in favor of models that include a particular effect in relation to models that do not include the effect. If we were testing whether spectral quality had an effect on RT, for instance, and we obtained $BF_{\text{Inclusion}} = 5$, then the data are 5 times more likely to come from a model with a spectral quality effect than a model without a spectral quality effect. Bayes factors greater than 3 or less than 1/3 will be considered positive evidence for the alternative and null hypothesis, respectively (Kass & Raftery, 1995). And Bayes factors less than 3 or greater than 1/3 will be considered inconclusive evidence for both hypotheses.

The LBA models, which we describe next, and the logistic regression models were evaluated on how well they predict future observations – the out-of-sample predictive error. The gold standard for estimating the out-of-sample predictive error of a model is cross-validation (Geisser & Eddy, 1979). Cross-validation is computationally expensive and so we used a computationally faster approximation: the widely applicable information criterion (WAIC; Watanabe, 2010; Gelman et al., 2014). The WAIC balances goodness of fit against model complexity. This measure is calculated from a model fit value and a model complexity penalty value, which approximates the number of effective parameters of the model. In this sense, WAIC is similar to the BIC (G. Schwarz, 1978) and AIC (Akaike, 1974) measures, but WAIC extends these by quantifying model complexity as across-sample variability in model fit rather than simply counting up the number of free parameters. The method we used is described in detail by Vehtari et al. (2016). The model with the lower WAIC value has better out-of-sample predictive error and is therefore the preferred model.

3.2.5 Linear Ballistic Accumulator Analysis

We used hierarchical Bayesian methods to estimate the parameters of the LBA model. The model fitting details are outlined in the Appendix B.

Here we give a brief description of the different LBA models we implemented. Each LBA model had two accumulators, one corresponding to /ɑ/ and another corresponding to /aː/. Each model had the parameters mean drift rate v, response threshold b, starting point variability A, non-decision time t_0, and drift variability s. The s parameter serves as the scaling parameter, which we set to 1 for the /ɑ/ accumulator and estimated for the /aː/ accumulator. Having a scaling parameter allows all other parameter values to be identified, because their values are now relative to the scaling parameter (Donkin, Brown, & Heathcote, 2009). All other parameters were estimated, but fixed across accumulators, with the exception of v and response threshold b.

To help describe all models considered, we present an LBA visualization in Figure 3.2. In the Figure, blue arrows represent changes in drift rate across the duration values, green arrows represent changes in drift rate across the spectral values, red arrows represent changes in non-decision time across the duration values, and gray arrows represent changes in response threshold across vowels.

The first model we tested, the equal drift model, allowed v to change across the spectral quality values and duration values. This model assumed that spectral quality and duration influenced the speed of information processing, but their effects were equal for /ɑ/ and /aː/ responses. In Figure 3.2, we would see that blue arrows

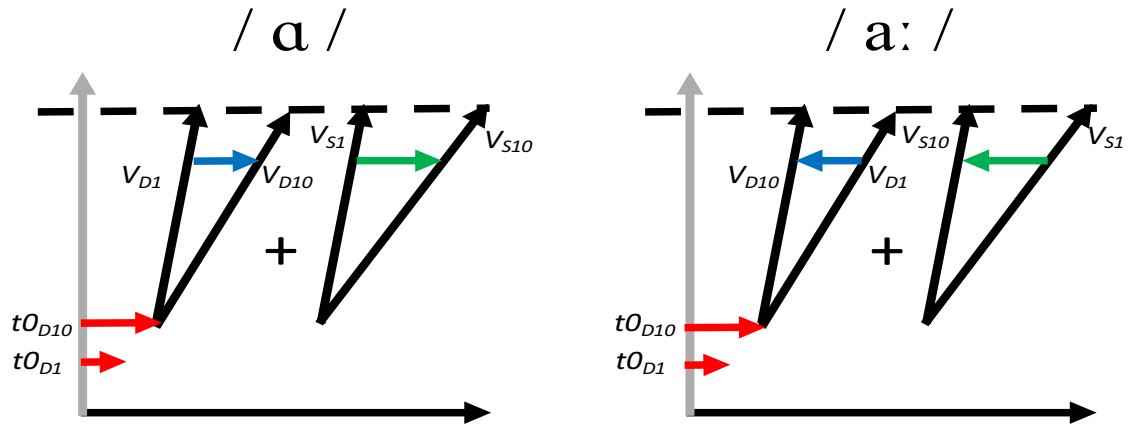


FIGURE 3.2: An example of a two-accumulator LBA model. The left panel shows the accumulator corresponding to the /ɑ/ response. The right panel shows the accumulator corresponding to the /aː/ response. In each accumulator we present 4 sloped lines that represent mean drift rates for the 1^{st} and 10^{th} duration values and 1^{st} and 10^{th} spectral values. The changes in slope correspond to changes in mean drift rate. The changes in mean drift that are induced by duration manipulations are depicted by the blue arrows, with arrows pointing in the direction of the change. Green arrows depict the spectral quality effects on mean drift rate. Red arrows show the changes in non-decision time that are due to different vowel durations – the further to the right the arrow extends the longer the non-decision time. Gray arrows show the response threshold heights for each accumulator.

are of equal length, green arrows are of equal length, red arrows are of equal length, and gray arrows are of equal length.

The second model, the unequal drift model, allowed v to vary across responses in addition to spectral quality values, and duration values. Like the equal drift model, this model assumed that spectral quality and duration influence the speed of information processing, however this effect was not fixed to be equal across /ɑ/ and /aː/ responses. In Figure 3.2, we would see that blue arrows are not equal length and green arrows are not equal length, but red arrows and gray arrows are still equal length. This model is theoretically interesting because it tests how much each cue contributes to the recognition of /ɑ/ or /aː/ individually. In contrast, traditional cue weighting measures based on categorization data only allow researchers

to infer about the contribution of duration or spectral quality cues to the overall vowel contrast, an approach that is captured by the equal drift model.

The third model we tested, the non-decision time model, was the same as the unequal drift model with one extension. The non-decision time model also had 10 separate t_0 parameters, one for each of the duration values. In Figure 3.2, we would see that blue arrows are not equal length, green arrows are not equal length, and the red arrows are not equal length, but gray arrows are of equal length. If longer stimulus durations delay the processing of duration information, since the participant must wait for the vowel to offset, then we should observe that longer stimulus durations induce systematic increases in non-decision processing time. Having 10 separate t_0 parameters for each duration value allows for longer durations to induce longer non-decision times. If we find evidence in favor of this model then this suggests that lengthening the vowel may induce longer perceptual encoding times (Reinisch & Sjerps, 2013).[4] However, if we find evidence in favor of another model, without different t_0 parameters, then this suggests that longer durations do not reliably delay the processing of duration information.

The final model we tested, the response bias model, was the same as the unequal drift model, but it allowed response thresholds to change across accumulators. This allowed the model to account for potential response bias in the data, i.e., responding /ɑ/ more often than /aː/ overall. Response bias is typically captured in response thresholds, unless the locus of bias is related to the stimulus (e.g., perceptual decision criterion, see White & Poldrack, 2014). In Figure 3.2, we would see

[4]Longer non-decision times could also mean longer motor response times, but we cannot disentangle effects of perceptual encoding time and motor response time. Here, we assume that longer durations do not have systematic effects on motor response time.

that blue arrows are not equal length, green arrows are not equal length, and gray arrows are not equal length, but red arrows are equal length.

Drift rates for each /aː/ response were defined as

$$v_{\text{aː}SD} = v_{\text{aː}} + \beta_{\text{aː}S}X_S + \beta_{\text{aː}D}X_D \tag{3.1}$$

where $v_{\text{aː}}$ represents the base drift rate for the /aː/ response. X_S corresponds to the $_Sth$ value spectral quality and X_D corresponds to the $_Dth$ value duration. The symbols X_S and X_D denote the stimulus spectral quality and duration steps, respectively (i.e., not the raw values in Hz/ms, but ordinal scale values from 1-10 shown in column 1 of Table 3.1). $\beta_{\text{aː}S}$ and $\beta_{\text{aː}D}$ denote parameters that describe the effect of the spectral and duration changes on drift rate for the /aː/ response, respectively.

Drift rates for each /ɑ/ response were defined as

$$v_{\text{ɑ}SD} = v_{\text{ɑ}} + \beta_{\text{ɑ}S}(11 - X_S) + \beta_{\text{ɑ}D}(11 - X_D) \tag{3.2}$$

where $v_{\text{ɑ}}$ represents the base drift rate for the /ɑ/ response. $\beta_{\text{ɑ}S}$ and $\beta_{\text{ɑ}D}$ denote parameters that describe the effect of the spectral and duration changes on drift rate for the /ɑ/ response, respectively. Note that for the equal drift model, the $\beta_{\text{ɑ}S}$ and $\beta_{\text{ɑ}D}$ are the same as the $\beta_{\text{aː}S}$ and $\beta_{\text{aː}D}$ estimates. The terms $(11 - X_S)$ and $(11 - X_D)$ ensure that higher spectral quality and duration conditions, which

correspond to atypical values for ɑ, produce smaller drift rates compared to lower spectral quality and duration conditions.

By estimating drift rate via Equations 3.1 and 3.2 we obtain four drift rate coefficient estimates – $\beta_{\text{a:}S}$, $\beta_{\text{a:}D}$, $\beta_{\text{ɑ}S}$ and $\beta_{\text{ɑ}D}$ – which represent the effect that our experimental cue manipulations have on the drift rate parameter. Note that Equations 3.1 and 3.2 produce linear effects on drift rate, but linear increases in drift rate also allow for non-linear effects on behavioral data (see e.g., van Ravenzwaaij, Brown, & Wagenmakers, 2011, Fig. 2).

We compared four versions of our LBA model: the equal drift model, the unequal drift model, the non-decision time model, and the response bias model using WAIC. In addition, we used Bayesian predictive p-values to statistically test for differences between the posterior distributions of the drift coefficient parameters (Meng, 1994). We used a criterion of .05 to determine if two posterior distributions are overlapping. Specifically, we calculated the difference between the two spectral quality posteriors, the two duration posteriors, the spectral /a:/ and the duration /a:/ posteriors, and the spectral /ɑ/ and the duration /ɑ/ posteriors. We calculated the probability (p-value) that the resulting difference distributions were equal to or less than 0. A p-value "is a measure of discrepancy between the observed data and the posited assumptions, among which the hypothesis being tested is only a part" (Meng, 1994, p. 1144). Thus, similar to the traditional p-value, a low predictive p-value indicates a low probability of observing this or more extreme data if the null hypothesis were true.

3.3 Results

The R code for all analyses and experimental data are available online at this paper's associated Open Science Framework page `https://osf.io/hp9xt/`.

3.3.1 Behavioral Data Results

We measured the choice proportion of participants for the two typical vowels: the minimum-duration step 1 and minimum-spectral step 1 stimuli (expected response was /ɑ/) as well as the maximum-duration step 10 and maximum-spectral step 10 stimuli (expected response was /aː/). Overall, the expected response was made 97% of the time, which suggests there was high consensus for the extreme stimuli. The percentage of expected responses ranged from 70% to 100% across participants. No participants were excluded from the analysis. Overall, there was a response bias as participants responded /ɑ/ 55% of the time, which we explore further in the LBA analysis.

Figure 3.3 shows the effects of both the duration and spectral manipulations as a heat map. The top right corner shows stimuli with typical /aː/ cue information (maximum-duration 10 and maximum-spectral 10) and the bottom left corner shows stimuli with typical /ɑ/ cue information (minimum-duration step 1 and minimum-spectral step 1). This plot shows the rate of change from responding /ɑ/ to responding /aː/ as a function of both the duration and spectral quality manipulations. As we move along the y-axis we can observe the change across spectral

quality and as we move along the x-axis we can see the change across duration. The diagonal of the graphic visualizes the listeners' perceptual boundary.

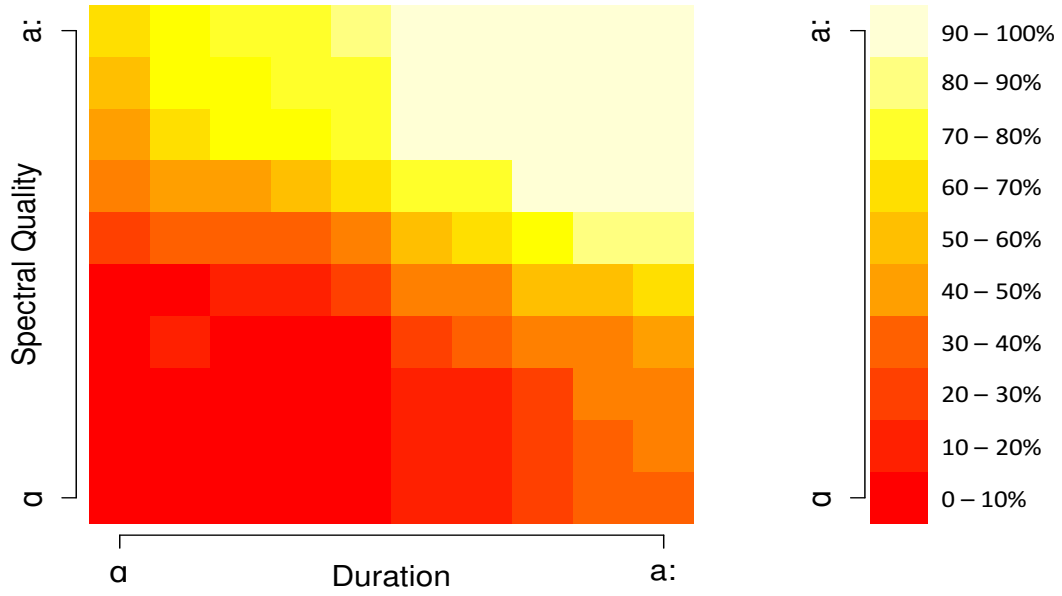


FIGURE 3.3: A heat map showing the overall effects of duration and spectral quality on categorization. The y-axis displays the spectral steps and the x-axis displays the duration steps, where each square represents a step. The text in the legend of this figure contains the percentage that participants responded /aː/ overall. The top right corner shows that participants predominantly respond /aː/ to stimuli with typical /aː/ values. The bottom corner shows that participants predominately response /ɑ/ to stimuli with typical /ɑ/ values.

Figure 3.4 shows the effects of spectral quality and duration on mean RT as a heat map. The top right corner shows stimuli that have typical /aː/ cue information and the bottom left corner shows stimuli that have typical /ɑ/ cue information.

We determined the effect of spectral cues and duration cues on categorization by subjecting the proportion of /aː/ responses to a Bayesian logistic regression. One coefficient of the regression corresponds to the spectral changes (β_{Spectral}) and another to the duration changes (β_{Duration}), and these represent each cue's influence on categorization. We regressed choice proportion on duration, spectral quality, and the interaction between the two. The model with main effects for both duration and

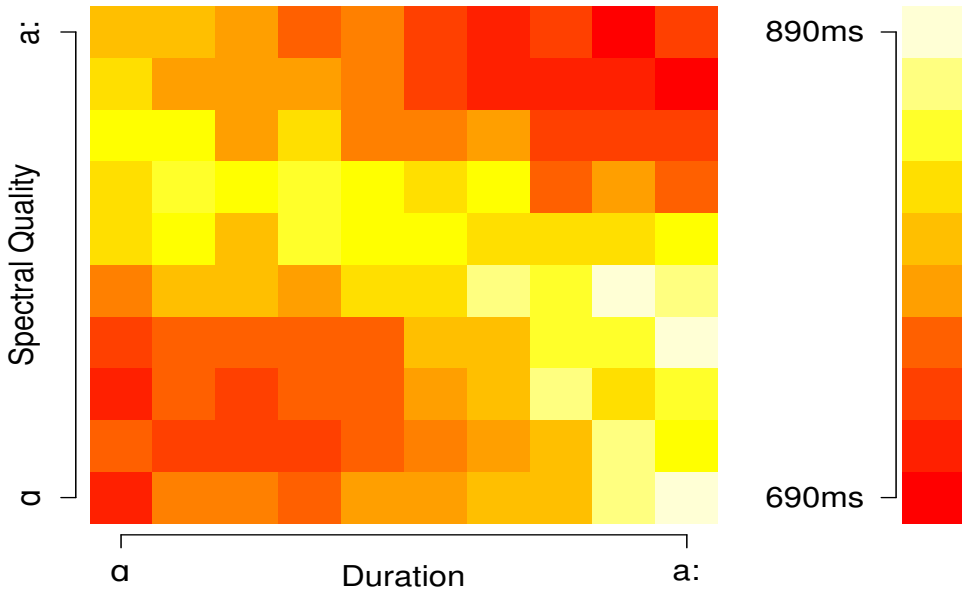


FIGURE 3.4: A heat map showing the effects of cue manipulation on RT, which has been collapsed over response. The y-axis displays the different spectral steps and the x-axis displays the different duration steps. The top right corner shows fast RTs for stimuli with typical /aː/ values. The bottom left corner shows fast RTs for stimuli with typical /ɑ/ values.

spectral quality and an interaction between the two had the lowest WAIC (12965.2). The null model with no effects, the duration only, spectral quality only, and the model with both main effects only had WAIC values of 20548.4, 19210.4, 15059.9, and 12970.5, respectively.

The spectral manipulation had a larger effect on categorization than the duration manipulation. Each step increase in duration increased the log odds of responding /aː/ by $\beta_{\text{Duration}} = .403$ [95% Credible Interval: 0.389, 0.449]. The percentage of /aː/ responses for the 1^{st} duration step was 22.9% and this increased to 66.4% for the 10^{th} duration step (t-test: $BF_{10} = 1.24\times10^{7}$). Each step increase in spectral quality increased the log odds of responding /aː/ by $\beta_{\text{Spectral}} = .664$ [95% Credible Interval: 0.650, 0.709]. The percentage of /aː/ responses for the 1^{st} spectral quality step was 11.7% and this increased to 84% for the 10^{th} spectral quality step

(*t*-test: $BF_{10} = 2.66\times10^{11}$). The interaction between spectral quality and duration showed that for the 1^{st} duration step, the percentage of /aː/ responses increased from 20% to 58% from the 1^{st} to the 10^{th} spectral quality step (*t*-test: $BF_{10} = 1.20\times10^{7}$). For the 10^{th} duration step the percentage of /aː/ responses increased from 28% to 96% from the 1^{st} to the 10^{th} spectral quality step (*t*-test: $BF_{10} = 2.95\times10^{7}$).

We regressed RTs on duration, spectral quality, response, and all interactions and found main effects for response ($BF_{\text{Inclusion}} > 10^{15}$), duration ($BF_{\text{Inclusion}} = 2.42\times10^{11}$), and spectral quality ($BF_{\text{Inclusion}} > 10^{15}$). Moreover, the model included an interaction between spectral quality and response ($BF_{\text{Inclusion}} > 10^{15}$), duration and response ($BF_{\text{Inclusion}} = 303$), and spectral quality and duration ($BF_{\text{Inclusion}} = 112$). RTs for the 1^{st} spectral quality step were faster than RTs for ambiguous spectral quality steps, such as 5 (*t*-test: $BF_{10} = 2.16$) or 6 (*t*-test: $BF_{10} = 10.99$), but the difference in RTs between the 1^{st} and 5^{th} spectral quality step is inconclusive. RTs for the 10^{th} spectral quality step were faster than RTs for the 5^{th} (*t*-test: $BF_{10} = 42.98$) or the 6^{th} spectral quality step (*t*-test: $BF_{10} = 165.95$). RTs for the 1^{st} duration step were equal to RTs for ambiguous duration steps, such as 5 (*t*-test: $BF_{10} = .203$) or 6 (*t*-test: $BF_{10} = .226$). RTs for the 10^{th} duration were slower than RTs for the 5^{th} (*t*-test: $BF_{10} = 1.88$) and 6^{th} (*t*-test: $BF_{10} = .80$) duration step, but the evidence for these differences was inconclusive.

As shown in the left panel of Figure 3.5, there was a crossover interaction between spectral quality and response. The evidence for differences in RT between the /ɑ/ and /aː/ responses for the 1^{st} spectral quality was inconclusive (*t*-test: $BF_{10} = 1.05$). But, RTs for /aː/ responses were faster than /ɑ/ responses for the 10^{th} spectral quality (*t*-test: $BF_{10} = 64.47$). As shown in the right panel of Figure 3.5,

RTs for /ɑ/ responses were faster than RTs for /aː/ responses for the 1^{st} duration step, but the evidence was inconclusive (t-test: $BF_{10} = 2.40$). In contrast, RTs for /ɑ/ responses were slower than RTs for /aː/ responses for the 10^{th} duration step (t-test: $BF_{10} = 14.63$). RTs for /aː/ responses were slower for the 5^{th} duration step (t-test: $BF_{10} = 0.949$) and faster for the 6^{th} duration step (t-test: $BF_{10} = 0.899$), but the evidence was inconclusive for both. There were no differences in RT between the /aː/ and /ɑ/ responses for the 5^{th} (t-test: $BF_{10} = 0.19$) and 6^{th} (t-test: $BF_{10} = 0.29$) duration steps. The absolute difference in RTs for the 1^{st} and 10^{th} spectral qualities was larger for stimuli with longer duration values (114ms) than shorter duration values (108 ms), but the evidence was ambiguous (t-test: $BF_{10} = 1.38$).

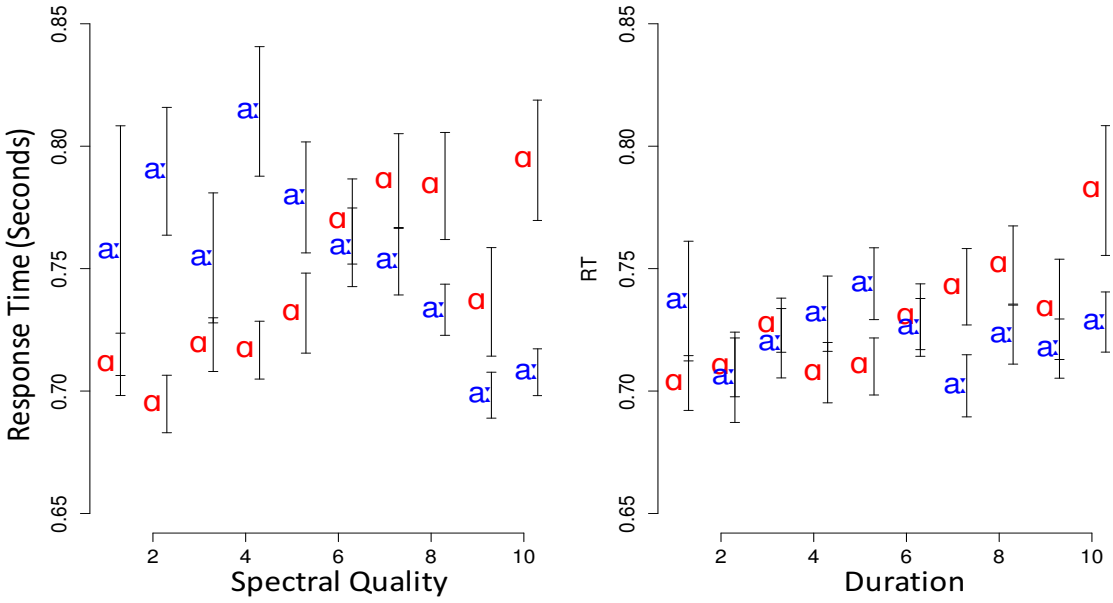


FIGURE 3.5: The interaction between the spectral quality and response (left) and duration and response (right). The /ɑ/ responses are plotted in red and the /aː/ responses are plotted in blue. Interval bars represent 1 standard error of the mean.

3.3.2 Linear Ballistic Accumulator Results

We found that drift rates (speed of processing) were affected more by spectral quality than by duration. We also found that the effects of spectral quality on speed of processing were not equal for both for phoneme responses. Given the findings of McMurray et al. (2008) and Reinisch and Sjerps (2013), who found that cues that appeared later in the signal affected eye-tracking behavior later in a trial, we expected longer non-decision times for stimuli with longer vowel durations – in particular, because participants were assumed to not be processing duration information until the vowel offset. However, our analysis showed that increased duration of vowels did not produce any systematic increases in non-decision processing time.

Specifically, to compare the equal drift, unequal drift, and non-decision time models we assessed which model was selected by WAIC. The unequal drift model (WAIC = 629.2) was preferred over the equal drift model (WAIC = 918.7), the non-decision time model (WAIC = 1083.2), and the response bias model (WAIC = 946.9). The unequal drift model also provided good fits to the empirical data, which are presented in the Appendix B.

Figure 3.6 shows the group level mean posteriors of spectral and duration drift coefficients from the unequal drift model. Bayesian p-values suggest that changes in spectral cues induced larger changes in drift rate than duration. The effects of spectral quality on drift rates are not identical for both vowels, where changes in spectral quality affect drift rates more for /aː/ than /ɑ/ responses. The effects of duration on drift rate are equal for both vowels.

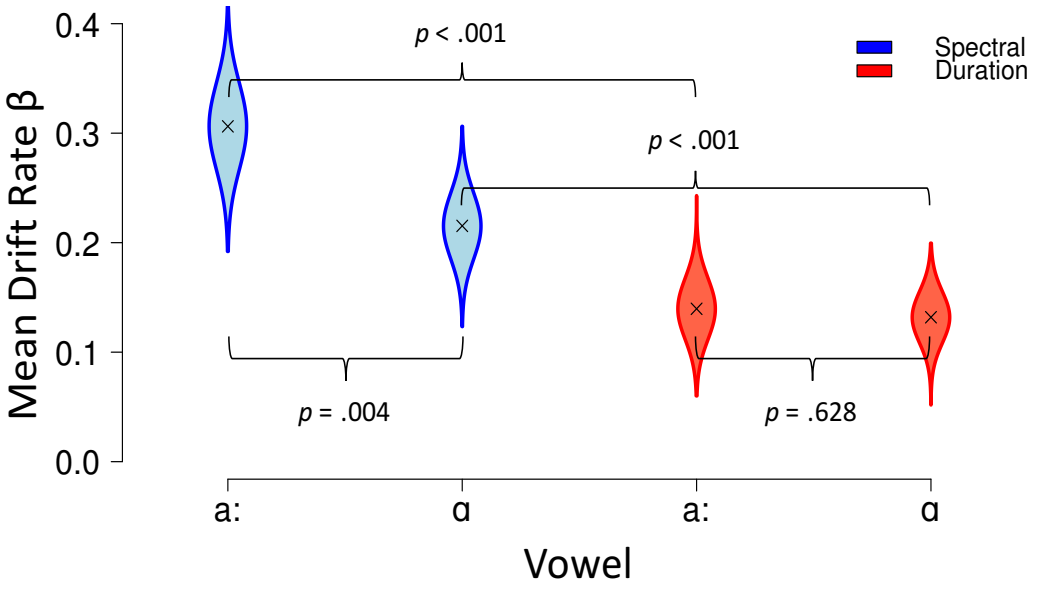


FIGURE 3.6: Group-level mean posterior distributions for mean drift rate β for both the /ɑ/ and /aː/ responses.

The group-level mean posteriors for all other LBA parameters are shown in Table 3.2. The base drift rate for /ɑ/ is higher than for /aː/. This explains the overall bias we observed, where participants responded /ɑ/ 55% of the time. We tested whether the bias is in response thresholds, instead of drift rate, by fitting the response bias model with separate thresholds for each vowel. The response bias model performed worse compared to the unequal drift model. Therefore the bias is better captured in drift rate.

The combined effect of base drift rate and spectral quality drift coefficients can be better understood by looking at the mean drift rates in each of the 10 spectral quality steps. To calculate these mean drift rates we inserted the group-level mean posterior median (see Table 3.2) for the base drift rate and drift coefficient parameters into Equations 3.1 and 3.2. This resulted in the mean drift rates in each of the 200 conditions. To arrive at the mean drift rates in each of the 10 spectral quality steps we averaged over the duration conditions. Zooming in on the different

Table 3.2: Group-level Mean Posteriors.

Parameter	Median	Credible Interval (2.5%, 97.5%)
b	2.125	(2.417, 2.720)
A	0.953	(0.841, 1.062)
s	0.911	(0.866, 0.958)
t_0	153ms	(61ms, 192ms)
$v_{\text{ɑ}}$	2.70	(2.41, 2.98)
$v_{\text{aː}}$	2.37	(2.10, 2.65)
$\beta_{\text{aː}S}$	0.306	(.252, .360)
$\beta_{\text{aː}D}$	.140	(.103, .176)
$\beta_{\text{ɑ}S}$	.215	(.174, .256)
$\beta_{\text{ɑ}D}$	.132	(.100, 1.64)

effects of drift rate across spectral quality for /ɑ/ and /aː/ (cf. blue posterior distributions in Figure 3.6), Figure 3.7 shows that for stimuli with atypical spectral quality values, which is the 1^{st} step for /aː/ and the 10^{th} step for /ɑ/, participants had higher drift rates for /ɑ/ compared to /aː/. The difference between mean drift rates between the two vowels decreases as the stimuli approach the typical spectral quality.

3.4 Discussion

In this study, Dutch listeners categorized sounds as /ɑ/ and /aː/ in an experiment in which we manipulated the spectral and duration properties of the sounds. Both the categorization and the response times (RTs) of participants were recorded and independently analyzed using Bayesian linear models. We then applied

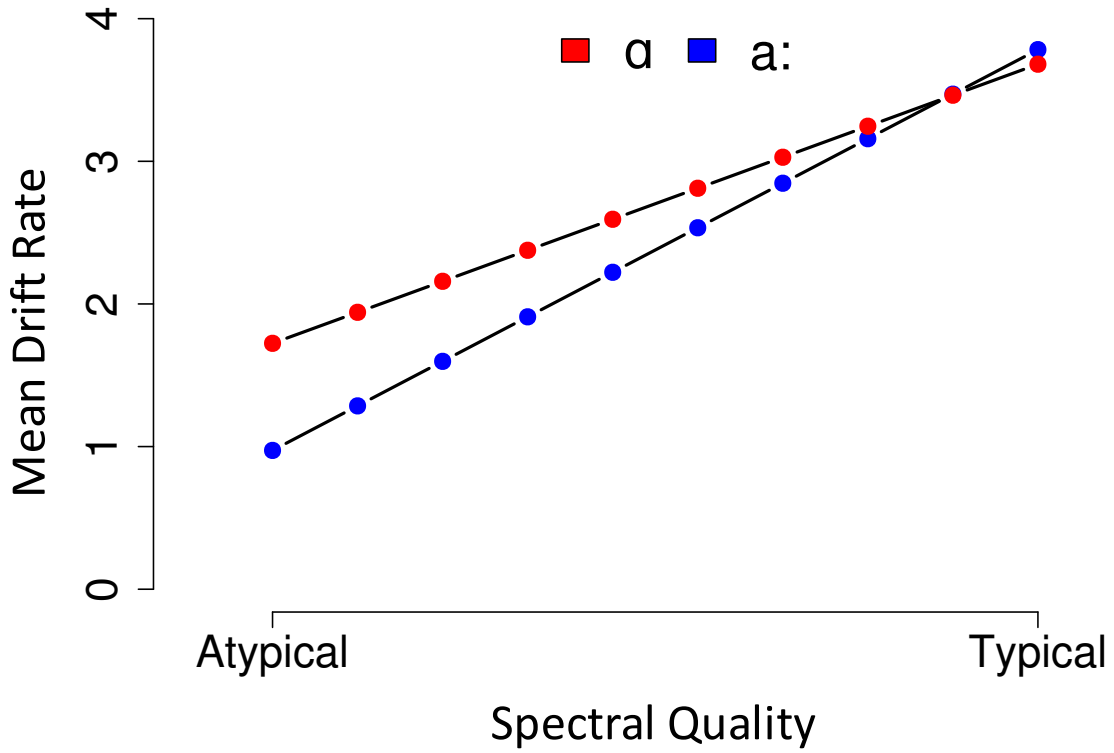


FIGURE 3.7: Mean drift rates across the 10 spectral qualities for /ɑ/ (red) and for /aː/ (blue). Mean drift rates are shown along the y-axis. Along the x-axis are the 10 different spectral quality values ranging from the atypical values, which is the 1^{st} value for /aː/ and the 10^{th} value for /ɑ/, to the typical values, which is the 10^{th} value for /aː/ and the 1^{st} value for /ɑ/.

a mathematical model of decision-making – the linear ballistic accumulator model (LBA; S. D. Brown & Heathcote, 2008) – which analyzed both streams of data simultaneously. The model allowed us to investigate how changes in acoustic cues affect latent cognitive processes that underpin phoneme decisions.

In terms of behavioral data, we were able to replicate the relatively heavy spectral cue weighting compared to duration that is typically observed in vowel categorization tasks involving the Dutch contrast between /ɑ/ and /aː/ (Nooteboom & Cohen, 1984; van Heuven et al., 1986; Escudero et al., 2009; van der Feest & Swingley, 2011; Reinisch & Sjerps, 2013). The LBA analysis addressed fundamental issues that arise when using categorization data to learn about the mapping

between acoustic cues and phonemes. The model posits an evidence accumulation process, which helps explain how continuous acoustic information can result in discrete phoneme decisions. With the LBA we observed that changes in spectral and duration cues lead to changes in drift rates (i.e., speed of information processing) for both /ɑ/ and /aː/ responses. In addition, changes in spectral quality had larger effects on the behavioral data than changes in duration and this was driven by differences in speed of information processing. Furthermore, when the stimuli had more atypical spectral qualities, which is a short duration and darker spectral quality for /aː/ or a long duration and clearer spectral quality for /ɑ/, listeners accumulate evidence faster for /ɑ/ responses compared to /aː/. This asymmetry in processing speed explains why listeners' responded /ɑ/ 55% of the time.

We also argued that if duration processing can only start at vowel offset, as suggested by Reinisch and Sjerps (2013), we would observe relatively longer perceptual encoding times for longer vowel durations. Non-decision time in the LBA is made up of the time needed to perceptually encode the stimuli and execute a motor response. If we assume that longer sound durations have no systematic effects on motor response times, then changes in non-decision time would be due to differences in perceptual encoding time. However, we found that different vowel durations did not affect listener's non-decision processing time.

3.4.1 What Precedes an Evidence Accumulation Process?

With the LBA we provide part of the picture of how continuous acoustic information leads to discrete phoneme decisions. In the past, researchers used categorization data to look at associations between cues before-processing and overt responses, which overlooks the decision processes required for categorization. With the LBA we specifically investigate the decision processes, showing that the relatively heavy weighting of spectral quality compared to duration is not merely a result of the timing of cue availability, but a property of the decision-making process.

While the LBA analysis presented here is a considerable advance on traditional separate analyses of response proportion and RT, it does not explicitly explain how continuous acoustic information becomes evidence that drives the decision process. Recall that the drift rate of the decision process is considered a measure of the quality of the evidence. But what is the evidence to this accumulation process? One possibility is to complement the LBA with a formal model of a cue mapping process that takes the continuous acoustic information and maps it onto exemplar clusters or discrete representations of phonetic categories. The output of this cue mapping model could serve as evidence that inputs into the LBA processes (see Ratcliff, Gomez, & McKoon, 2004, for a similar concept in lexical decision-making). The drift rate in the LBA would then be a measure of the speed, efficiency, or certainty with which a certain acoustic cue is mapped onto a phoneme category. In summary, the continuous acoustic information enters the cognitive system, the acoustic information then maps onto a phonetic category, the mapping processes outputs evidence that enters an LBA process, which accumulates evidence until a

response threshold is reached and an overt response is made.

But what could cause differences in speed of information processing? In our case, we found behavioral evidence that spectral quality is weighed heavier than duration for the /aː/ and /ɑ/ contrast. The LBA analysis showed that changes in spectral quality cause larger changes in speed of processing than changes in duration. Perhaps the differences in drift rate were due to spectral quality cues being mapped more efficiently or more certainly onto the categories /aː/ and /ɑ/ compared to duration cues. Similarly, the higher spectral quality drift rate for /aː/ than for /ɑ/ suggests that the mapping between spectral quality information and /aː/ provides better evidence than the mapping between spectral quality and /ɑ/. Finally, the higher drift rate for atypical values of /ɑ/ than for atypical values of /aː/ suggests that the mapping between atypical acoustic cues and categories is more efficient or more certain for /ɑ/ than it is for /aː/. Asymmetric mappings between cue values and categories could be instrumental in explaining asymmetries in vowel perception (Polka & Werker, 1994), as well as provide a phonetic basis for the notion of phonological under-specification (Lahiri & Reetz, 2010).

3.4.2 Future Directions

In this study, we investigated duration and spectral quality, which are cues that are processed separately and by different pathways in the brain (e.g., Zatorre & Belin, 2001). The LBA model is not constrained by this independence of acoustic cues. For instance, researchers could examine phoneme decisions about sounds that

are cued by two spectral dimensions (F1 and F2) rather than one spectral and one temporal dimension.

Moreover, the LBA can be used to model how listeners deal with correlated acoustic dimensions, such as F1 and inherent vowel duration (House & Fairbanks, 1953; Peterson & Lehiste, 1960; Lehiste & Lass, 1976) or F2 and spectral tilt for /i/ and /u/ (Ito, Tsuchida, & Yano, 2001). These correlations can be modeled by the LBA by changing the way parameters vary across conditions. For instance, if spectral quality and duration were positively correlated, Equations 3.1 and 3.2, which estimate drift rates for each condition, could be extended by including an interaction term. The mean drift rates for each /aː/ response could be defined as

$$v_{\text{aː}SD} = v_{\text{aː}} + \beta_{\text{aː}S}X_S + \beta_{\text{aː}D}X_D + \beta_{\text{aː}SD}X_SX_D \tag{3.3}$$

where $\beta_{\text{aː}SD}$ denotes the parameter that describes the interaction effect of the spectral and duration changes on mean drift rate for the /aː/ response. A formal model comparison (i.e., comparing a model fit with and without the interaction terms in their ability to fit the data) can shed light on the way drift rates change with both dimensions. Note that this approach allows one to investigate the effect of both dimensions independently, even though they covary in practice. The LBA analysis with the added interaction terms can be incorporated for phonetic distinctions cued by one dimension, two dimensions, three dimensions, or more by augmenting Equation 3.3 with the relevant terms.

The LBA could also be used to address other long standing questions in

speech perception. For example, the literature so far does not contain a definitive explanation of an asymmetry first observed by Nooteboom and Doodeman (1980, p. 277): reducing the duration of /aː/ can lead to listeners responding 100% /ɑ/; yet, increasing the duration of /ɑ/ does not lead to a consistent /aː/ response. Van der Feest and Swingley (2011) proposed two possible explanations of this phenomenon. Firstly, lengthening sounds might show weaker effects because it occurs in natural language as prosodic effects, such as the application of emphatic stress (Ko, Soderstrom, & Morgan, 2009). Secondly, lengthening sounds may facilitate perceptual access to vowel quality, whereas the spectral quality in shortened vowels may be harder to evaluate, leading to reliance on duration for short vowels. The LBA could model the second explanation by letting the drift rate for spectral quality change within a trial as the stimulus duration increases. Fortunately, a non-constant drift rate over the course of the trial is something that is possible to explore in the LBA (Holmes, Trueblood, & Heathcote, 2016).

Allowing non-constant drift rate within a trial opens the door to investigating how time-variant acoustic cues cause changes in LBA parameters that also change over time. For example, the cues to pre-voiced /b/ become available in a sequence: first pre-voicing, which provides information about phonological voicing, then the burst, which provides information about the place of articulation and voicing, and then the formant trajectories into the vowel, which listeners rely on most to determine place of articulation. Another example of a time-variant acoustic cue is vowel-inherent spectral change (Morrison, 2013). These dynamic cues can be mapped onto latent cognitive processes by linking them to dynamic LBA parameters. Dynamic LBA parameters may be able to disentangle the effects of cues that

are available earlier (in time) in the speech from cues that are simply mapped more strongly to phonetic categories.

Overall, our study demonstrates a successful and novel application of sequential sampling models to phoneme categorization tasks. These models allow researchers to investigate latent cognitive processes by analyzing behavioral data. Given the merit of using sequential sampling models, we hope to see them applied to other research questions embedded in the speech perception literature, some of which we have outlined here.

Chapter 4

Extending the Diffusion Model of Lexical Decision

Reference

Tillman, G., Osth, A., van Ravenzwaaij, D., & Heathcote, A. (Under Review). A Diffusion Decision Model Analysis of Evidence Variability in the Lexical Decision Task. Psychonomic Bulletin & Review

4.1 Introduction

The lexical-decision task, which involves identifying letter strings as words or non-words, has been used extensively in psycholinguistic research to understand reading, and to develop cognitive models (e.g., Grainger & Jacobs, 1996; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Norris, 2006). Discrimination of words from non-words in the lexical-decision task has typically been understood using signal-detection theory (SDT, e.g., Norris, 1986; Balota & Chumbley, 1984). In SDT, both words and non-words are assumed to have continuously distributed evidence of word-likeness, with observers using a criterion on the evidence axis as a basis for their decision.

Simple elaborations of SDT (e.g., Balota & Spieler, 1999) do not correctly predict the shapes of response time (RT) distributions (Yap, Balota, Cortese, & Watson, 2006) in an "information controlled" lexical-decision task, where participants control the time at which they make their choice. For this reason, researchers have applied the Diffusion Decision Model (DDM; Ratcliff, 1978) to the lexical-decision task (Ratcliff, Gomez, & McKoon, 2004). The DDM is an sequential sampling model that can account for both choice proportion and RTs. Decisions between two alternatives are based on the accumulation of noisy evidence from a stimulus. Evidence accumulates until one of two decision boundaries is reached and the associated response is made. The mean rate, or drift-rate, of evidence accumulation varies from trial to trial. This trial-to-trial variability in drift-rate is analogous to the continuously distributed variability of evidence in SDT (Ratcliff, 1978, 1985).

Previous fits of the DDM to lexical-decision data have assumed equal drift-rate variability for both words and non-words (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers et al., 2008). However, studies incidentally using lexical decisions to look at practice effects (Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Dutilh, Krypotos, & Wagenmakers, 2011) and post-error slowing (Dutilh et al., 2012) have reported larger drift-rate variability for words than non-words. Moreover, the Retrieving Effectively from Memory model of lexical-decision (REM–LD; Wagenmakers et al., 2004) makes the a-priori prediction that evidence-strength variability differs across stimulus classes. Note, however, that REM-LD has only been applied to "time-controlled" lexical-decision tasks, where participants must respond at a deadline specified by the experimenter.

Taken together, these findings warrant an investigation of whether evidence variability differs among words and non-words of different types in the lexical-decision task. To do so, we fit the DDM to previously published information-controlled lexical-decision data sets from Ratcliff, Gomez, and McKoon (2004) and Wagenmakers et al. (2008). We also compare the DDM based evidence mean and variability estimates to predictions that we derive from REM–LD.

4.1.1 The Diffusion Decision Model

The DDM provides a comprehensive account of data from the lexical-decision task (Ratcliff, Gomez, & McKoon, 2004), not only for the choices made, but also for the associated RTs. In the DDM, evidence begins to accumulate at the starting

point z, which is sampled from a uniform distribution with width s_z. Evidence accumulation is noisy, with a drift-rate that is sampled from a normal distribution with mean v and standard deviation η. Evidence continues to accumulate until it hits either an upper boundary (a, corresponding to a 'word' response) or a lower boundary (0, corresponding to a 'nonword' response). The boundary that is reached first determines the decision, and the time taken to reach the boundary is the decision time. Non-decision time, T_{er}, which quantifies the time taken to encode stimuli and execute a motor response, is estimated as the remainder of each RT. Non-decision time is assumed to have a uniform distribution with range s_t. Figure 4.1 illustrates the DDM.

Ratcliff (1978) introduced drift-rate variability to model item differences in a recognition-memory task, but it is also useful because it enables the DDM to predict the common finding of slower error than correct responses. When the start point is unbiased, a somewhat counter-intuitive prediction for a diffusion process with no trial-to-trial variability is that correct and error decision times have equal distributions. When drift-rate variability is included, lower drift-rates, which cause slow decisions, are more likely to produce errors, while higher drift-rates, which cause fast decisions, are more likely to produce correct responses. Slow errors result because, on average, correct responses have faster drift-rates than incorrect responses (Ratcliff & McKoon, 2008).

Using the DDM, Ratcliff, Gomez, and McKoon (2004) modeled the effects of different stimulus classes in the lexical-decision task using differences in mean drift-rate alone. Higher frequency words had larger drift-rates, which accounted for

their greater accuracy and faster RT. However, all item types were assumed to have the same drift-rate variability.

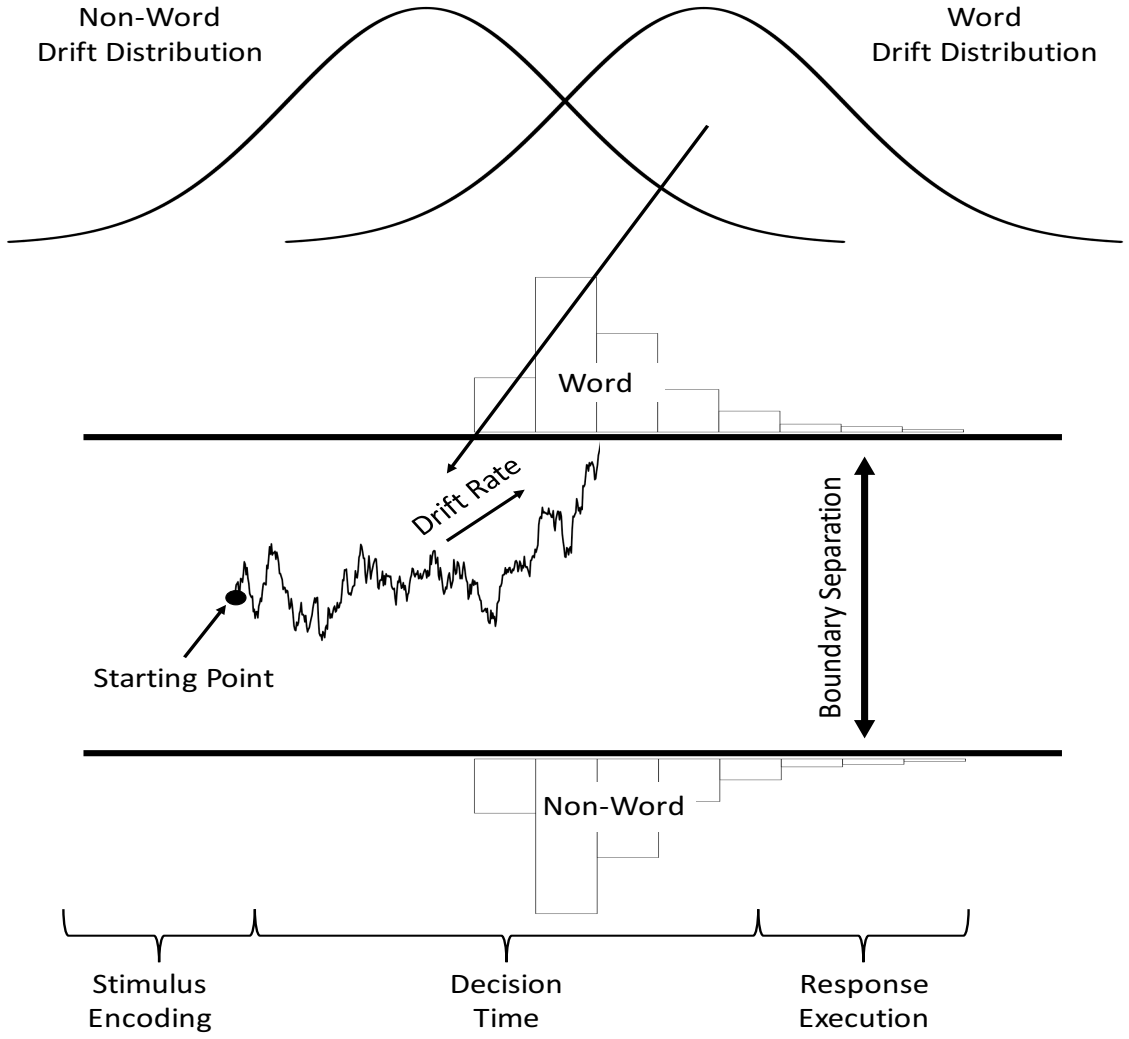


FIGURE 4.1: DDM conceptualization of a two choice decision between 'word' and 'non-word' in the lexical-decision task. The top panel shows distributions of drift-rates for both words and non-words. The drift-rate on each trial is sampled from a drift distribution.

4.1.2 From Signal Detection to Diffusion Decision Models of Evidence Variability

SDT is often fit to receiver operating characteristic (ROC) data. ROCs plot the hit rates as a function of false alarm rate across different levels of bias. Bias can

be manipulated in a number of ways. The most frequently used method is confidence, assuming that different confidence ratings are produced by placing multiple criteria on the evidence-strength axis. SDT predicts a curvilinear ROC, but z-transforming hit rates and false alarm rates produce a linear ("zROC") plot with a slope equal to the ratio of the standard deviation (SD) of the noise distribution and the signal-plus-noise distribution. Therefore, a zROC slope different from 1 suggests unequal variance evidence distributions.

Jacobs, Graf, and Kinder (2003) extended the multiple read-out model (MROM) of lexical-decision to allow it to generate ROC functions. They found that in a time-controlled lexical-decision task with a 500ms response deadline, MROM predicted zROCs slopes of 1, which suggests equal evidence variability for words and non-words (see also S. D. Brown & Steyvers, 2005).

However, the notion that zROC slopes equal the ratio of SDs in evidence strength has been contested by Ratcliff and Starns (2009). Ratcliff and Starns used a race model (RTCON) to explain both confidence judgments and RT. Similar to the DDM, decisions in RTCON are made on the basis of accumulation of noisy evidence to a response boundary. RTCON assumes several additional sources of variability beyond that of SDT, including variability in response boundaries for each potential confidence response. Because the zROC functions generated by RTCON are dependent on multiple sources of variability, Ratcliff and Starns suggested that zROC slopes cannot be used as a valid measure of the ratio of SDs in evidence strength because it does not take account of RT.

In a two-choice recognition memory task, Starns and Ratcliff (2014) showed

that the DDM could be used to estimate the ratio of evidence SDs in a way that takes account of RTs, but without requiring ROC measurement (see also Starns, 2014). We aim to use the DDM in a similar manner to investigate evidence variability in the lexical-decision task, and to compare these results to the predictions of REM–LD, addressing not only word vs. non-word variability, but also variability differences among different types of words and non-words. Before we present our analysis, we will briefly overview the REM–LD model, and investigate its predictions about evidence variability.

4.1.3 Retrieving Effectively from Memory – Lexical Decision

In REM–LD, lexical representations for words are vectors of features. The elements of these vectors can encode semantic, phonemic, and orthographic information about the words that are experienced.[1] During a lexical-decision task the features of the probe are matched in parallel to features of lexical traces. The number of features that are available for this matching process increases as the probe is processed for longer. Not having all features available for the matching process results in there being mismatching features as well as matching features, even when the probe is the same as the trace.

The number of active features, and the number of matches between the probe and traces, are binomially distributed. REM–LD computes the probability that r probe features are active given some amount of processing time, and from

[1]The REM–LD model also considers contextual information, however, Wagenmakers et al. (2004) advocates that contextual information is less relevant in lexical-decision.

that, calculates the likelihood of obtaining the observed number of matches between a memory trace and the probe when the probe is the same as the trace (i.e., the stimulus is a word). It then computes the likelihood of the observed number of matches given that the trace is not the same as the probe (i.e., the stimulus is a non-word). The two types of likelihoods are used to calculate odds ratios that are averaged over all memory traces, yielding a posterior odds ratio that is used to predict the probability of responding 'word' vs. 'non-word' in a deadline paradigm.

REM–LD and the DDM both posit that evidence in the decision process accumulates over time. Indeed, Wagenmakers et al. (2004) note that REM—LD is closely related to the random walk model (e.g., Link & Heath, 1975), which is a diffusion process in discrete time. Wagenmakers' et al. did not extend the REM–LD model to an information-controlled paradigm. There are several ways this might be achieved. For example, if evidence in the DDM is equated with the logarithm of the REM–LD posterior odds calculated over some fixed encoding time (which has an asymptotic normal distribution), then REM–LD might serve as a "front-end", providing an input to the DDM (for a similar idea see Ratcliff, Gomez, & McKoon, 2004).

We used the equations from Wagenmakers et al. (2004) to derive deadline predictions of REM–LD about the log posterior odds ratio distributions of the stimulus classes used in Ratcliff, Gomez, and McKoon (2004) and Wagenmakers et al. (2008) (for details see Appendix C). These categories include words of high-frequency (HF), low-frequency (LF), and very-low-frequency (VLF), along with two types of non-words, pseudo-words (PW), and random letter (RL) strings.

The mean (top left panel) and SD (top right panel) of the log posterior odds ratio distributions are plotted in Figure 4.2. For all deadlines greater than 250ms, which is the starting point for the decision process, the SD is largest for HF, followed by LF, RL, VLF, and finally PW stimuli. The variability of log posterior odds ratio distributions is determined by variability in its constituent odds ratios. For non-words the odds-ratio distribution comes from traces that are not the same as the probe, with an SD that decreases as the probability of a match increases. As the probability of a match is less for RL than the word-like PW stimuli, the latter have a smaller SD. For words the SD is strongly influenced by the lexical trace corresponding to the probe word, which causes the odds ratio distribution to be skewed. Higher frequency words skew the distribution more, producing a more variable odds-ratio distribution. Therefore, although REM–LD generally predicts greater variability with a larger evidence mean, there is a dissociation for RL stimuli, which have the lowest mean but the second highest SD.

Although we outlined one fairly direct way of linking REM–LD to the DDM, there are clearly alternatives, such as Wagenmakers et al.'s (2004) suggestion that moment-to-moment variability in REM–LD directly causes moment-to-moment variability in within-trial evidence accumulation. This also implies that, in contrast to the DDM's assumptions, evidence is non-stationary, growing in both mean and variance over the course of accumulation. Exploring this alternative in detail, let alone quantitatively fitting them to data, would be a major undertaking. Instead, we tested the predictions made by REM–LD about the mean and SD of the evidence distribution in a qualitative rather than quantitative way, by comparing its predicted ordering to the ordering of η estimates from the DDM model fits.

4.1.4 The Present Study

We examine whether evidence variability is different across stimulus classes via model selection. Specifically, we test whether a model with separate rate-variability (η) parameters for each stimulus class accounts for data better than a model with only one η for all stimulus classes. We then extract the estimates of rate means (v) and η from the former model, and compare them to REM–LD's predicted ordering.

4.2 DDM Analysis

We used hierarchical Bayesian methods to estimate the parameters of the DDM; the fitting routine, the specific model parameterization for each data set, and the results of a parameter-recovery study validating our estimates are provided in Appendix C.

4.2.1 Data Sets

Table 4.1 provides data-set details. All data sets contained a word frequency manipulation. In data set 1 (Experiment 1; Wagenmakers et al., 2008), participants were instructed to respond either quickly or accurately. For data set 2 (Experiment 2; Wagenmakers et al., 2008), one condition contained 25% word stimuli and another condition contained 75% word stimuli. Data sets 3-5 (Experiments 1,2, and 4, respectively, from Ratcliff, Gomez, & McKoon, 2004) all contained just a word

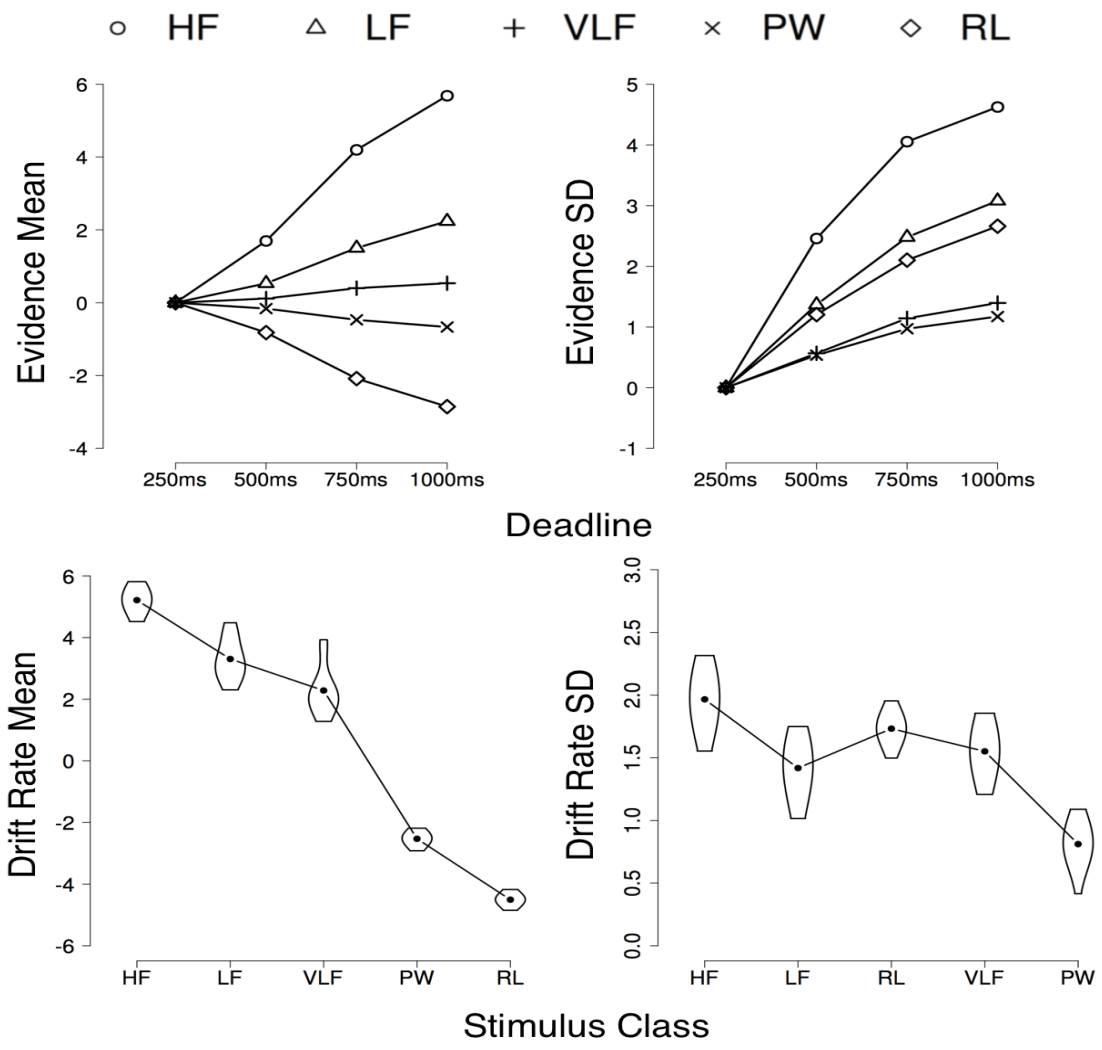


FIGURE 4.2: The top row plots the mean (top left panel) and SD (top right panel) of the log-posterior-odds-ratio distributions from the REM–LD model at four different deadlines (250ms, 500ms, 750ms, 1000ms). HF = high-frequency, LF = low-frequency, VLF = very-low-frequency, and PW = pseudo-word, and RL = random letter strings. The minimum processing time was 250ms, and the rate of increase in probability of activation was .0025. The probability of a feature match when encoding the same item was HF = .85, LF = .75, and VLF = .65. The probability of a feature match when encoding a different item was PW = .5, and RL = .35. The bottom row plots the drift-rate mean (v, bottom left panel) and SD (η, bottom right panel) group-level mean posterior distributions from DDM fits to the five lexical-decision experiments. The distributions for each stimulus class are the concatenation of the posterior distributions across all 5 experiments. The posterior distributions are displayed as violin plots, which show the median of the posterior (black dot) and a rotated kernel density mirrored on either side. The violin plots are truncated to contain the 95% highest density interval. The stimulus class labels along the x-axis are ordered from left-right in the same order as REM–LD's predicted ordering from highest to lowest.

frequency manipulation, but the researchers changed the characteristics of the non-word stimuli, using either pseudo-words (pronounceable letter strings in data-set 3, and created by randomly replacing all vowels in words by other vowels, in data sets 1 and 2), or unpronounceable random-letter strings (data-sets 4 and 5).

TABLE 4.1: Data sets.

Data Set	Source	N	Obs.	Variables
1	Wagenmakers et al. (2008) Exp1.	17	1844	Emphasis (Speed or Accuracy)
				Word Frequency (high, low, very low, pseudo-words)
2	Wagenmakers et al. (2008) Exp.2	19	1915	Proportion (25% Word or 75% Word)
				Word Frequency (high, low, very low, pseudo-words)
3	Ratcliff et al. (2004) Exp.1	16	2057	Word Frequency (high, low, very low, pseudo-words)
4	Ratcliff et al. (2004) Exp.2	14	2070	Word Frequency (high, low, very low, random letter strings)
5	Ratcliff et al. (2004) Exp.4	17	1477	Word Frequency (high, low, random letter strings)

Note. N = number of participants; Obs. = the mean observations for each participant.

4.2.2 Model Selection

We selected among models using WAIC, a measure of out-of-sample prediction error (Watanabe, 2010; Gelman et al., 2014), where lower values indicate better out-of-sample prediction. WAIC is similar to AIC (Akaike, 1974), but also takes account of functional-form complexity, and is more stable than the most common model-selection metric for hierarchical Bayesian models, DIC (Spiegelhalter et al., 2002).

We compared two versions of the DDM: the "equal model" and the "unequal model". The equal model had one η parameter for all different stimulus conditions, whereas the unequal model had a separate η parameter for each. Table 4.2 shows that the unequal model is substantially preferred for data sets 1-3 and clearly preferred for 4-5. Note that a difference in WAIC of greater than 3 provides positive evidence and

a difference of 10 or more strong evidence. We now examine whether the preferred models provide a good account of the data, in which case their parameters provide an accurate understanding of performance in the LDT.

TABLE 4.2: WAIC results for the equal and word frequency DDMs.

Data Set	Source	Equal Model	Unequal Model	Equal - Unequal
1	Wagenmakers et al. (2008) Exp1.	-16154.9	**-16320.7**	165.8
2	Wagenmakers et al. (2008) Exp.2	-21897.0	**-22022.4**	125.4
3	Ratcliff et al. (2004) Exp.1	-20440.7	**-20616.8**	176.1
4	Ratcliff et al. (2004) Exp.2	-33395.4	**-33402.3**	6.9
5	Ratcliff et al. (2004) Exp.4	-29059.4	**-29066.9**	7.5

Note. Bold WAIC values indicate the preferred model for the corresponding data set.

4.2.3 Model Fit

We checked fit by generating posterior-predictive data from the unequal models, simulating 100 data sets of the same size as the empirical data from 100 parameter-vector samples from joint-posterior distributions for each participant in each experiment. Figure 4.3 plots summaries of the observed and predicted data. To summarize the RT distributions, we present five quantile estimates (10%, 30%, 50%, 70% and 90%). The 10%, 50%, and 90% quantiles represent the leading edge, median, and tail of the distribution, respectively. These plots also indicate the proportion of correct (green) and incorrect (red) responses along the y-axis.

The top two panels in Figure 4.3 show empirical and predicted values for data sets 1 and 2 from Wagenmakers et al. (2008); the unequal model fits both well. The middle two panels and the bottom panel of Figure 4.3 displays the same for data sets 3-5 (Ratcliff, Gomez, & McKoon, 2004). The fits are good except for

consistent misses of the tail of the error RT distribution. This misfit is likely due to the low rate of errors and relatively high variability in the 90% quantile for error RTs. However, average speed of error vs. correct RT is still captured well.

4.2.4 Drift Rate Parameters

Figure 4.2 shows the mean of the posterior distributions of the group-level mean drift-rate and η estimates from the unequal DDM. The distributions for each stimulus class are the concatenation of the posterior distributions across all 5 experiments. The ordering of mean drift-rates are in agreement with REM–LD's evidence means within words and within non-words. The ordering of η for the DDM is mostly in agreement with the evidence variability predictions of REM–LD, with the exception of LF words. Drift variability was highest for HF words, followed by RL, VLF, LF, then PW. REM–LD predicted that evidence variability was highest for HF words, followed by VLF, LF, RL, then PW.

We used Bayesian predictive p-values to assess the probability that the difference between two posteriors is equal to or less than 0 (Meng, 1994). Small p-values in Table 4.3 suggest that the DDM and REM-LD are in agreement and larger p-values suggest that the two models are in disagreement. They mostly agree, except in regards to LF, RL, and VLF stimuli. The predicted order is reversed between LF and RL, and either reversed or equivocal between VLF and LF or RL.

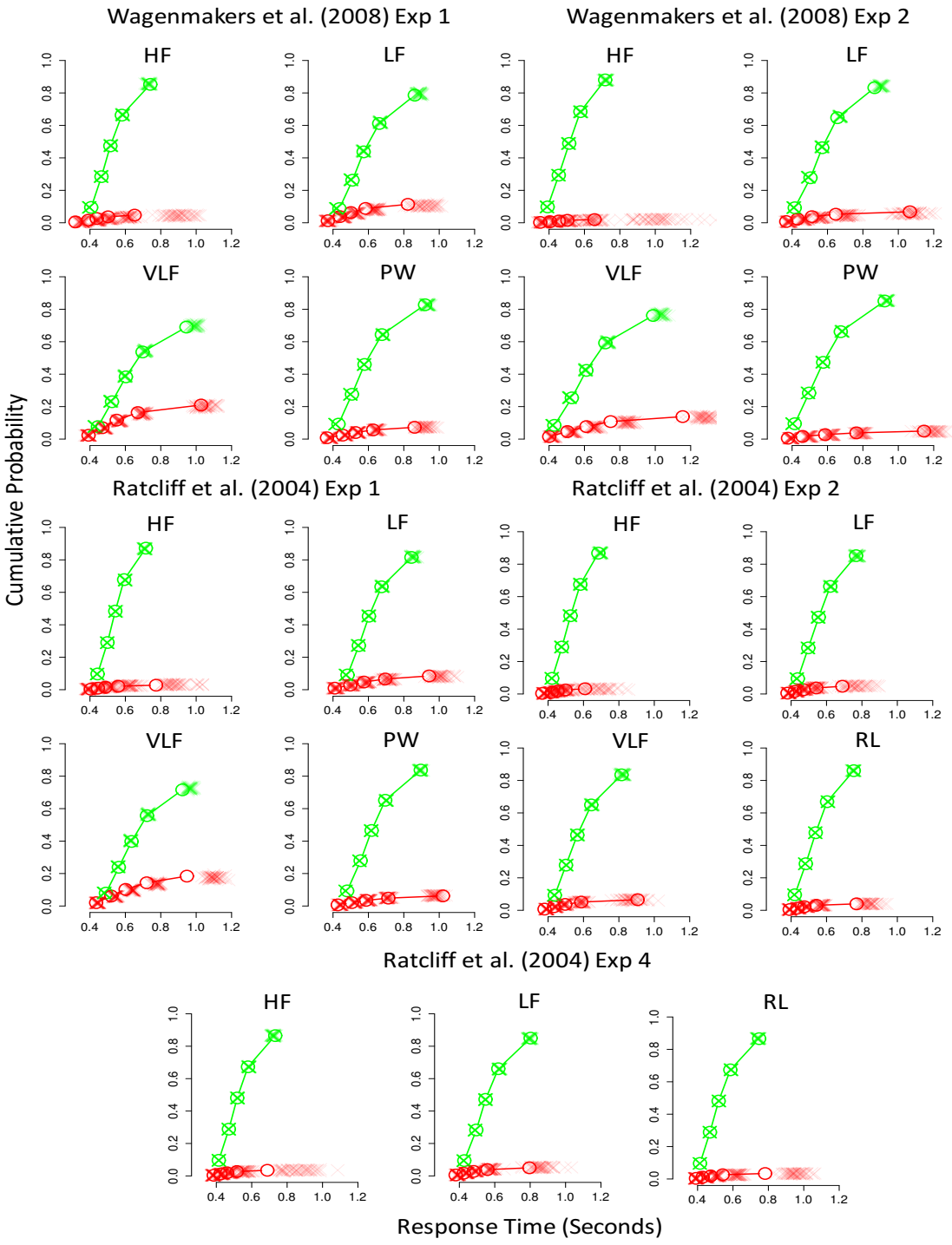


FIGURE 4.3: Defective cumulative distribution plots of the predicted RTs from the unequal model and empirical RTs for each stimulus condition. HF = high-frequency, LF = low-frequency, VLF = very-low-frequency, and PW = pseudo-word, and RL = random letter strings. The circles represent the empirical data and the crosses represent the predicted data. Note the predicted data consists of 100 separate data sets superimposed on the empirical data. The green points are correct responses and the red points are incorrect responses.

TABLE 4.3: Bayesian predictive p-values for drift variance ordering.

Data Set	HF ≤ LF	HF ≤ RL	HF ≤ VLF	HF ≤ PW	LF ≤ RL	LF ≤ VLF	LF ≤ PW	RL ≤ VLF	VLF ≤ PW
1	< .001	-	< .001	< .001	-	.985	< .001	-	< .001
2	< .001	-	.01	< .001	-	.972	.008	-	< .001
3	.012	-	.040	< .001	-	.679	< .001	-	.001
4	.011	.1	.015	-	.875	.556	-	.556	-
5	< .001	.074	-	-	.978	-	-	-	-

Note. Low p-values suggest that the DDM and REM–LD are in agreement.

4.3 General Discussion

The lexical-decision task has often been conceptualized as a specific case of signal detection theory (SDT; Norris, 1986; Balota & Chumbley, 1984), with decisions based on a continuously distributed evidence variable (i.e., word-likeness). The outcomes of decisions depend on both the mean and variance of evidence, but previous studies have assumed that these evidence distributions are equally variable for words and non-words (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers et al., 2008) with some supporting evidence from choice data (Jacobs et al., 2003; S. D. Brown & Steyvers, 2005). This implies that performance depends purely on evidence-distribution means. However, the latter investigations did not consider response times (RTs), which could potentially support different conclusions (Ratcliff & Starns, 2009).

This turned out to be the case, with our analysis based on both RTs and accuracy clearly rejecting the equal variance assumption (see also Dutilh et al., 2009, 2011, 2012). These results imply that researchers should take account of factors that affect the variability in evidence as well as its mean. For example, number of letters, orthographic neighborhood size, average base-word frequency, and average base-word number of syllables are factors known to affect between-item

variability in response times and accuracy (Yap, Sibley, Balota, Ratcliff, & Rueckl, 2015). Estimation of inter-trial drift variability is sensitive to variability in RT and accuracy, and so, it seems likely that these item level differences will be influential on the magnitude of inter-trial evidence (i.e., drift rate) variability.

We also investigated the Retrieving Effectively from Memory model of Lexical-Decision (REM–LD). REM–LD is based on a general model architecture that provides a comprehensive explanation of human memory. In REM–LD, stronger matches between the probe and trace skew the evidence distribution, which produces greater evidence variability for words than non-words, particularly for higher frequency words. Using typical parameter settings, we showed that REM–LD makes the prediction that the evidence variability will be largest for high-frequency words, followed by low-frequency, random letter strings, very-low-frequency, and finally pseudo-words.

We fit the Diffusion Decision Model (DDM; Ratcliff, 1978) to free-response lexical-decision data and examined the parameter estimates of inter-trial drift rate variability, which is analogous to evidence variability in SDT and REM–LD (Ratcliff, 1978, 1985). We found that the predictions of REM–LD were comparable to the DDM's evidence variability estimates for all word frequency conditions except low-frequency words. Specifically, the DDM predicted drift variability was highest for high-frequency words, followed by random letter strings, very-low-frequency, low-frequency, then pseudo-words. Overall, our results are encouraging because two prominent models of lexical-decision mostly agreed about predictions of word and non-word evidence variability.

Evidence variability occurs because items in the same category do not have

the same word-likeness value, or in terms of the DDM, the same drift rate. Intuitively, one might assume that higher frequency words are less variable than lower frequency words; perhaps because people might not know the definitions to some lower frequency words, making them more like non-words and inflating the variability. Despite this intuition, we observed that higher frequency words are more variable. Under REM–LD, the reason that higher frequency words are more variable is because of the way lexical retrieval operates by comparing a probe cue to all of the traces in the participant's lexical memory. When the probe cue is a word, it produces a strong match to its own trace and a weak match to all of the other traces in lexical memory. When these matches are averaged together, the contribution from the strong match skews the posterior odds ratio distribution, producing greater variability for words than non-words and greater variability for higher frequency words relative to lower frequency words.

Our results parallel findings from the recognition memory literature, where inter-trial drift rate variability is higher for studied (i.e., stronger) items (Ratcliff & Starns, 2009; Starns & Ratcliff, 2014; Starns, Ratcliff, & McKoon, 2012; Osth, Dennis, & Heathcote, in press). Models of recognition memory employ the same retrieval structure as REM–LD and predict higher variability for studied items for a similar reason: recognition is carried out by matching a cue vector against each memory, calculating the similarity, and making a decision based on either the summed or averaged similarity. Findings about evidence variability have played a crucial role in developing a theoretical understanding of recognition memory (Wixted, 2007; Osth & Dennis, 2015; Shiffrin & Steyvers, 1997), and our results suggest that they may play a similar role for theories of lexical memory.

Chapter 5

Comparing Prominent Sequential Sampling Models

All sequential sampling models have the same underlying assumptions about the decision-making process: they assume there is an initial amount of evidence for all response options and more evidence is continually sampled from the environment until an evidence criterion is reached for a particular response. When this criterion is reached, a decision is made. Beyond these assumptions, each sequential sampling model describes the decision process differently. The problem is that we have no way to identify the correct model – as "all models are wrong, but some are useful" (Box & Draper, 1987, p. 424) – and with many models being used to describe the processes underlying decision-making, we must at least be sure that the conclusions being made are not dependent on what model is used.

Previous work has been insightful on the consistencies between two of the more prominent sequential sampling models – the Diffusion Decision model (DDM; Ratcliff, 1978) and the Linear Ballistic Accumulator model (S. D. Brown & Heathcote, 2008) – in terms of the conclusions they draw about underlying psychological processes. Researchers have found differences in how the models account for practice effects in a lexical decision task (Heathcote & Hayes, 2012) and performance in a reward maximization task (Goldfarb, Leonard, Simen, Caicedo-Núñez, & Holmes, 2014). In a comprehensive comparison between the DDM and LBA, Donkin, Brown, Heathcote, and Wagenmakers (2011) observed no differences between the models in empirical data, but in simulated data, the authors found differences in response caution between both models. For the most part, these models have been found to be in agreement with each other.

This paper highlights two differences between the DDM and the LBA that have practical implications. The first distinction is the LBA's capacity to predict

an overall increase in mean drift rate toward all response boundaries. The DDM can produce equivalent effects by increasing the within trial noise of the diffusion process. The second distinction is the estimation of non-decision time processes for both models. Before I discuss both these differences, I briefly outline the DDM and LBA models.

5.1 Increasing Overall Drift Rate

The DDM is part of a more general class of sequential sampling models that assume relative evidence accumulation (Link & Heath, 1975; Laming, 1968; Stone, 1960). This means that evidence for one response option is evidence against the other response option. On the other hand, the LBA is part of the racing accumulator framework (S. D. Brown & Heathcote, 2005; Verdonck & Tuerlinckx, 2014), where each accumulator independently accumulates evidence toward the response threshold.[1] The overall drift rate effect I discuss next distinguishes classes of models that assume relative evidence accumulation from models that assume independently racing accumulators.

The LBA can estimate an overall speed up toward both the correct and incorrect response boundary. In the DDM, a speed up of mean drift rate toward the correct boundary is a slow down of the mean drift rate toward the incorrect boundary. In Figure 5.1, are the effects of increasing overall mean drift rate for correct and incorrect accumulators for the LBA. Initially, the mean drift rate for the

[1]Note that when there is competition between accumulators, and if competition is strong enough, then independent accumulation resembles relative accumulation (Usher & McClelland, 2001; Teodorescu & Usher, 2013; Teodorescu, Moran, & Usher, 2015).

correct accumulator is 2 and the mean drift rate for the incorrect accumulator is 0, making for an average mean drift rate of 1. When both correct and incorrect mean drift rates are increased, then the overall mean drift speed up reduces the spread of both the correct and incorrect response time (RT) distributions. In particular, the tail, or .9 quantile (q.9), shifts closer to the median as overall mean drift rate increases. We can quantify the associated change in the skew of the distributions, where skew is defined as (q.9 - q.5) - (q.5 - q.1). The change in skew from an average drift of 1 to 3 is .456 for the correct RTs and 1.26 for the error RTs.

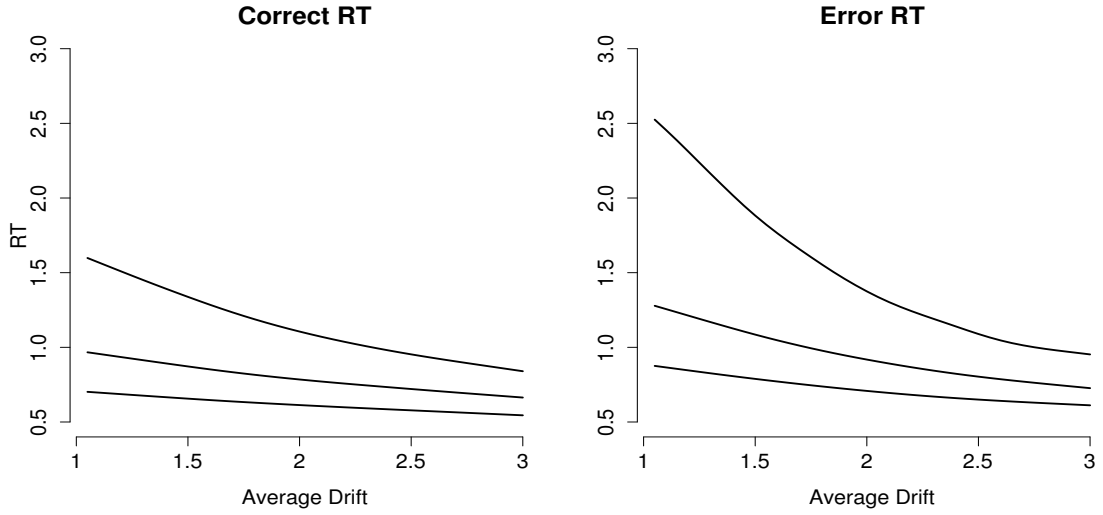


FIGURE 5.1: Effects of increasing overall mean drift rate on correct and error RT distributions when other LBA parameters are held constant. The x-axis shows the average of the correct and incorrect mean drift rate, which increases as both correct and incorrect mean drift rates increase. In both panels the top, middle, and bottom lines represent the .9, .5, and .1 quantiles, respectively. Values of the other parameters were $A = 1$, $b = 2$, $s = 1$, $t_0 = 300$ ms.

In practice, this speed up toward both thresholds could arise in speed vs accuracy manipulations. For example, Rae et al. (2014) presented results from fitting the LBA to experiment 1 from Wagenmakers et al. (2008). In this experiment participants identified whether a string of letters was a word or a non-word. A key manipulation was the emphasis on speed or accuracy. In the speed emphasis

condition, participants were told to respond as quickly as possible. In the accuracy emphasis condition, participants were told to respond as accurately as possible. Importantly, the mean drift rates were allowed to vary over speed and accuracy emphasis.

The mean drift rates from the LBA fit in Rae et al. (2014) are shown in Figure 5.2. The Figure shows the correct (dots) and incorrect (crosses) mean drift rates in the speed condition (in gray) both increase relative to the accuracy condition, demonstrating an overall mean drift rate increase.

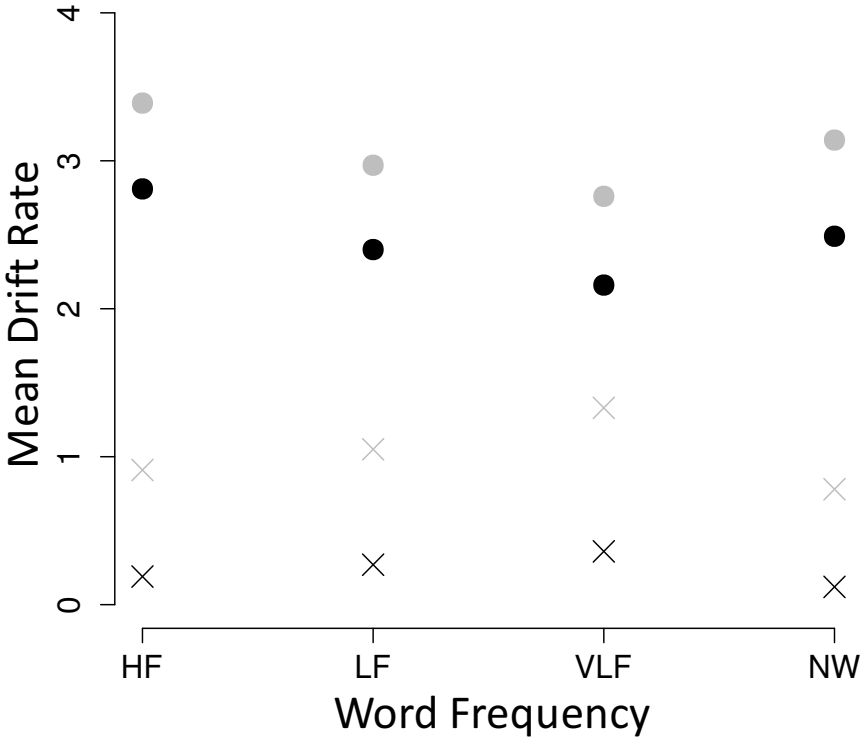


FIGURE 5.2: Mean drift rate estimates taken from Table 5 from Rae et al (2014). Drift rate parameters are from an LBA model fit to Experiment 1 of Wagenmakers et al. (2008). The dots are for the correct response and the crosses are for the incorrect response. Grey points are for the speed emphasis condition and black dots are for the accuracy emphasis condition.

But how does DDM explain a speed up toward both boundaries? To answer this I simulated data from the LBA with the following parameters: $A = 1$, $b = 2$,

$s = 1$, $t_0 = 300$ms, $v1_{correct} = 3$, $v2_{correct} = 3.5$, $v1_{incorrect} = 1$, $v2_{incorrect} = 1.5$. The main manipulation is an overall increase in mean drift rate for both correct and error accumulators, where the first condition has an average mean drift (averaged over accumulators) of 2 and the second condition has and average mean drift of 2.5.

In Figure 5.3 are two sets of 20 simulated diffusion processes. The figure shows that increasing s increases the variability of the evidence accumulation processes, which results in faster accumulation towards both boundaries and relatively more errors (i.e., terminations at the bottom boundary). Increasing overall mean drift rate in the LBA leads to a similar decrease in RTs and an increase in errors.

To test if changes in s can account for overall mean drift rate increases in the LBA, I fit the DDM using hierarchical Bayesian methods to data generated from the LBA. I fixed the s parameter to 1 in the average drift = 2 condition and estimated s in the average drift = 2.5 condition. I found that an overall increase in mean drift rate in the LBA results in an increase in the s parameter in the DDM: the mean group-level mean s was 1.29 in the average drift = 2.5 condition. As shown in Figure 5.4, the DDM with a free s parameter provides an adequate fit to the data generated by the LBA.

The LBA can predict an increase in mean drift rate in all accumulators. The DDM can attribute this speed up to an increase in the within trial noise of evidence accumulation. However, because s scales the parameters of the DDM, increasing s is equivalent to decreasing all of a, v_{DDM}, and sv by multiplying them by $\frac{1}{s}$.[2] Meaning that another, yet more complicated, interpretation of the diffusion

[2]Only a, v_{DDM}, and sv need to be scaled if z and sz are estimated as relative to a, which is the case here.

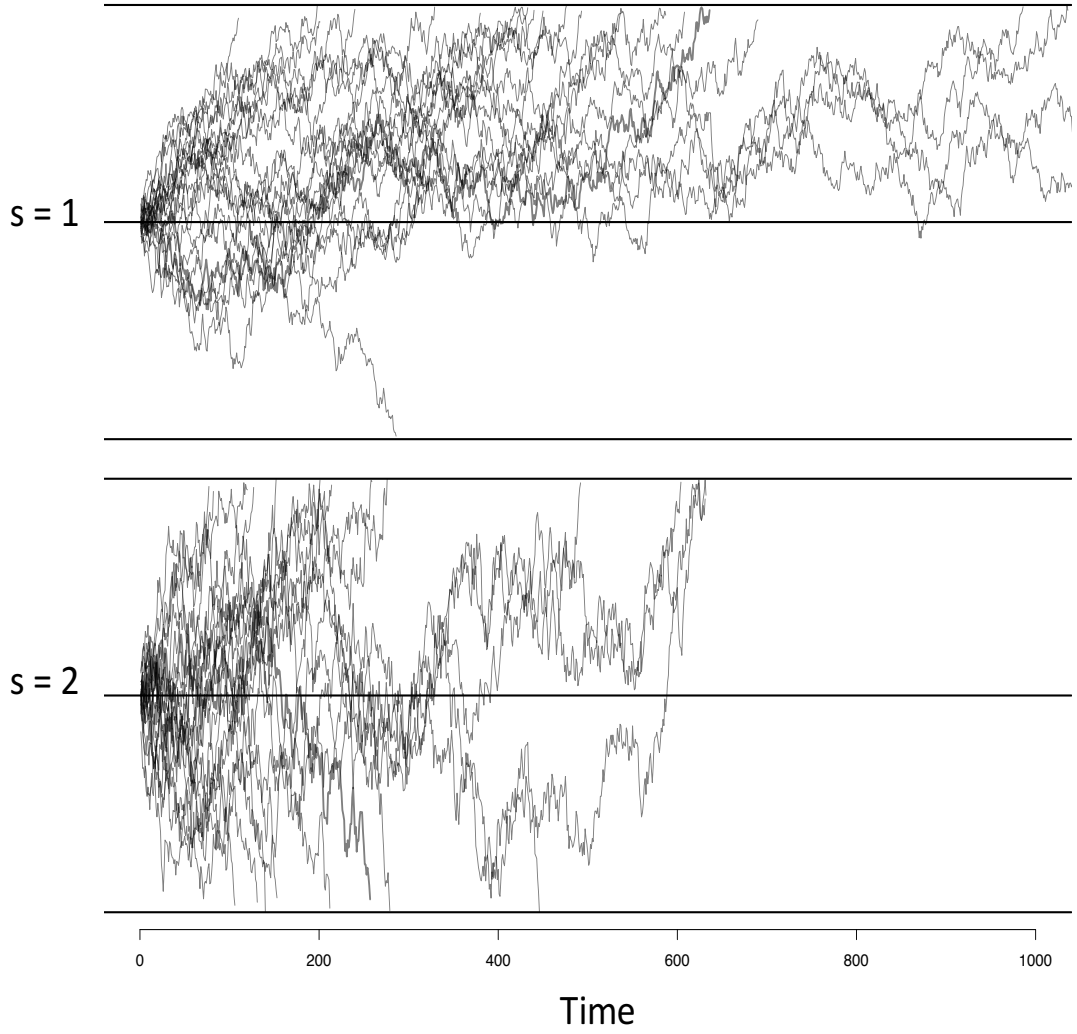


FIGURE 5.3: Simulated diffusion processes. Each panel shows 20 processes simulated by random walks. For both simulations, $a = 2$, $z = \mathrm{a}/2$, and $v = 2$. For the top panel $s = 1$ and for the bottom panel $s = 2$. In both panels, the upper boundary is for the correct response and the bottom panel is for the incorrect response.

model results, is that the increased rush toward both decision boundaries is caused by lower boundary separation, lower mean diffusion drift rate, and lower between trial variability in drift rate.

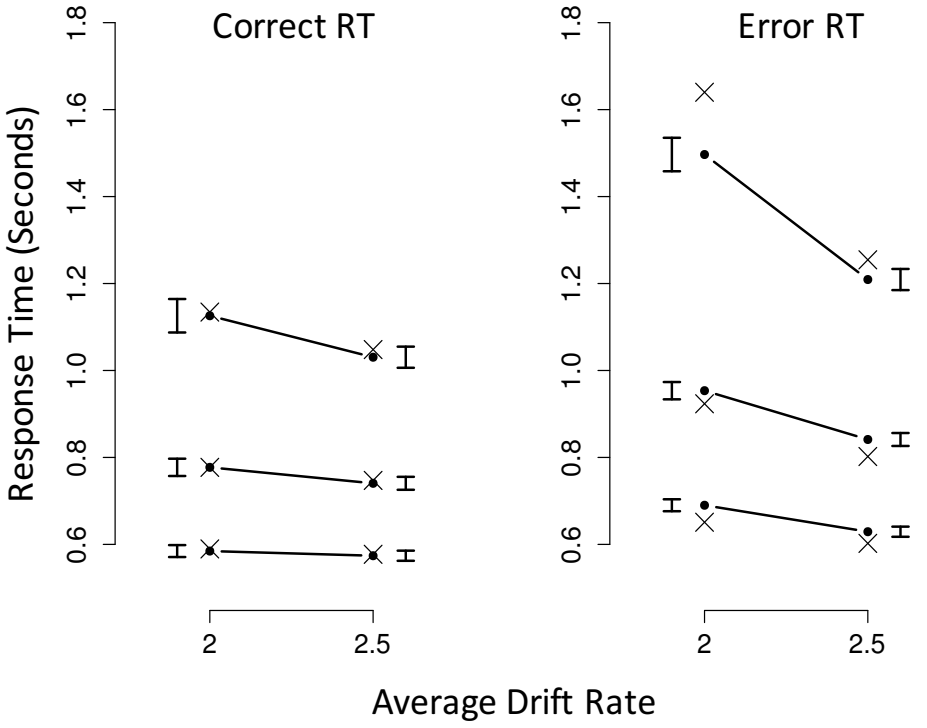


FIGURE 5.4: Simulated RT distributions from the LBA mode (black dots) and predicted RT distributions from the DDM model (black crosses). RT distributions are summarized as by their .1, .5, and .9 quantiles. Error bars represent 1 standard error of each quantile. Predicted data from the DDM was generated from the median group level mean posterior estimates.

5.2 Estimating Non-Decision Time

Both the DDM and LBA explicitly model the decision process, which allows for researchers to determine the effects of sub-processes (e.g., drift rate) on decision time. The time needed for processes outside of the decision process is represented by the non-decision time parameter. The non-decision time parameter determines the smallest possible RT and accounts for shifts in the entire RT distribution.

Researchers who use the DDM typically assume that non-decision time is variable across trials. For example, Ratcliff, Gomez, and McKoon (2004) found that including non-decision time variability allowed the DDM to better fit the .1 quantile of RT distributions from a lexical decision task (Ratcliff, 2002; Ratcliff & Smith,

2004). However, assuming a constant non-decision time provides a substantial computational speed up, which is one reason why a constant value is assumed by the more parsimonious LBA. Whether a constant value of non-decision time is overly simplistic is yet to be formally investigated – in fact, the distributional properties of non-decision times are not well understood (Ratcliff, 2013; Verdonck & Tuerlinckx, 2015).

Perceptual encoding time for simple visual stimuli can be as fast as 50ms (Bompas & Sumner, 2011), while key-press responses can be as fast as 66ms (Smith, 1995, p. 585). More recently, researchers have related measurements of electrophysiology in monkeys (Cook & Maunsell, 2002), and MEG and EEG in humans (Amano et al., 2006; Tandonnet, Burle, Hasbroucq, & Vidal, 2005; Vidal, Burle, Grapperon, & Hasbroucq, 2011) to the RTs to detect simple visual stimuli. By doing so, researchers could partition out the time needed for visual perception, which is approximately 150-200 ms. Given that non-decision time is the sum of perceptual encoding and response production time, for tasks with a key press response modality and stimuli as complex as motion dots, for example, we should expect non-decision time to be at least 200ms.

Here I test if the DDM and LBA models agree about the time needed for non-decision processes. In a simulation study, Donkin et al. (2011) found that the DDM reliably estimates larger non-decision times than the LBA. In practice, the DDM also typically produces higher estimates than the LBA. For instance, when the LBA was fit to data from Ratcliff and Rouder (1998), the model produced non-decision time estimates (S. D. Brown & Heathcote, 2008, Table 2) between 27-66ms lower than the diffusion model (Ratcliff & Rouder, 1998, Table 1). In

some instances, the DDM estimates non-decision times greater than 400ms (Gomez, Ratcliff, & Childers, 2015), leaving only 62-118ms of the mean RTs to be accounted for by the decision process.

The difference between the LBA and DDM in terms of non-decision time can have practical implications. For example, Heathcote and Hayes (2012) demonstrated that non-decision time decreases with practice in a lexical decision task, but only for the DDM, not the LBA. van Ravenzwaaij et al. (2012) analyzed data from a motion dots task with the DDM and found that alcohol consumption impaired cognitive and motor/perceptual encoding capacity. I refit the data here with both the DDM and LBA model using hierarchical Bayesian methods.

We calculated the non-decision time values from the mean of the group-level posterior mean. With the DDM, non-decision time was 278 ms for placebo doses and this increased to 288 ms for high alcohol doses.[3] With the LBA, non-decision time was 107 ms for placebo doses and this decreased to 100ms for high alcohol doses. Therefore, if researchers had used the LBA, they would conclude that alcohol consumption does not lead to motor/perceptual encoding deterioration and that alcohol doses reduce motor/perceptual encoding time. Finally, the estimates of the LBA are below the expected 200 ms inferred from previous work.

[3]Note that the original authors found an increase of 19 ms rather than 10 ms in non-decision time between placebo and high alcohol doses.

5.3 Discussion

Sequential sampling models are useful because they translate response time (RT) data into interpretable psychological phenomena. These models allow researchers to draw conclusions about the speed of information processing, a priori response bias, response caution, and non-decision time. The increase in use of these models demonstrates that more and more researchers have discovered the merits of these analysis tools. But are the conclusions that are drawn dependent on what model is used?

Previous research has shown that the conclusions drawn from both the diffusion model (DDM) and linear ballistic accumulator (LBA) are mostly in agreement (Donkin et al., 2011). In the current paper, I highlight two key discrepancies between the DDM and LBA. Firstly, the LBA can predict an overall speed-up in mean drift rates for both responses. I show that this overall speed-up in drift rate occurs when participants are asked to prioritize speed over accuracy. The simplest DDM account for this overall drift speed up is an increase in within-trial drift variability. Secondly, the DDM estimates higher values of non-decision time than the LBA, which can lead to the two models disagreeing about non-decision time effects. Both these discrepancies could cause researchers to draw different psychological conclusions from the LBA or DDM.

As I have demonstrated in this thesis, both models have merit for their application in specific modeling applications. However, I suggest that researchers use caution when drawing conclusions based on either absolute drift rates or non-decision time. Fortunately, the models seem to agree more often than they disagree. The

conceptual differences highlighted in the current chapter are aimed to be informative for researchers who need to select between the DDM and LBA for inferring about the underlying processes of decision-making.

Chapter 6

A Novel Measure of Cognition Applied to the Stroop Task

Reference

Tillman, G., Eidels, A., & Finkbeiner, M. (2016). A Reach-To-Touch Investigation on the Nature of Reading in the Stroop Task. Attention, Perception, & Psychophysics.

6.1 Introduction

Participants in the Stroop task (Stroop, 1935) are presented with a color-word printed in color and must respond to the color and ignore the word. They are faster, on average, to name the print color of a congruent color-word stimulus (e.g., the word RED printed in red) than an incongruent stimulus (e.g., GREEN in red). The Stroop effect is calculated as the difference in response time (RT) between congruent trials and incongruent trials, and demonstrates the unintended influence of the word. It is one of the most replicated experimental effects in cognitive psychology, yet despite years of research there is no agreed theoretical resolution as to the cause of the effect (MacLeod, 1991; Eidels, Townsend, & Algom, 2010; Eidels, 2012).

Theoretical accounts of the Stroop effect (e.g., Palef & Olson, 1975; Logan, 1980; Cohen, Dunbar, & McClelland, 1990; Melara & Algom, 2003) must assume that participants process the meaning of the printed words despite instructions to ignore them and focus on the print color; otherwise the time to respond 'red' should be the same for any word printed in that color, regardless of whether it is congruent or incongruent – and hence, there would be no behavioral Stroop effect.

RTs have been the preferred dependent variable in many psychological experiments (see Luce, 1986), including the Stroop task. Researchers used RTs to determine that the Stroop effect is contingent on 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).

Despite their benefits, RTs provide only a single estimate of processing duration at the end of each trial. Meaning there are limitations to what RTs can tell us about the time course of an experimental effect. For example, an RT of 500ms on a given trial of the Stroop task suggests that it had taken 500ms to perceptually encode a stimulus, process and decide on the color of the stimulus, and execute a behavioral response. However, we do not know how long each of these sub-processes take.

There are statistical methods that provide insight into the time course of experimental effects. Parametric studies can fit sequential sampling models to RT distributions and estimate perceptual encoding time and rate of processing, but these models are esoteric in cognitive psychology (e.g. Ratcliff & McKoon, 2008; S. D. Brown & Heathcote, 2008). Alternatively, more general graphical exploration methods of RT distributions, such as the delta plot, inform researchers about the time course of experimental effects (De Jong, Liang, & Lauber, 1994).

6.1.1 Delta Plots of Stroop Data

Delta plots display graphically how an experimental effect changes across different points of two RT distributions. For instance, the left panel of Figure 6.1 shows congruent and incongruent RT distributions of a hypothetical Stroop task. For delta plots, instead of calculating the Stroop effect as the difference between mean RTs of incongruent and congruent conditions, a researcher calculates the effect at a

desired number of percentiles (e.g., at each decile of the two distributions). They could then plot the effect at each decile against the mean RT of the two distributions at each decile.[1] The resulting delta plot is shown in the right panel of Figure 6.1. This function is always above 0, meaning that RTs in the incongruent distribution are slower than RTs in the congruent distribution for every decile. The positive slope suggests that the difference between the incongruent and congruent RTs is bigger for slower RTs than faster RTs.

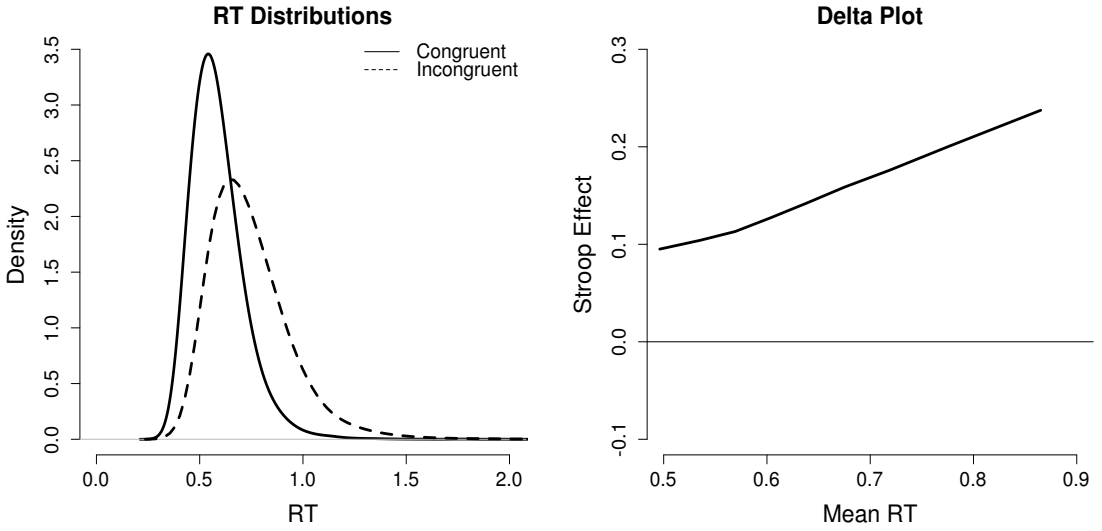


FIGURE 6.1: The left panel depicts RT distributions for the congruent and incongruent conditions of a Stroop task. The right panel depicts the resulting delta plot.

Pratte, Rouder, Morey, and Feng (2010) used delta plots to investigate the distributional properties of the Stroop and Simon effects, and found delta plots with different slopes. Specifically, the slope for the Stroop effect delta function was positive, with small values for fast responses and larger values for slower responses. In contrast, the slope of the Simon effect delta function was negative, with large values for the fast responses and smaller values for slower responses. However, the

[1]The benefit of plotting percentile effects as a function of percentile means is that the delta plot will be linear (Speckman, Rouder, Morey, & Pratte, 2008).

delta plot slope depends on the exact nature of the task (cf. Proctor, Vu, & Nicoletti, 2003; Proctor & Shao, 2010; Dittrich, Kellen, & Stahl, 2014). The negative slope for the latter suggests that the Simon effect results from a conflict at the motor response stage, which decays over time. The positive slope for the Stroop effect delta function suggests the effect results from a conflict at the processing stage, which grows in magnitude as the participant processes the stimulus for a longer duration.

A potential limitation of the delta plot method is its sensitivity to the difference in variance between the distributions in question. This point is illustrated in Figure 6.2, where we show delta plots that compare gamma distributions with different means and standard deviations (SD) to a gamma distribution with fixed arbitrary parameters – mean = 12 and SD = 3.[2] The middle panel serves as a benchmark and shows the delta plot of two identical gamma distributions with mean = 12 and SD = 3, resulting in a flat line at 0. Each column represents distributions with a different variance and each row represents distributions with a different mean. The effect of changes in variance on the slope of the delta plot, while the mean is held fixed, can be observed by moving along the columns within any given row. Critically, if one RT distribution had the same mean but larger variance than the other RT distribution, then the slope of the delta plot will be positive, regardless of the mean RT. Note that for empirical RT distributions the standard deviation of RT typically increases linearly with the mean (Wagenmakers & Brown, 2007), although

[2] Typically the gamma distribution is parameterized with the shape and rate parameter. Given that the mean of the gamma distribution = $\frac{shape}{rate}$ and the SD of the gamma distribution = $\frac{\sqrt{shape}}{rate}$, the shape and rate parameters we used to generate data were shape = $\frac{mean^2}{SD^2}$ and rate = $\frac{mean}{SD^2}$ (e.g. Kruschke, 2011, p. 170).

there are cases where this trend does not hold, such as the Simon task (Pratte et al., 2010).

There are limitations to investigating the time course of the Stroop effect using mean RTs, as they rely on a single measurement of latency at the end of each trial and ignore distributional information. The delta plot method makes use of the entire RT distribution, but effects of mean and variance are hard to discern (see Figure 6.2). Moreover, RT distributions can have different shapes and be shifted in time. Ideally researchers would like a measure that produces identically shaped distributions so that they can compare responses across the two distributions at the same points in time. We offer an alternative to the delta plot method. A method that allows researchers to look at experimental effects across identical distributions at the same points in time.

6.1.2 Reach-To-Touch Paradigm

A promising method in cognitive science is the reach-to-touch paradigm (Finkbeiner, Coltheart, & Coltheart, 2014). For instance, in the Simon task literature, the reach-to-touch paradigm has already been used to investigate temporal properties of the effect (Porcu, Bölling, Lappe, & Liepelt, 2016; Buetti & Kerzel, 2010; Finkbeiner & Heathcote, 2016). In a typical design, participants may be presented with a cognitive task that requires a speeded choice between two or more response alternatives. Participants execute their response by reaching out to designated spatial locations, say, left for color green and right for color red. The arm-movement trajectories are recorded and serve as the dependent measure.

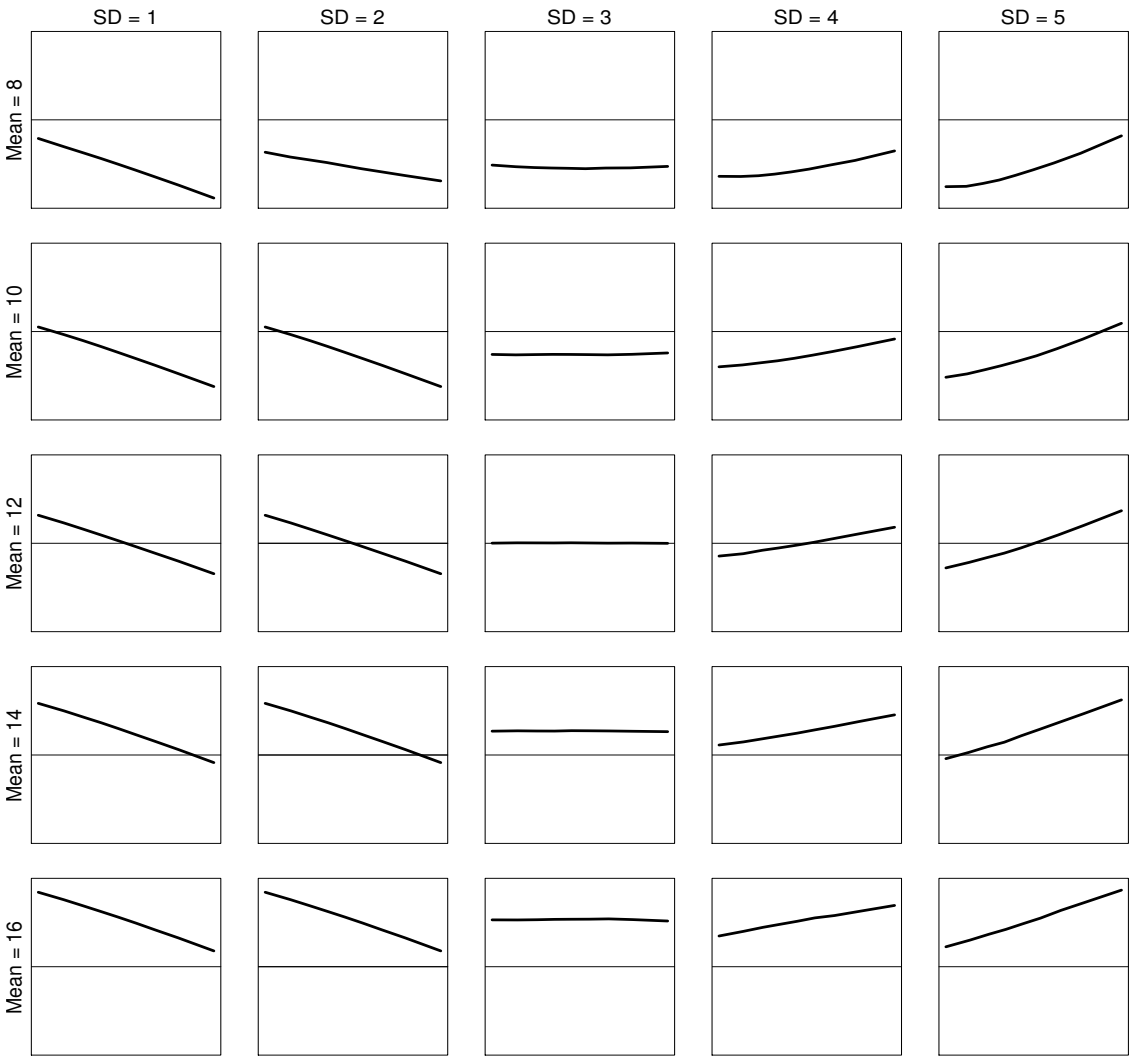


FIGURE 6.2: Simulated delta plots. Each delta plot is calculated by comparing gamma distributions with different means and SDs to a gamma distribution with mean = 12 and SD = 3. Each column shows a distribution with different SD and each row shows a distribution with different mean.

There are two key components to the reach-to-touch paradigm. First, it is a continuous response measure that can reveal experimental effects as they emerge over time. Arm movements in the reach-to-touch paradigm have been considered a window into cognitive processes (Song & Nakayama, 2009; Spivey, Grosjean, & Knoblich, 2005). More recently, Finkbeiner and colleagues (Finkbeiner et al., 2014; Quek & Finkbeiner, 2013, 2014) pointed out that this continuous response measure should be used with the second key component, the response signal procedure (Reed,

1973, 1976).

The current study instructs participants to initiate their movement within 300ms of an imperative 'go' signal. This go signal is the final beep in a sequence of 3 beeps. Importantly, on each trial the 3 beeps occur randomly so that the final 'go' beep appears at different points in time relative to the onset of the target stimulus. The time at which the participant begins moving relative to stimulus onset is the movement initiation time (MIT). For example, MIT = 0 means that the subject started to move their finger at the same time as the stimulus was presented. Similarly, MIT = 300 indicates that the subject lifted their finger from the start point 300ms after the stimulus onset. A negative MIT means that the subject starting moving their finger before having seen the stimulus. MITs represent movements that commence at a range of different stimulus processing times. In the Stroop milieu, we can examine the magnitude of the Stroop effect for various processing times (i.e., is the observed effect larger on late lift-off trials, which presumably allow more time for processing).

6.1.3 The Forced-Reading Stroop Task

As well as the statistical methods discussed, recent experimental methods have shed light on the nature of the Stroop effect. Eidels, Ryan, Williams, and Algom (2014) employed a novel forced-reading Stroop task and found the standard Stroop effect is only a proportion of the Stroop effect that could be observed. In the standard task participants are 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. Eidels et al. found a larger Stroop effect in the benchmark forced-reading task compared to the standard Stroop task, and suggested that the nature of reading occurring in the two tasks is not comparable.

6.1.4 The Current Study

In the present study we use both the standard and forced-reading Stroop tasks in conjunction with measurements of arm-reaching trajectories to understand the time course of the Stroop effect.[3] The forced reading task is a useful benchmark as it yields a Stroop effect from fully processed words.

The key aspect of our study is that distributions of MITs do not differ across conditions in our experiment (see Figure 6.3). There were no differences in the means or the SDs of how long subjects view and presumably process the stimulus before initiating their movement. Therefore, we compared arm reaching trajectories across two identically shaped Stroop distributions to see if the Stroop effect unfolds at a different rate, for the standard and forced task, at the same points in time. Our analysis is not compromised by differences in variances or shapes between the

[3]The term 'standard Stroop task' is a neutral term we use to refer to a Stroop task in which participants are not ensured to read on each and every trial. Standard Stroop tasks typically use response time as the dependent measure, have a vocal mode of responding, and have more than two color stimuli (but see MacLeod, 1991).

congruent and incongruent distributions, thus our study addresses concerns with the delta plot.

We address four key research questions in our study. First, we expect that participants will be more informed of the correct response with additional processing time. So do participants get a better idea of how to respond at later MITs? Second, there is an increased task demand in the forced-reading Stroop task because participants are required to read each and every word. But does this task demand result in the decision process unfolding faster in the standard Stroop task compared to the forced-reading Stroop task? Third, researchers have inferred from delta plots that Stroop interference grows over time (Pratte et al., 2010). This conclusion is also in line with extant theories of the Stroop effect (Cohen et al., 1990; Melara & Algom, 2003). However, given the limitations of the delta plot, we investigate whether the Stroop effect (when it exists) grows over time in the reach-to-touch-paradigm. Finally, the standard Stroop has previously been found to be a proportion of the benchmark forced-reading Stroop effect (Eidels et al., 2014). With our method we look at whether the Stroop effect grows in magnitude in the forced-reading task more than the standard Stroop task as stimulus-processing/viewing time increases.

6.2 Method

6.2.1 Participants

Twenty psychology students from Macquarie University participated in the study in return for course credit. All participants were native English speakers with

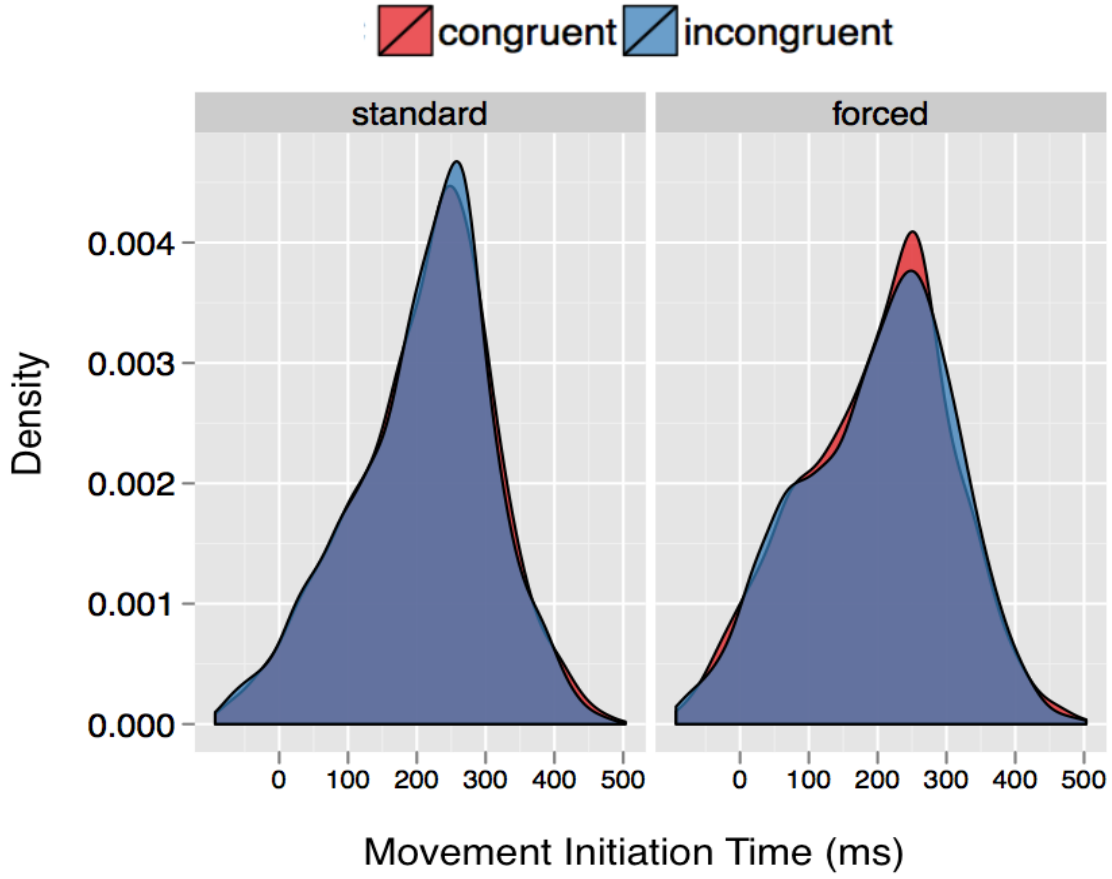


FIGURE 6.3: Distribution of MITs for congruent and incongreunt conditions in the standard and forced Stroop tasks. The four MIT distributions of interest do not differ in location or scale. Zero value on the x-axis means that the participant initiated movement at the same time as the stimulus onset. The figure shows that the majority of responses were initiated after stimulus onset. In the standard task the mean MIT was 172ms in the congruent condition and 169ms in the incongruent condition. In the forced task the mean MIT was 166ms in both the congruent and incongruent conditions.

normal or corrected to normal vision, intact color vision, and reported to be right handed. All participants took part in both the standard and forced-reading Stroop tasks.

6.2.2 Apparatus

A schematic of the experimental apparatus is presented in Figure 6.4, with important materials labeled with numbers. Participants sat in front of a table and

placed their right index finger on a small Velcro square (marked '0' in Figure 6.4), which marks the starting position and the return position for every trial. Stimuli were presented on a 27" Samsung LCD/LED monitor using the software 'Presentation'. The monitor was situated 1m away from the participants and centered with their body mid-line. Lateral response boards (30cm x 9cm) were placed to the left (1) and right (2) of the monitor, 75cm apart and 50cm from the front of the desk. A third response location (3) was marked on the desk between the participant and the monitor, 50cm away from the front edge of the desk. A small motion-tracking sensor was taped to the tip of the right index fingertip of each participant. A Polhemus Liberty (240Hz) electromagnetic motion tracking system was used to record the participants arm trajectories during the experiment. Participants wore headphones adjusted to a comfortable volume level, which were used to present a sequence of beeps.

6.2.3 Stimuli

The standard Stroop task and the forced-reading Stroop task used the same stimuli. The stimuli were the color-words: RED and GREEN; and the non-color-words were: ROD, BED, RENT, QUEEN, GRAIN and GREED. These non-color stimuli were specifically selected to ensure that participants would not base their responses on local cues. The non-color stimuli were the orthographic neighbors of the color-words with the closest frequency, such that each non-color-word shared all but one or two letters with a color-word (see Eidels et al., 2014). All words were printed in either the color red or green (with RGB values of 220/0/0 and 0/170/0,

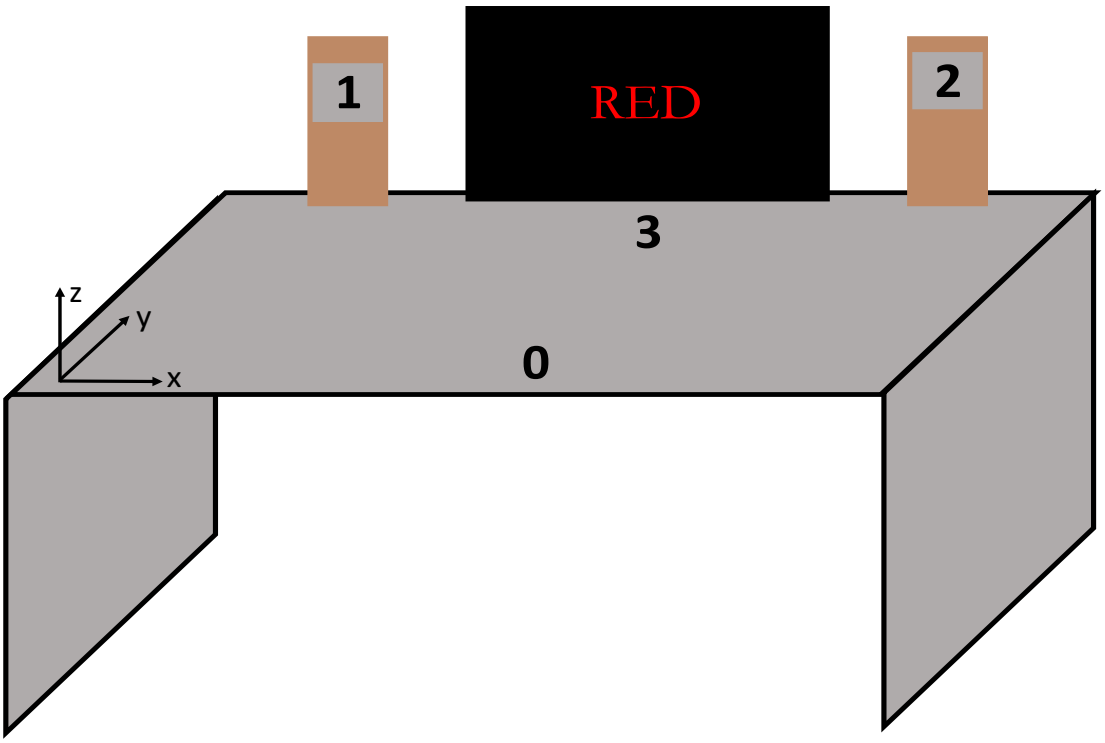


FIGURE 6.4: A front facing view of the apparatus used for the current experiment. Subjects placed their index finger on position 0 to start the trial. On each trial participants reached toward the color response options, denoted by 1 and 2. In the forced-reading task, participants could also reach towards a neutral response option, denoted by 3.

respectively) and were written in uppercase Garamond font, which at a viewing distance of 1m allowed for a visual angle of 4 degrees. Each of the color-words could be congruent to the font color (e.g., RED printed in the color red) or incongruent (e.g., RED printed in the color green). All non-color-words can be considered neutral to the font color, whether they were printed in red or green (but see T. L. Brown, 2011).

6.2.4 Design and Procedure

Each participant attended two experimental sessions: the standard Stroop task and the forced-reading Stroop task. Sessions were separated by a minimum of 1

day and a maximum of 7 days. The order of task administration was counterbalanced across participants so that half of the participants performed the standard Stroop task first, and the remaining half performed the forced-reading Stroop task first. The order of word presentation was random for each participant. For each session, the participant performed in 840 trials. These trials were partitioned into seven blocks of 120 trials each. There were 2-minute breaks between each block administration. In each block, color-words were presented 15 times per combination of color × word (RED in red, RED in green, GREEN in red, and GREEN in green), which made for 60 color-word trials. The six non-color-words were presented 5 times per combination, making for 60 non-color-word trials within the same block.

In the standard Stroop task the participant classified the color of all the words presented by reaching out to the left or right lateral response boards ('1' and '2' in Figure 6.4). The left and right response boards corresponded to a red or a green color and were counter balanced across participants. In the forced-reading Stroop task, participants classified the color of color-words but did not classify the color of non-color-words. For non-color-words, participants responded by reaching towards a neutral response location ('3' in Figure 6.4).

On each trial, a single word in color was presented at the center of a black screen. The timing of stimulus presentation was relative to the sound of three auditory beeps that were played through the participant's headphones. The stimulus was randomly presented, with equal probability, at one of five different times prior to the third beep (300, 230, 150, 70, or 0ms before the third beep). In four of the five timing conditions (i.e., 80% of trials) the stimulus was presented before the onset of the third beep, whereas in the 0ms condition (20% of trials) the stimulus and

the third beep were presented simultaneously. This procedure controls for participant's anticipation of stimulus display. In both tasks, participants had to initiate their movement between 100ms before and 200ms after the third beep, meaning all movement begun within a 300ms window around the third beep. Two example trial-sequences are presented in Figure 6.5. If participants failed to initiate movement within the allotted time-window they would receive a loud buzzing sound and visual feedback to indicate they had responded 'Too Early!' or 'Too Late!'. Once a movement was initiated, participants were required to maintain a continuous forward motion. Failing to do so terminated the trial and participants were provided with a buzz and appropriate visual feedback. Trials that were terminated via movement errors were repeated at a later stage of the block. The presentation of the trial terminated when the participant responded via the response points. The next trial followed after the sensor was returned to the start point.

6.2.5 Data Analysis

From the trajectories (Figure 6.6) we calculated the velocity along the x axis (x-velocity), which serves as our dependent measure. X-velocity quantifies how fast a participant is moving in the correct direction at any time during the trial. X-velocity is positive for movements towards the correct direction and negative for movements toward the incorrect direction. Thus, x-velocity provides data that ranges between fast movement in the correct direction (large positive values) and fast movement in the incorrect direction (large negative values). It is a more informative measure

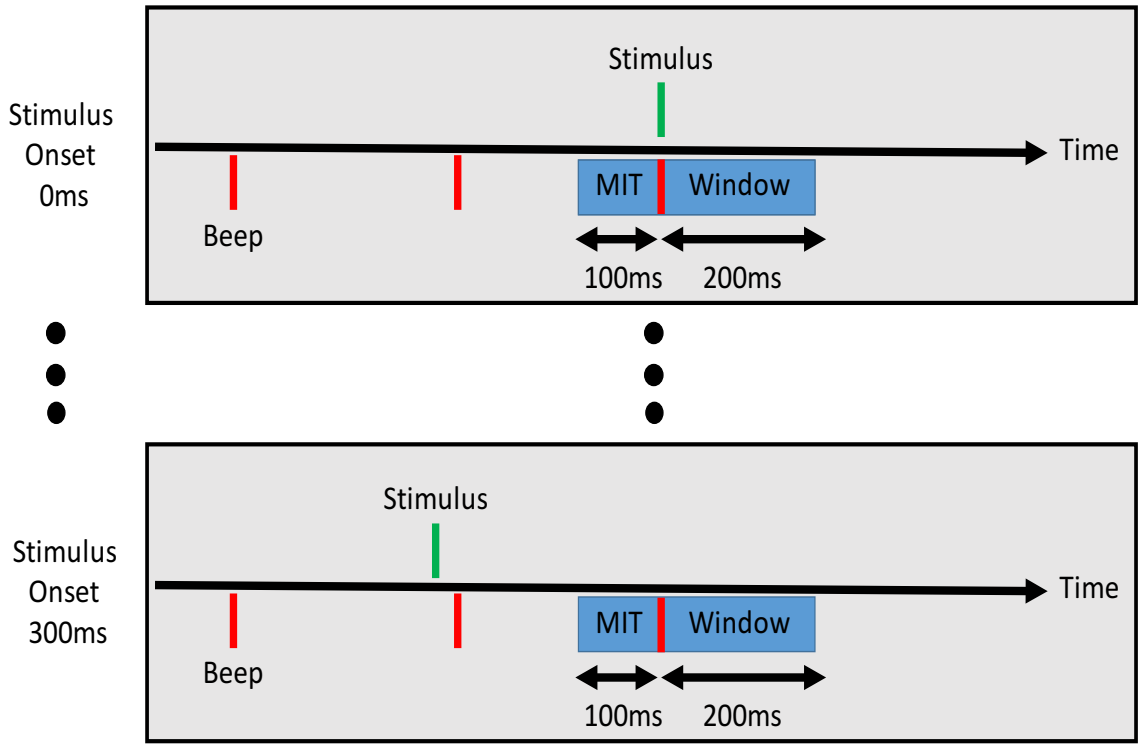


FIGURE 6.5: Example trial sequences for trials in which stimuli were presented simultaneously with the third beep (top panel; 0ms gap between the onset of the stimulus and the third beep) and 300ms before the third beep (bottom). The red vertical bars below the time line indicate the onset of the three auditory beeps, the green bar above the time line indicates stimulus onset, and the blue box shows the 300ms window in which participants begun their movements. In addition to the 0 and 300ms trial types there were also trials in which stimulus onset preceded the third beep by 70, 150, or 230ms (not shown in the figure).

compared to nominal accuracy rates (correct/incorrect) or RTs, which range from 'slow' to 'fast' in only a positive direction.

Before calculating x-velocity, the positional data taken from the Polhemus Liberty device was filtered with a two-way low-pass Butterworth filter at 7Hz, which reduced noise in the data. Then, x-velocity was derived from the numerical differentiation of the filtered positional data. The onset of movement was identified as the first of 20 consecutive samples in which the tangential velocity exceeded 10cms/sec. The offset of movement was identified as the first of 20 consecutive samples of tangential velocity that occurred after peak velocity and that were less than 10cms/sec.

For our analysis, we first improved the signal to noise ratio of the trajectories with a modified version of orthogonal polynomial trend analysis (OPTA). The OPTA procedure used here has been described in detail in Finkbeiner et al. (2014) and Finkbeiner and Heathcote (2016). In summary, OPTA uses a regression model with x-velocity as the dependent variable and MIT (with polynomial terms up to the 15^{th} order) as the predictor variable. Terms that did not explain significant variance were removed from the model, leaving only significant coefficients to predict x-velocity for each trial. After the OPTA analysis, we calculated the mean predicted x-velocity values from the first 350ms of the reaching movement (initial x-velocity; Finkbeiner et al., 2014). We limit our dependent measure to the first 350ms because the initial part of the trajectory represents the motor plan participants had formulated just prior to initiating their movement. The MIT latencies were used to group the initial x-velocity profiles into 20 equal bins (i.e., semi-deciles). Finally, the mean predicted initial x-velocity values were then subjected to a linear mixed-effects model with MIT semi-decile included as a fixed effect.

6.3 Results

6.3.1 Accuracy

Overall, across all participants, 91% of the responses were correct and valid. Mean error rate amounted to a negligible 1%. Invalid responses consisted of responding too early (3%), responding too late (5%), and not moving fast enough (2%). None of the participants were excluded from analysis due to accuracy.

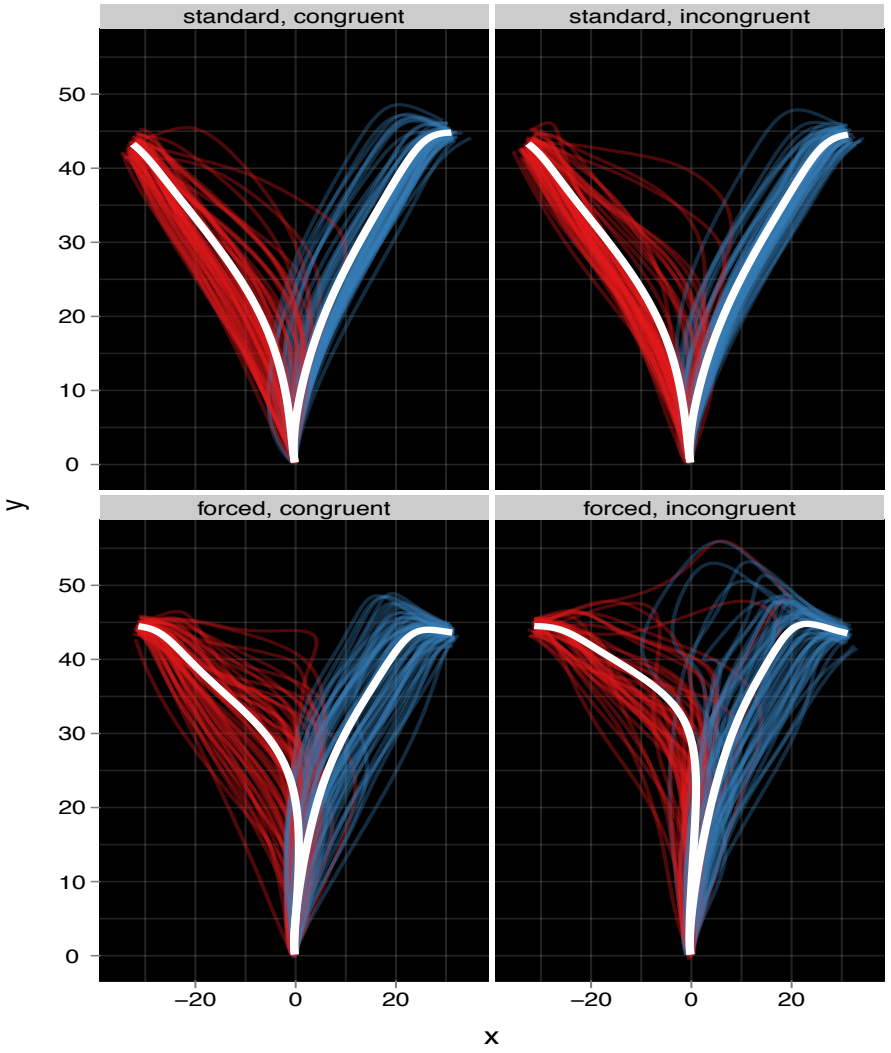


FIGURE 6.6: Arm trajectories and mean arm trajectories of a single participant. The four panels include arm trajectories related to the four possible conditions obtained from crossing Task by Congruence. X and Y labels refer to the movement planes presented in Figure 6.4. The Y axis denotes forward motion and the X axis denotes lateral motion. Trajectories only include correct responses. Thus, any differences between the left and right tracks are natural deviations in how the hand moves to a target situated to the left versus right of mid-line.

6.3.2 Linear-Mixed Effects Analysis

The Linear-Mixed Effects analysis on predicted initial x-velocity (x-velocity hereafter) was conducted only for correct responses. We used a model comparison approach with the Bayesian information criterion (BIC G. Schwarz, 1978), which selects the best fitting model while penalizing for complexity (i.e. number

of parameters). The best fitting model included Task (forced, standard), Condition (congruent, incongruent), and MIT (semidecile) as fixed effects. The model also included subjects as a random effect. The relationship between x-velocity and MIT was curvilinear and so the model included up to 3rd order terms for MIT. Here we report the coefficients (b), standard errors, and t-values of the best fitting model. The criterion for significance is a coefficient magnitude of at least twice the corresponding standard error. For the 'condition' factor, the congruent condition was used as a baseline meaning that negative coefficients represent smaller x-velocities relative to the congruent condition. For the 'task' factor, the standard task was used as a baseline meaning that negative coefficients represent smaller x-velocities relative to the standard task.

X-velocity was smaller in the forced Stroop task compared to the standard task (b = -34.80, SE = 0.32, t = -109.32). There was also a smaller x-velocity in the incongruent condition compared to the congruent condition (b = - 4.07, SE = 0.32, t = -12.87). X-velocity increased as a function MIT semidecile (b = 1606.03, SE = 21.07, t = 76.21). There was a significant interaction between task and condition, where the difference in x-velocity between congruent and incongruent trials was bigger in the forced task than the standard task (b = - 2.76, SE = 0.46, t = - 6.05). There was an interaction between task and MIT semidecile (b = - 825.75, SE = 30.10, t = - 27.43), but no interaction between condition and MIT semidecile (b = 59.50, SE = 29.85, t = 1.99). Finally, there was a three-way interaction between task, condition, and MIT semidecile (b = - 665.88, SE = 43.08, t = - 15.46).

To understand the nature of the three-way interaction we ran paired t-tests (congruent vs. incongruent) at each of the 20 MIT semideciles for both the standard

Stroop task and the forced Stroop task (Figure 6.7). We corrected for an inflated type I error rate with Bonferroni corrected p values. This analysis showed the Stroop effect unfolding over time. In the standard task the Stroop effect was not significant for any of the 20 MIT semideciles.[4] However, in the forced task the Stroop effect was significant for movements that commenced at the 7^{th} MIT semidecile ($\sim$ 133ms) through to the 20^{th} and final MIT semidecile ($\sim$ 338ms).

6.4 Discussion

Participants performed in both a standard and forced-reading Stroop task. The dependent measure for both tasks were the reaching trajectories. Using arm-reaching trajectories coupled with a signal-to-respond procedure allowed us to compare Stroop effects that are calculated from two identically shaped distributions. This way we could compare Stroop effects at the same points in time and presumably equivalent processing times. At each point in time we observed how fast the participant initially moved towards the correct response – initial x-velocity.

First, we wanted to know if participants get a better idea of how to respond with increased stimulus processing/viewing time (*processing time* for brevity). Initial x-velocity significantly increased as a function of MIT. Thus, the participant moved faster toward the correct response when they had more processing time. This finding might not be surprising as the participant would be more informed of the correct response with additional time. Nonetheless, this finding supports our claim that the

[4]We ran 20 Bonferroni corrected t-tests across the semi-deciles separately for the two different sessions. We found that the results of the standard Stroop effect was not dependent on the session order as no standard Stroop was found for either session.

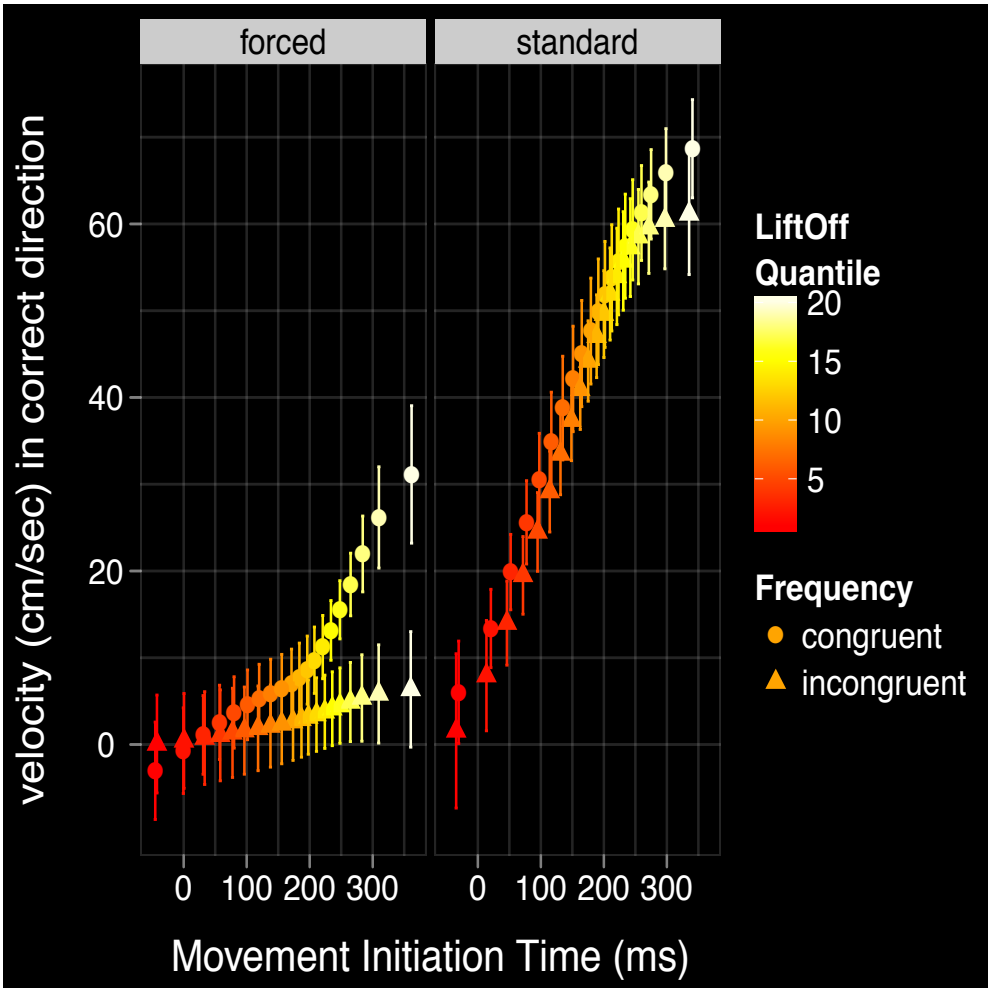


FIGURE 6.7: Initial x-velocity by MIT and Condition (congruent and incongruent) in standard and forced Stroop task. Error bars represent within-subjects 95% CIs. MIT values indicate the time delay between stimulus onset and the beginning of the response movement. The Stroop effect is represented by the vertical difference in height between the congruent (circles) and incongruent (triangles) markers at each quantile. Stroop effect is null on early quantiles of the forced task, but emerges later on. It is effectively null in the standard task, for all quantiles.

impact of increased processing time can manifest in our initial x-velocity dependent measure.

Second, we looked at whether their was a difference in overall performance in the forced-reading Stroop task compared to standard Stroop task, as the forced task had a greater task demand. We found that initial x-velocity increased more quickly as a function of MIT for the standard task than the forced task. This suggests that

the participant's decision process unfolded at a faster rate over time in the standard task compared to the forced task.

Finally, we wanted to know if the Stroop magnitude emerged with more processing time and if the effect grew in the forced-reading task more than the standard Stroop task. We found that the Stroop effect was not evident in neither the standard nor forced tasks prior to approximately 133ms of processing time. Yet, after 133ms the Stroop effect was only evident in the forced task and not the standard task. In the forced task, the Stroop effect continued to grow in magnitude after 133ms. The lack of effect in the standard task suggests the standard Stroop effect is only a proportion of the benchmark forced-reading Stroop effect. Crucially, this finding does not depend on the amount of processing time – although some processing time, namely 133ms, is needed for significant differences between the standard and forced-reading Stroop task to emerge.

6.4.1 Validating Findings from Delta Plots and Forced-reading

Pratte et al. (2010) advocated the delta plot as a method for examining the time course of experimental effects, such as the Stroop effect. In their application of the delta function they found that the Stroop effect was small for fast responses and large for slow responses. Their finding suggested that the effect grows in magnitude as processing time increased. But, the slope of the delta plot is sensitive to the variance of the distributions in question, limiting its applicability. We showed that when the Stroop effect is observed, it grows in magnitude as processing time increases, even when assessed without the confounds of delta plots.

However, a significant Stroop effect only emerged in the forced Stroop task. The lack of a Stroop effect in the standard task is not a surprising result. Despite the reputation of the Stroop effect as a robust phenomenon, it has been shown to depend on design as well as other contextual factors. The effect appears only when certain conditions are met, but can be very small and even reversed given particular contextual factors (e.g., Kahneman & Chajczyk, 1983; MacLeod & Dunbar, 1988; Dishon-Berkovits & Algom, 2000; Besner et al., 1997; La Heij & Vermeij, 1987). In his comprehensive review of Stroop research, MacLeod (1991) listed set-size, mode of response, and relative speed of processing (among other factors) as factors that determine the magnitude of the effect. Since MacLeod, a substantial number of empirical papers have shown the malleable nature of the Stroop effect and how, with small set size and manual responses, it can be quite small and even vanish (see, e.g., Melara & Mounts, 1993; Dishon-Berkovits & Algom, 2000; Sabri, Melara, & Algom, 2001; Melara & Algom, 2003).

Our experimental design was limited to only two colors and to a manual (rather than than vocal) mode of response, both known to limit the magnitude of the Stroop effect (see also Eidels, Townsend, & Algom, 2010). Nonetheless, a marked Stroop effect was registered in the forced-reading task of the current study, suggesting that the effect can emerge even with two colors and a manual mode of responding. Its absence in the standard task does not merely reflect sensitivity to set size or to the mode of responding, but rather suggests that words in the standard Stroop task may not be fully processed, at least not to the same extent they are processed in the forced task.

The asymmetry in Stroop effects across the standard and forced tasks could

potentially be explained by the complexity of the forced-reading task. Specifically, Eidels et al. (2014) documented longer response times in the forced task, with the additional time allowing for the irrelevant word to interfere with color naming more (e.g., Melara & Algom, 2003).

The present study offers another way to expand on the findings of Eidels et al. (2014) by providing the means to directly examine the magnitude of the Stroop effect at the same points in processing time across the two tasks. Participants in the present study initiated their reaching responses in synchrony with an imperative go signal, as opposed to the target stimulus. Thus, we were able to equate the movement initiation times across the two tasks, despite the differences in task difficulty/complexity. When we compared the magnitude of the Stroop effect across tasks at similar points in stimulus-processing time, we observe a clear Stroop effect in the forced-reading version of the task at all time points greater than 133ms. In contrast, the magnitude of the effect is reduced at the corresponding time points in the standard version of the task. Expanding on Eidels et al. (2014) we show that the larger Stroop effect under forced-reading instructions is not an artifact due to longer processing time, but a genuine effect.

6.4.2 Theoretical Implications

A central result of the current study is the larger difference observed between the incongruent and congruent conditions (i.e., larger Stroop effect) at longer movement initiation times (see Figure 6.7) in the forced reading task. Existing theories of the Stroop effect may differ in their predictions concerning the magnitude

of the effect as processing time increases. We briefly survey three popular models and discuss whether they can predict this observed result.

The horse race model of the Stroop effect (Palef & Olson, 1975) suggests that activation of the word and color information accumulates in parallel. Word and color information accumulate toward a response channel, where task irrelevant word information arrives first. Because the word channel finishes first our cognitive system needs to wait for a response activated by the slower color information, which manifests as Stroop interference. This model has been criticized as it cannot account for data where the word information is delayed (e.g., Glaser & Glaser, 1982). In regards to our study, the horse race account cannot accommodate a Stroop effect that grows over time, which we observed in the forced reading task.

A current and popular account of the Stroop task is the parallel distributed processing model (Cohen et al., 1990). This model suggests that our system receives information (input) from different dimensions that travel down specific pathways to response mechanisms (output). Some of these pathways have stronger activation than others and the strength of this activation, not the speed, determines the output. In the Stroop task, the word pathway is considered stronger than the color pathway. Because word processing is more likely to reach the output node before color processing, additional activation needs to be recruited from task-specific nodes, which cause the system to run for many more processing cycles.[5] This account is in line with our results as longer processing times produce greater Stroop interference.

[5]See Botvinick, Braver, Barch, Carter, and Cohen (2001), who expanded the parallel distributed processing model to explain how our cognitive system monitors and regulates conflict.

Similarly, our results are in line with the tectonic theory of selective attention (Melara & Algom, 2003). In this model, evidence from target relevant information lead to the response required on the trial and values of the non-presented target lead to an incorrect response. A ratio of this evidence is calculated, and once the ratio reaches 1, a response is made. When there is more evidence for the non-presented target (i.e., when you have processed the word for longer) than the presented target, more processing steps are required to exceed the response threshold.

The fact that we found a Stroop effect in the forced task, but not the standard, sheds light on the nature of reading in the Stroop task. For instance, on any particular trial of the standard task, a participant might be processing the word to some extent or not reading the word at all. Eidels et al. (2014) posit a simple probability-mixture model to account for these results. Under this model, the empirical congruent and incongruent distributions we observe are binary mixtures of two unobserved distributions. A given trial is a sample drawn from the distribution associated with reading (with probability p) or the distribution free of word reading (with probability 1-p). The forced reading task increases the probability of reading to (p=1). This should lead to an inflated Stroop effect compared with the standard task, which is what we observe in our data.

6.4.3 Conclusions

Our study has methodological and theoretical implications. The arm reaching paradigm can potentially reveal how experimental effects emerge over time. We found that when the Stroop effect is observed, it grows in magnitude with more time

for processing – and this finding was demonstrated without the confounds of delta plots. We also showed that the nature of reading in the standard Stroop task is not comparable to a task in which we know the participant reads on every trial.

Chapter 7

General Conclusions

In this thesis, I examined perceptual decision-making from a sequential sampling model perspective. I addressed a number of theoretical issues related to cognitive load effects, speech perception, and lexical decision. This work demonstrates how sequential sampling models can account for both accuracy and response times (RT) from perceptual decision-making tasks, and by doing so, allow researchers to draw psychologically meaningful conclusions from behavioral data. In these studies I used Bayesian estimation to fit models to data. The Bayesian fitting routine allowed me to calculate posterior distributions of parameters, which contain information about what parameters are plausible and represent a natural measure of uncertainty. In each study I compared competing models by how well they predict future data. I show that predictive accuracy is one way that researchers can choose between competing theories instantiated as formal mathematical models. Despite the benefits of sequential sampling models, I demonstrated that these models can arrive at different conclusions when applied to the same data, and therefore, I advocate careful application of these models. I also demonstrated how researchers can use motion tracking technology and arm-reaching movements in perceptual decision-making tasks as a method for observing cognitive processes unfold online. For the rest of this chapter I will summarize the main conclusions of this thesis and then outline future directions for sequential sampling models of perceptual decision-making.

7.1 Summary of Results

In chapter 2, I investigated cognitive load effects on drivers and passengers of a motor vehicle. Cognitive load from secondary tasks, such as talking on a cell

phone, is a source of distraction, which is a significant cause of injuries and fatalities on the roadway. The Detection Response Task (DRT) is an international standard used to assess cognitive load on drivers' attention. I investigated whether decrements in DRT performance was due to changes to the speed of information processing, the response caution, or the non-decision processing of drivers and passengers. I had pairs of participants take part in the DRT while performing a simulated driving task. I manipulated cognitive load via the conversation between driver and passenger and observed associated slowing in DRT RT. Fits of single-bound diffusion model indicated that slowing was mediated by an increase in response caution. I proposed the novel hypothesis that, rather than the DRT's sensitivity to cognitive load being a direct result of a loss of information processing capacity to other tasks, it is an indirect result of a general tendency to be more cautious when making responses in more demanding situations.

In chapter 3, I used the Linear Ballistic Accumulator (LBA) to investigate how changes in acoustic cues affect latent cognitive processes that underpin phoneme decisions. In summary, I tested 30 Dutch listeners in a categorization experiment that required them to categorize speech sounds that varied in vowel quality (F1 and F2) and duration between typical /ɑ/ and /aː/. Using the LBA model, I found that the changes in spectral quality and duration cues lead to changes in the speed of information processing, and the effects were larger for spectral quality. For stimuli with atypical spectral information I found that listeners accumulate evidence faster for /ɑ/ compared to /aː/. Finally, longer durations of sounds did not produce longer estimates of perceptual encoding time.

In chapter 4, I applied the Diffusion Decision Model (DDM; Ratcliff, Gomez,

& McKoon, 2004) to the lexical-decision task. The lexical-decision task is among the most commonly used paradigms in psycholinguistics. In both the signal-detection theory and DDM frameworks, lexical-decisions are based on a continuous source of word-likeness evidence for both words and non-words. Previous applications of the DDM and studies using receiver operating characteristics assumed that evidence variability is equal across words and non-words. To test this assumption, I analyzed five lexical-decision data sets with the DDM. For all data sets, drift-rate variability changed across word frequency and non-word conditions. I also compared the results of the DDM analysis to the a-priori predictions of the REM–LD model of the lexical-decision task (Wagenmakers et al., 2004). There were some small discrepancies, but for the most part, we confirmed the predictions of REM–LD about the ordering of evidence variability across stimuli in the lexical-decision task.

In chapter 5, I compared the DDM and LBA and discussed conceptual differences between the two models that may lead to researchers drawing different psychological conclusions from the models. I argued that there are now many sequential sampling models, and although they share fundamental assumptions, they do not always draw the same psychological conclusions. I highlighted two key conceptual differences between two prominent sequential sampling models: the DDM and LBA. Firstly, the LBA can predict a speed up in mean drift rate for all accumulators, and the diffusion model compensates for this effect with changes in the diffusion coefficient (moment-to-moment drift rate fluctuation). Secondly, the DDM reliably estimates higher non-decision times than the LBA and the two models can disagree on whether non-decision time effects are present.

In chapter 6, I presented an arm-reaching method, which I used to investigate the nature of reading in the Stroop task. In a Stroop task participants are presented with a color name printed in color and need to classify the print color while ignoring the word. The Stroop effect is typically calculated as the difference in mean RT between congruent (e.g., the word RED printed in red) and incongruent (GREEN in red) trials. Arm-reaching trajectories allow for a more continuous measure for assessing the time course of the Stroop effect than RT. I compared arm movements to congruent and incongruent stimuli in a standard Stroop task and a control task that encourages processing of every word. The Stroop effect emerged over time in the control task, but not in the standard Stroop, suggesting words may be processed differently in the two tasks. Overall, chapter 6 demonstrated that the arm-movements of participants are a promising measure for investigating cognitive processes online.

7.2 Future Directions

For more than 50 years researchers have developed sequential sampling models of perceptual decision-making. The key feature of these models is their capacity to account for behavioral data from perceptual decision-making tasks while re-expressing the data as meaningful latent cognitive processes. With the advancement of modern neuroimaging technology, researchers have also been able to investigate perceptual decision-making by measuring brain activity and have been able to draw meaningful conclusions from the neural data.

Traditionally, mathematical modeling of cognition and neuroscience have been two separate and non-interacting methods of inquiry. Two key limitations to this independent approach is the lack of empirical grounding of mathematical models of cognitive processes to neural activity that may give rise to processes themselves. On the other hand, researchers find it difficult to make inferences about cognitive processes from the neural data alone. Because of these limitations, mathematical modeling and neuroscience have begun to converge on an interdisciplinary field known as model-based cognitive neuroscience (Forstmann & Wagenmakers, 2015). This field attempts to offer a detailed explanation of human behavior by combining our understanding of cognitive processes through modeling and our understanding of neural processes through brain measurements.

One aim of model-based cognitive neuroscience is to determine if the neural-activity of brains corresponds to processes posited by cognitive models. For example, do drift rates in sequential sampling models correspond to firing patterns in certain neuron populations? However, the links between cognitive processes and brain activity is not straight forward, and these links fall along a continuum, where multiple theoretical levels of linking exist (de Hollander, Forstmann, & Brown, 2015). At a qualitative level, the neural dynamics uncovered by brain imaging studies has inspired researchers to develop neurally plausible sequential sampling models (Usher & McClelland, 2001; Verdonck & Tuerlinckx, 2014). Neural data have also been used by researchers to investigate what decision-making processes may look like at the level of neuron populations (Gold & Shadlen, 2001, 2002, 2007) and to choose between cognitive models that cannot be discriminated from behavioral data alone, where the neural data serves as a qualitative constraint on the cognitive model

(Ditterich, 2010).

At a quantitative level, researchers have found that adjusting response thresholds in sequential sampling models correlates with activation in brain regions associated with caution (Forstmann et al., 2008, 2010). Advancements in Bayesian methods have allowed researchers to build complex sequential sampling models that are statistically constrained by both the behavioral and neural data (Turner, Forstmann, et al., 2013; Turner, van Maanen, & Forstmann, 2015; van Ravenzwaaij et al., 2016). Finally, researchers have directly mapped the firing rates of the frontal eye field visual and movement neurons onto the evidence accumulation processes of sequential sampling models (Purcell et al., 2010; Purcell, Schall, Logan, & Palmeri, 2012; Cassey et al., 2014). Such direct links are dependent on the assumption that the macro time scale of response times and sequential sampling models map onto the micro time scale of firing patterns of large networks of neuron populations, yet simulation studies have shown that explanations at the two different time scales are in agreement (Zandbelt, Purcell, Palmeri, Logan, & Schall, 2014).

So far, cognitive psychologists have made many exciting developments in both mathematical models and in neuroscience. This thesis demonstrated the utility of the former approach. I have briefly outlined the type of research being conducted in model-based cognitive neuroscience, which is a promising direction for mathematical models of perceptual decision-making. In future we should see the increased collaboration between scientists in mathematical psychology and neuroscience in an attempt to develop a complete explanation of human behavior.

Appendix A

Appendix: Chapter 2

A.0.1 Converging Evidence

Here we report fits from another model that can account for simple RT: the log-normal race model (Heathcote & Love, 2012). The model cannot tell drift rates from thresholds but can inform us whether there are rate/threshold and T_{er} effects. In Table A.1, we can see that a model with rate/threshold and non-decision time effects is preferred over a model with only one of the effects, which is in line with our single-bound diffusion model results.

Table A.1: WAIC results for the log-normal race models.

Model	Effective Parameters	WAIC
a/v $\sim$ F & $\eta \sim 1$ & $t_{er} \sim 1$	122.8	-18882.8
a/v ~ 1 & $\eta \sim 1$ & $t_{er} \sim$ F	207.4	-18631.0
a/v $\sim$ F & $\eta \sim 1$ & $t_{er} \sim$ F	189.1	**-19205.9**

Note. Bold WAIC value indicates the preferred model.

F = Cognitive Load Manipulation

A.0.2 Model Fitting Method

We estimated posterior distributions of parameter values using the Markov Chain Monte Carlo (MCMC) method (see van Ravenzwaaij et al., 2015, for a tutorial). To generate proposals for the MCMC algorithm we used differential evolution (DE-MCMC; Ter Braak, 2006). DE-MCMC has been shown to efficiently estimate parameters of hierarchical versions of models similar to the single-bound diffusion model (e.g., Turner et al., 2015; Turner, Sederberg, Brown, & Steyvers, 2013). For all model fits in the paper we ran the DE-MCMC algorithm with 40 chains. The starting points of these chains were drawn from the following distributions: $a \sim N(1, .1)$, $v \sim N(2, .2)$, $\eta \sim N(1, .1)$, and $T_{er} \sim N(.4, .04)$, where $N(m, sd)$ indicates a normal distribution with mean m and standard distribution sd.

Part of approximating posterior distributions via sampling is deciding when convergence has been obtained, at which we are confident that samples represent the posterior distribution. All samples prior to convergence are discarded. To decide the point of convergence we both visually inspected the chains and discarded all samples prior to the $\hat{R}$ statistic being less than 1.01 (Gelman & Rubin, 1992). Upon reaching the $\hat{R}$ criterion, we drew 5000 additional samples for each chain. To save memory during computing, and given the high auto-correlation within-chains, we thinned the posterior by only keeping every 10^{th} iteration. These 20000 (i.e. 40 chains $\times$ 500 iterations) samples constituted our posterior distribution estimates.

Each parameter for each subject was stochastically dependent on a group level distribution, ϕ_θ, where the subscript θ denotes the subject level parameter. We assumed that each group level distribution ϕ_θ had a truncated normal distribution,

where $\phi_\theta \sim N(\mu, \sigma) \mid (0, \infty)$ (where the numbers after the | indicate the distribution range). We set priors on the group level parameters where the mean of ϕ_θ had a truncated Normal prior $\sim N(2, 2) \mid (0, \infty)$, and the standard deviation had a Gamma prior $\sim \Gamma(1.01, 1)$. For the T_{er} parameter, the group level mean had prior $\sim N(.4, .4) \mid (0, 1)$ and the standard deviation had a Gamma prior $\sim \Gamma(1.01, 1)$. The subject level and group level parameters were estimated simultaneously. We also conducted a parameter recovery exercise (Heathcote, Brown, & Wagenmakers, 2015), which is presented in the following section, that demonstrates that our best fitting model's parameters can be recovered.

A.0.3 Parameter Recovery

For each participant in each data set, we repeatedly generated new data by simulating the single-bound diffusion with a sample from each participant's joint posterior distribution. We fit all of the simulated data sets with the single-bound diffusion using the same model parameterization. Presented in Figure A.1 are scatterplots of generating and recovered parameter values.

A.0.4 Model Fit

To assess how well the 10 models account for empirical trends in the data, we simulated data from each model. We sampled a set of parameter values from the joint subject-level posterior distributions for each participant. With these parameters we generated a data set the same size as the empirical data. This process was repeated 100 times, resulting in 100 simulated data sets. Figure A.2 summarizes the fit of the

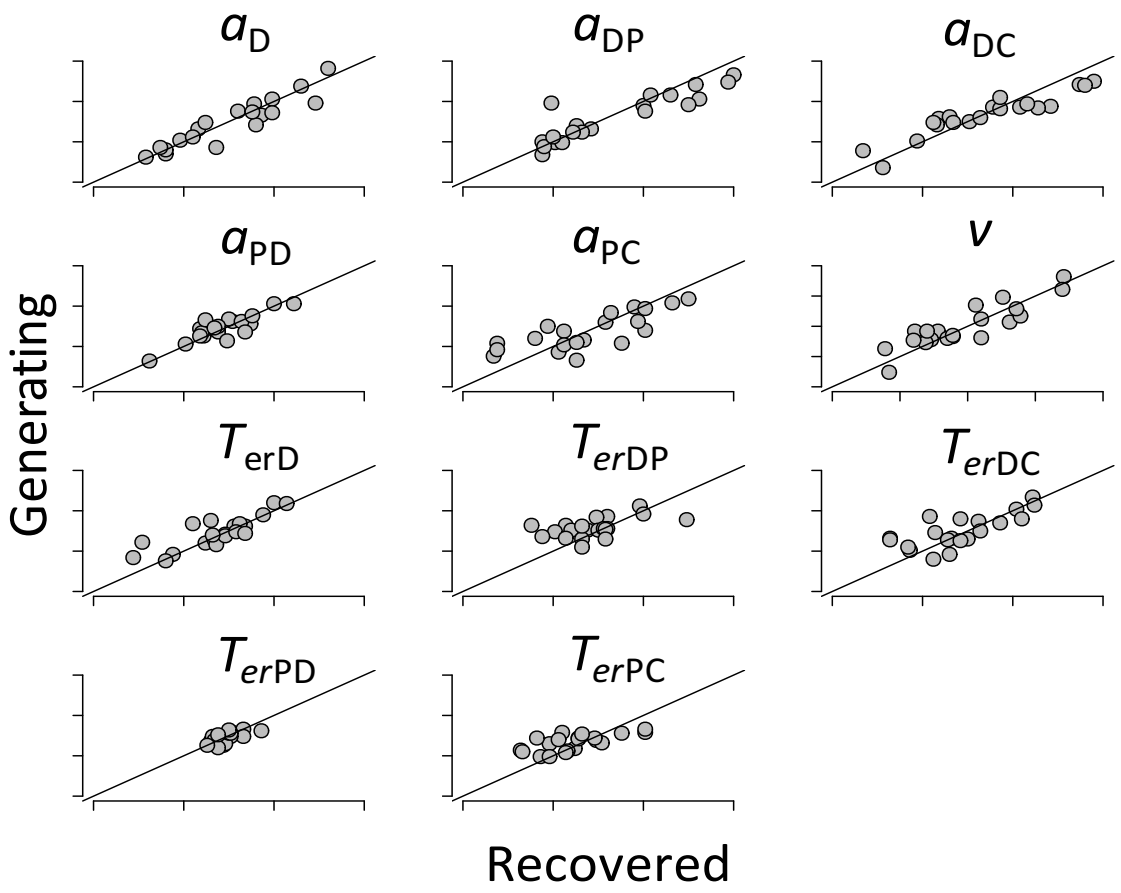


FIGURE A.1: Mean of subject-level posteriors of generating parameters values plotted as a function of the mean of subject-level posteriors of recovered parameter values.

10 models by superimposing each of the 100 simulated data sets on the empirical data. RT are summarized by their 10%, 50%, and 90% quantile. From visual inspection, all models appear to fit the data well.

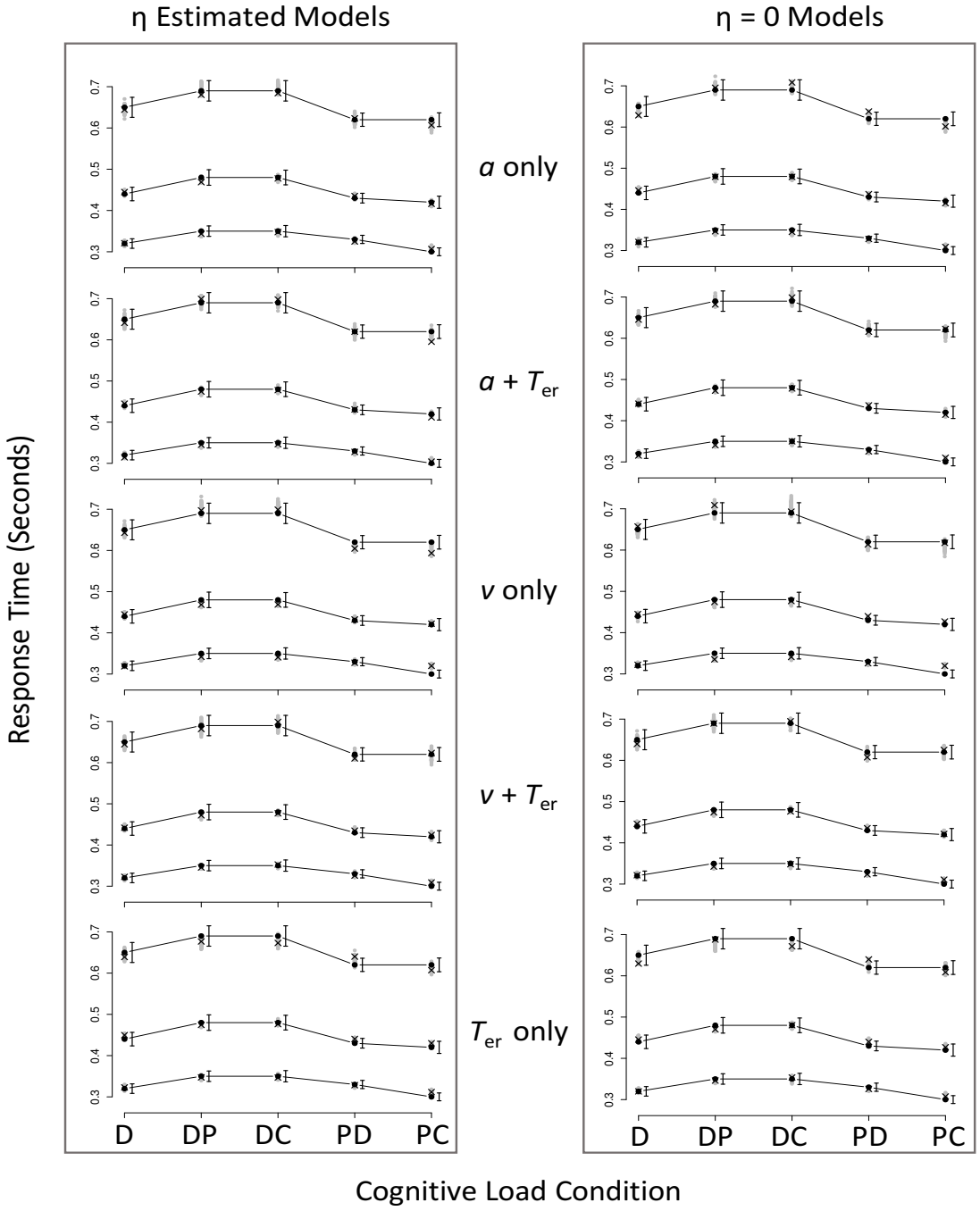


FIGURE A.2: Empirical and predicted .1, .5, and .9 quantiles for the response time (RT) distributions across five conditions. Predicted data is generated from the a model, $a + T_{er}$ model, v model, $v + T_{er}$ model, and T_{er} model – all with (left panels) and without (right panels) between trial drift variability. The black dots are the empirical data and the black error bars represent 1 standard error of the mean. The black ×s are the predicted RTs generated from the median group-level mean posterior distributions. The grey ×s are generated from posterior samples from the group-level mean posterior distributions.

A.0.5 RT Hazard Functions

The behavior of the single-bound diffusion model with negative rates is illustrated by its hazard function, a plot of the probability that a response will occur in the next unit of time, at each time point, given that a response has not yet occurred. Formally, hazard functions are defined as $h(t) = f(t)/[1 - F(t)]$, where $f(t)$ and $F(t)$ are the probability density function and cumulative density function, respectively. In the diffusion model with a single positive rate the hazard function increases to a plateau at longer times. A mixture or rates causes the hazard function to increase then decrease for longer times. The tail of the hazard function remains above zero when rates are strictly positive, but it decreases all the way to zero when the mixture contains negative rates, consistent with failures to respond. In both cases, including a mixture of rates makes the right tail of the RT distribution longer (i.e., it increases the proportion of slow responses).

Ratcliff and Strayer (2014) suggested that drift variability is needed to account for hazard functions that increase initially but then fall to a low asymptotic. In order to derive analytic likelihoods for our model we truncated the trial-to-trial drift rate distribution to positive values (Desmond & Yang, 2011). Figure A.3 shows that this truncation still allows our model to predict hazard functions with an initial increase than a fall towards the tail of the distribution.

The plots show no evidence of the pronounced dip in the right tail of the empirical hazard functions that is associated with substantial trial-to-trial rate variability, further confirming our selection of the simpler model with $\eta = 0$. The sharp

increase in the tail of the observed hazard function is likely due to the difficulty in approximating hazard functions from sparse data.

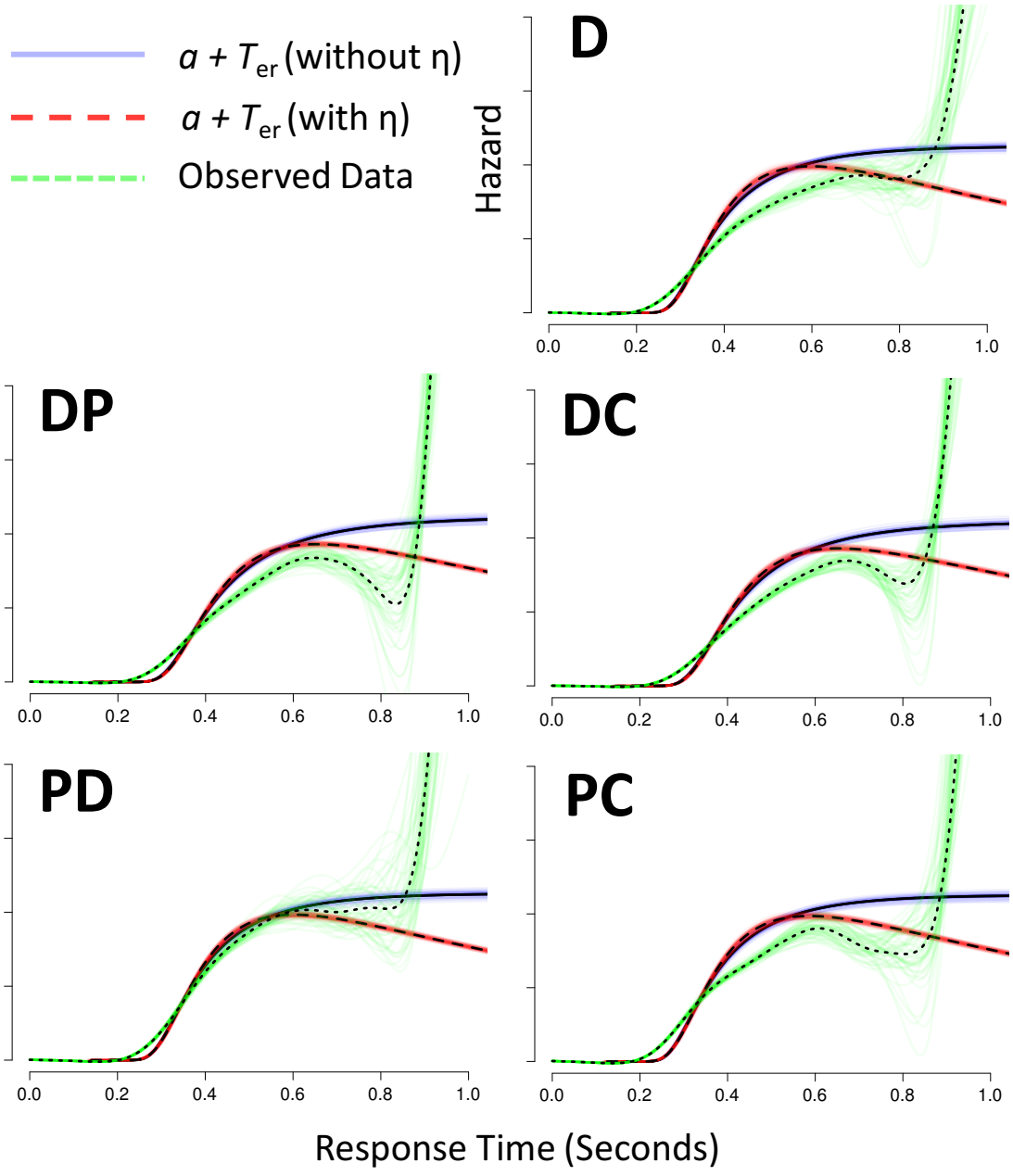


FIGURE A.3: Hazard functions for $a+T_{er}$ model with and without drift variability and for the empirical data. Red and blue lines are hazard functions generated from additional samples from the joint subject level posterior of the fitted models. Green lines are hazard functions generated from boot-strapped samples of the observed data.

Appendix B

Appendix: Chapter 3

B.0.1 Model Fitting Method

We use Bayesian methods to fit a hierarchical version of the LBA model (Kruschke, 2011; Gelman, Carlin, Stern, & Rubin, 1995; Lee & Wagenmakers, 2013). Two different statistical distributions were used for priors of the group level parameters. The first distribution was a truncated normal distribution, $N(\mu, \sigma) \mid (0, Inf)$, with mean parameter μ, standard deviation parameter σ, lower bound and upper bound. The second distribution we used was a gamma distribution, $\Gamma(\alpha, \beta)$, with shape parameter α and rate parameter β. We assumed that parameters (θ) for each subject came from the group level distribution ϕ, where $\phi \sim N(\mu, \sigma) \mid (0, Inf)$. We set priors on the group level parameters for subject level parameters θ, so that $\theta\mu \sim N(2, 2) \mid (0, Inf)$, and $\theta\sigma \sim \Gamma(1.01, 1)$. For the β parameters, the prior on the group level mean was $\sim N(0, .2) \mid (-Inf, Inf)$ and the prior on the group level SD was $\sim \Gamma(1.01, 1)$.

For hierarchical models, we cannot derive posteriors analytically, therefore we approximated the posterior distributions via sampling. We used the differential evolution Markov Chain Monte Carlo algorithm (DE-MCMC; Ter Braak, 2006; Turner, Sederberg, et al., 2013). For a tutorial on Markov Chain Monte Carlo see van Ravenzwaaij et al. (2015). For all model fits in the paper we ran the DE-MCMC algorithm with 48 chains in parallel. The first 2000 samples were considered part of a burn-in period and were discarded. For the last 1000 samples of the burn-in period, we ran with a migration algorithm to deal with any stuck chains (see Turner, Sederberg, et al., 2013). After the burn-in period, 7500 additional samples were run for each chain. This sample was considered enough to compensate for the autocorrelation present within each chain. Thus, only every 15th sample was kept. This resulted in 500 samples for each chain leaving 24000 (i.e. 48 chains $\times$ 500 iterations) samples, which constituted our posterior distribution.

Starting points for the Markov chains were drawn from the following distributions: both $v \sim N(2, .2) \mid (0, Inf)$, all $\beta \sim N(.5, .05) \mid (0, Inf)$, $A \sim N(1, .1) \mid (0, Inf)$, $b \sim N(1, .1) \mid (0, Inf)$, $s \sim N(1, .1) \mid (0, Inf)$, $t_0 \sim N(3, .03) \mid (0, Inf)$. The tuning parameters of the differential evolution proposal algorithm were set to the values used in (Turner, Sederberg, et al., 2013). The MCMC chains generated proposals separately for each participant's parameters and also blocked the group-level parameters in μ and σ pairs.

To assess convergence we visually inspected the chains. Figure B.1 demonstrates an example of a single participant in which we deemed converged. To assess convergence more objectively we also calculated $\hat{R}$, which quantifies how much the dispersion of the posteriors may reduce if we continue sampling (Gelman et al., 1995,

p. 285). $\hat{R}$ values closer to 1 indicate better convergence. $\hat{R}$ was less than 1.1 for all hyper parameter samples and the largest mean $\hat{R}$ for any one participant was 1.014, which suggests that chains were converged. The minimum effective number of independent draws ($\hat{\eta}_{\text{eff}}$) for hyper parameters, calculate as described by Gelman et al. (1995, p. 286), was $\hat{\eta}_{\text{eff}} = 19279.30$ (of 24000), suggesting that we have estimated our posterior distributions with high precision.

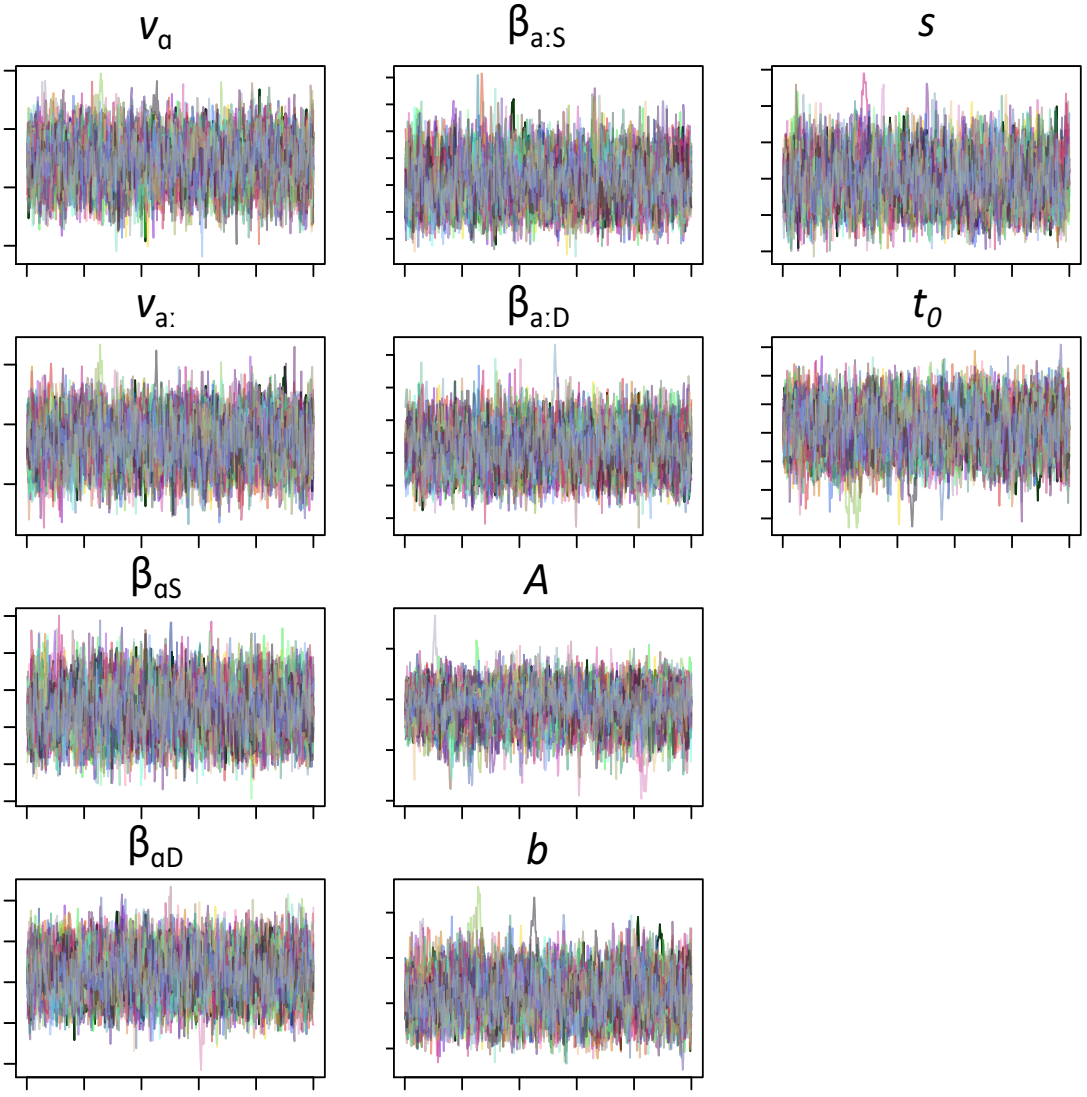


FIGURE B.1: Trace plot of Markov chains for an individual subject. These chains are considered converged with regards to visual inspection. Note that the trace plot here contains 48 Markov chains, which have been run in parallel.

B.0.2 Model Fit Results

To assess how well our model captures the empirical trends in the data, we simulated data from our unequal drift LBA model, which was the best fitting model. This process involved sampling a set of parameter values from the subject-level posterior distributions for each participant. With these parameters we generated a data set the same size as the empirical data. This process was repeated 100 times, resulting in 100 simulated data sets. Figure B.2 plots the empirical data and superimpose all 100 of the simulated data sets.

The left panel of Figure B.2 shows the RT distributions for both the /ɑ/ and /aː/ responses across the 10 duration values. To summarize the RT distributions, we present five quantile estimates (10%, 30%, 50%, 70% and 90%). The 10% and 90% quantiles represent the leading edge and tail of the distribution, respectively. The 50%, or median, is a measure of central tendency. These plots are defective cumulative distributions, meaning that the relative heights of the /ɑ/ (red points) and /aː/ (blue points) distributions show the proportion of /ɑ/ and /aː/ responses, respectively. The colored circles show the observed quantile estimates and the colored crosses show the LBA's predicted quantile estimates. Across all 10 duration values we can see that the observed data and the predicted data are mostly in agreement. However, for the 1^{st} duration step for /aː/ responses and for the 10^{th} duration step for /ɑ/ responses, the model misses the 90% quantile. Specifically, the tail of the empirical data extends out further than the model predicts. Note that when responses are rare (as is for 1^{st} duration /aː/ responses and 10^{th} duration /ɑ/ responses), the high quantiles are often not captured well by response time models.

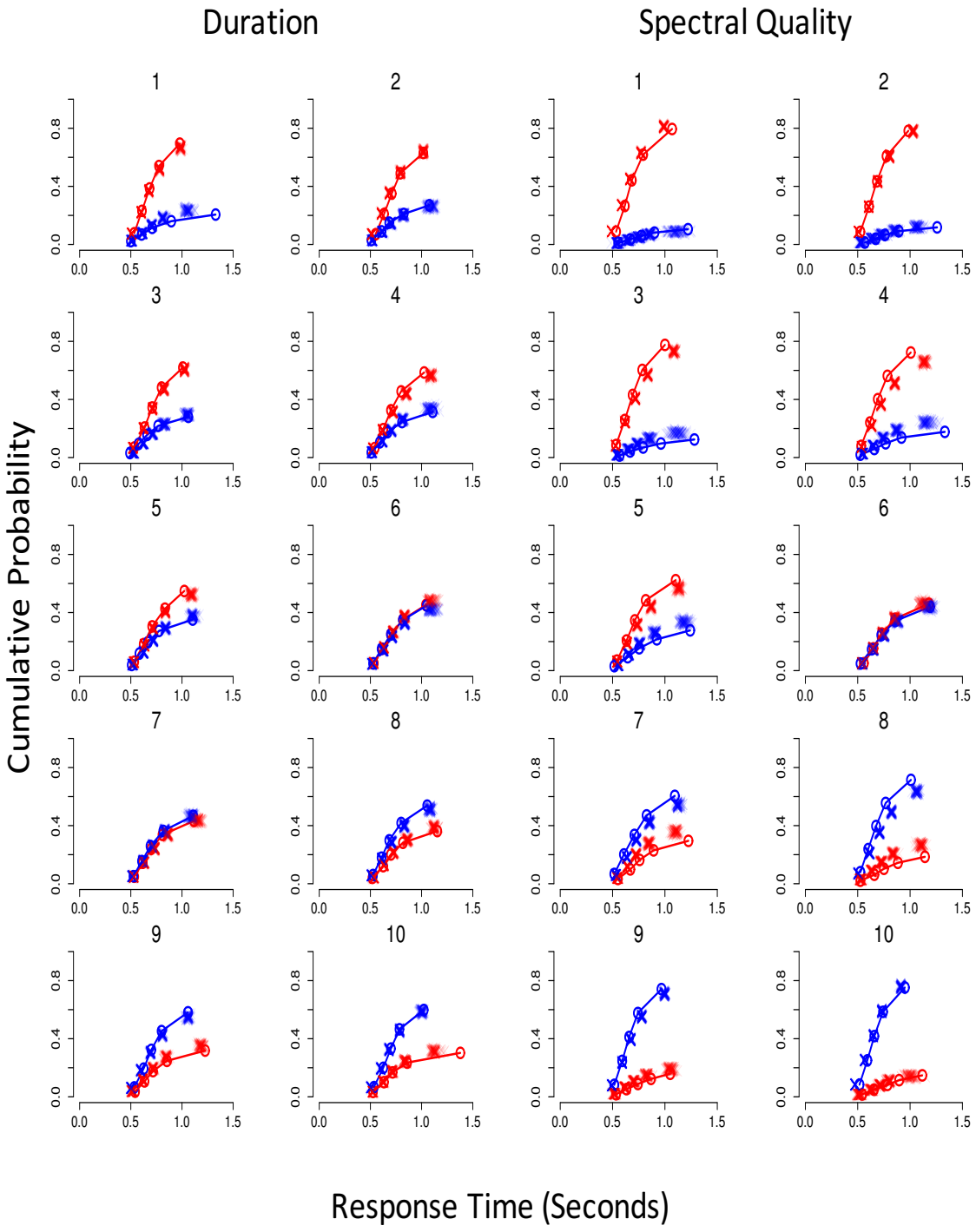


FIGURE B.2: Predicted and empirical RTs for each duration step (left panel) and each spectral quality step (right panel). The circles represent the empirical data and the crosses represent the posterior predicted data. The red points are /ɑ/ responses and the blue points are /aː/ responses. Numbers 1 though to 10 at the top of each plot represent the condition level as presented in Table 1 in the original paper.

The right panel of Figure B.2 shows RT distributions for both the /ɑ/ and /aː/ responses across the 10 spectral qualities. Again, the observed data and the predicted data are mostly in agreement. The LBA captures response times well, but produces some misses to the response proportions for the 3^{rd}, 4^{th}, 5^{th}, 7^{th} and 8^{th} spectral qualities. Overall, the model fits well given that only 10 parameters are estimated to capture effects of categorization and RTs in 200 conditions.

B.0.3 Parameter recovery check

For each participant in each data set, we repeatedly generated new data by simulating the LBA with a sample from each participant's joint posterior distribution. We fit all of the simulated data sets with the LBA using the same model parameterization. Presented in Figure B.3 is the deviation between the posterior distributions used to generate simulated data and the posterior distributions recovered in the fits. All of the histograms are centered on zero (red vertical line) suggesting that we have recovered our model.

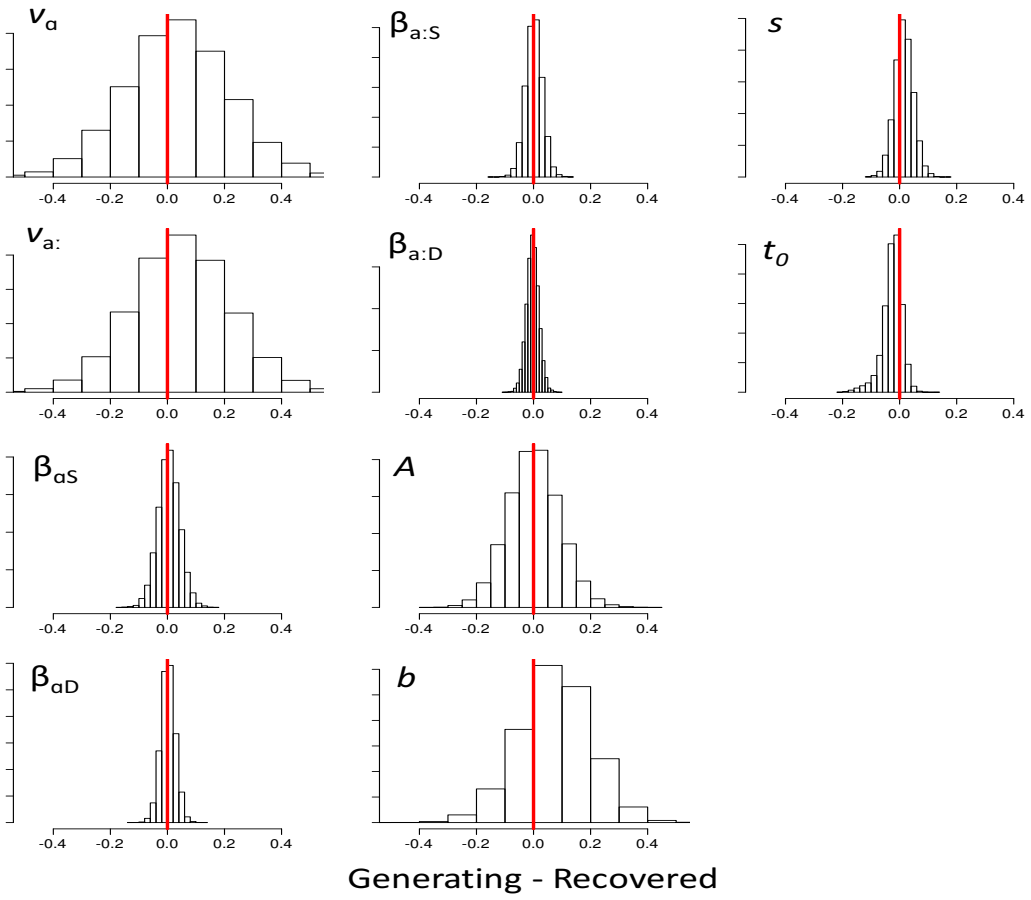


FIGURE B.3: Deviation between group-level mean posterior values used to generate simulated data and the recovered posterior values in fits. The vertical red lines show that all histograms are centered on zero suggesting a good parameter recovery and little estimation bias.

Appendix C

Appendix: Chapter 4

C.0.1 REM–LD Predictions

To derive predictions from REM–LD we use equations from Wagenmakers et al. (2004). The probe and memory traces all consisted of 30 features. We also assume that the probe is compared to 100 lexical traces in memory. The probability of a feature match when encoding the same item was .85 for high frequency words, .75 for low frequency words, and .65 for low frequency words. The probability of a feature match when encoding a different item was .5 for pseudo-words and .35 for random letter strings. The rate of increase in probability of activation was .0025.

We calculated the log posterior odds ratio for 1000 trials at four different deadlines (250ms, 500ms, 750ms, 1000ms). The deadline was when we told the system to respond and it reflects the nature of evidence at different response times. The system had a minimum processing time of 250ms, during which no comparisons between the probe and trace are made.

C.0.2 Model Parameterization

The diffusion model parameterization for each data set is shown in Table C.1. The parameterization are for the unequal model, which had a separate η for each lexical stimulus. The equal model had the same parameterization as the unequal model, but only included one η for all word frequency conditions.

TABLE C.1: Data sets that were fit with the diffusion model.

Data Set	Source	Parameterization
1	Wagenmakers et al. (2008) Exp1.	a ~ E & v ~ W*E & z ~ E & sz ~ E & η ~ W & t_{er} ~ E & st ~ E
2	Wagenmakers et al. (2008) Exp.2	a ~ 1 & v ~ W & z ~ P & sz ~ 1 & η ~ W & t_{er} ~ W & st ~ 1
3	Ratcliff et al. (2004) Exp.1	a ~ 1 & v ~ W & z ~ P & sz ~ 1 & η ~ W & t_{er} ~ W & st ~ 1
4	Ratcliff et al. (2004) Exp.2	a ~ 1 & v ~ W & z ~ P & sz ~ 1 & η ~ W & t_{er} ~ W & st ~ 1
5	Ratcliff et al. (2004) Exp.4	a ~ 1 & v ~ W & z ~ P & sz ~ 1 & η ~ W & t_{er} ~ W & st ~ 1

E = Speed or Accuracy Emphasis

P = Word/Non-Word Proportion Manipulation

W = Word Frequency Manipulation

C.0.3 Model Fitting Method

We use Bayesian methods to fit an hierarchical version of the diffusion model (Kruschke, 2011; Gelman et al., 1995; Lee & Wagenmakers, 2013). Two different statistical distributions were used for priors of the group level parameters. The first distribution was a truncated normal distribution, $N(\mu, \sigma) \mid (lower, upper)$, with mean parameter μ, standard deviation parameter σ, lower bound and upper bound. The second distribution we used was a gamma distribution, $\Gamma(\alpha, \beta)$, with shape parameter α and rate parameter β. We assumed that parameters (θ) for each subject came from the group level distribution ϕ, where $\phi \sim N(\mu, \sigma) \mid (0, Inf)$. We set the following priors on the group level parameters: σ for all diffusion parameters was distributed as $\sim \Gamma(1.01, 1)$; $a_\mu \sim N(2, 2) \mid (0, 10)$; $v_\mu \sim N(2, 2) \mid (-10, 10)$;

$z_\mu \sim N(.5, .5) \mid (0, 1)$; $sz_\mu \sim N(.1, .1) \mid (0, 1)$; $T_{er\mu} \sim N(.3, .3) \mid (0, 1)$; $st_\mu \sim N(.1, .1) \mid (0, 1)$; $\eta_\mu \sim N(1, 1) \mid (0, 10)$.

For hierarchical models, we cannot derive posteriors analytically, therefore we approximated the posterior distributions via sampling. We used the differential evolution Markov Chain Monte Carlo algorithm (DE-MCMC; Ter Braak, 2006; Turner, Sederberg, et al., 2013). For a tutorial on Markov Chain Monte Carlo see van Ravenzwaaij et al. (2015). For all model fits in the paper we ran the DE-MCMC algorithm with 48 chains in parallel. The first 2000 samples were considered part of a burn-in period and were discarded. For these first 2000 samples we ran with a migration algorithm to deal with any stuck chains (see Turner, Sederberg, et al., 2013). After the burn-in period, 5000 additional samples were run for each chain. This sample was considered enough to compensate for the autocorrelation present within each chain. Thus, only every 10th sample was kept. This resulted in 500 samples for each chain leaving 24000 (i.e. 48 chains $\times$ 500 iterations) samples, which constituted our posterior distribution. Starting points for the Markov chains were drawn from the group level prior distributions. The tuning parameters of the differential evolution proposal algorithm were set to the values used in (Turner, Sederberg, et al., 2013). The MCMC chains generated proposals separately for each participant's parameters and also blocked the group-level parameters in μ and σ pairs.

To assess convergence we visually inspected the chains and also calculated $\hat{R}$, which quantifies how much the dispersion of the posteriors may reduce if we continue sampling (Gelman et al., 1995, p. 285). $\hat{R}$ values closer to 1 indicate better convergence. $\hat{R}$ was less than 1.1 for all hyper parameter samples, which suggests that chains were converged.

C.0.4 Model Recovery

We ensure that the trends in our parameter estimates are not the product of estimation bias (see Starns & Ratcliff, 2014). To do this we simulate data from the diffusion model with only 1 η parameter and fit this data with the model with separate η parameters for each word frequency condition.

The η parameter of the diffusion model is estimated from subtle effects in the empirical data. Therefore, it is important to check whether any systematic changes we observe in η are not due to our fitting routine – i.e., estimation bias. To check for estimation bias we generate a new data set for each participant. To do this we generating data from a set of parameters sampled from their corresponding posterior distribution of the equal variance diffusion model. Each of these data sets are then fit using the unequal diffusion model. Because the data are generated from the equal variance model, any systematic differences in the η parameters for the unequal model fit to this data is due to estimation bias.

In Figure C.1 we display the actual generating η values minus the recovered η values. The values are the group level mean posterior estimates of η. Distributions centered at zero indicate an accurate recovery and suggests that there is no estimation bias. There was an accurate recovery for all data sets, which is indicated by the 95% credible intervals of the posteriors overlapping with zero.

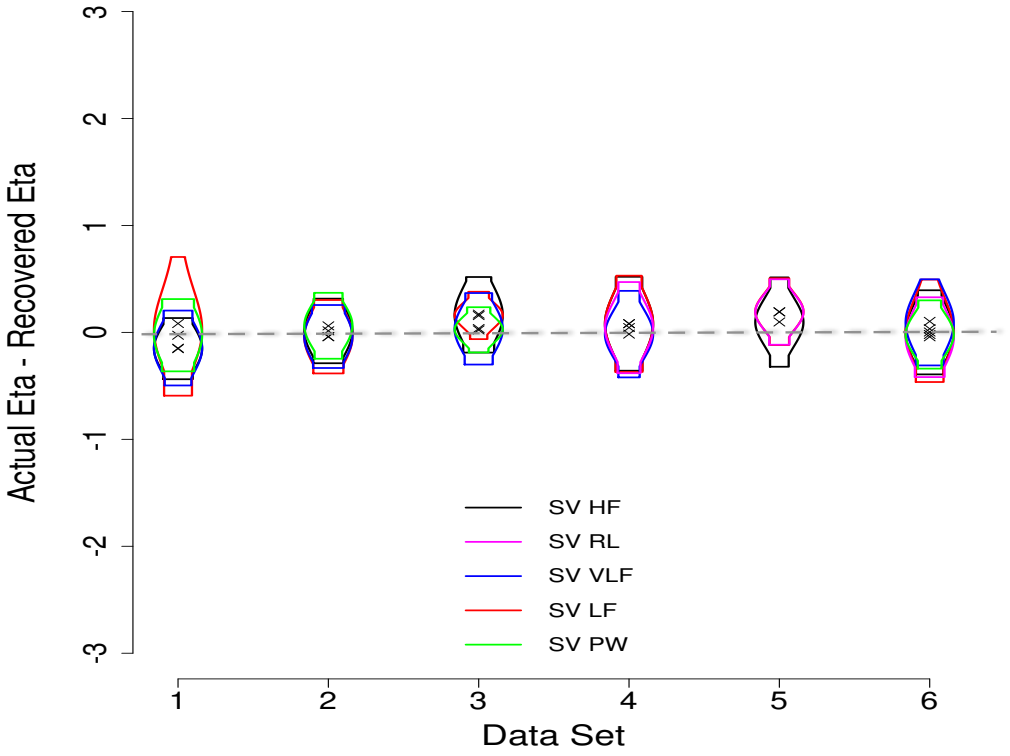


FIGURE C.1: Deviation between the η values used to generate simulated data and the η values recovered in fits. Results for high frequency (HF), low frequency (LF), very low frequency (VLF), and pseudo-words/random letter strings (NW) from each data set (DS) are displayed. The vertical lines mark the average deviation across the simulated data sets, so a vertical line at zero indicates no estimation bias. The y-axis is the probability density and so the histogram has a total area of one.

References

Adank, P., Van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of northern and southern standard dutch. *The Journal of the Acoustical society of America*, *116*(3), 1729–1738.

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

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, *38*(4), 419–439.

Amano, K., Goda, N., Nishida, S., Ejima, Y., Takeda, T., & Ohtani, Y. (2006). Estimation of the timing of human visual perception from magnetoencephalography. *The Journal of neuroscience*, *26*(15), 3981–3991.

Anders, R., Alario, F., & van Maanen, L. (2016). The shifted wald distribution for response time data analysis. *Psychological Methods*.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3), 144–151.

Ball, K., & Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science*, *218*(4573), 697–698.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human perception and performance*, *10*(3), 340-357.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, *128*(1), 32-55.

Berger, J. O. (2006). Bayes factors. *Encyclopedia of statistical sciences*.

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

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9/10), 341–345.

Boersma, P. (2007). Some listener-oriented accounts of h-aspiré in french. *Lingua*, *117*(12), 1989–2054.

Bohn, O.-S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in l2 vowel perception. *Applied Psycholinguistics*, *11*(03), 303–328.

Bompas, A., & Sumner, P. (2011). Saccadic inhibition reveals the timing of automatic and voluntary signals in the human brain. *The Journal of neuroscience*, *31*(35), 12501–12512.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. (2001). Evaluating the demand for control: Anterior cingulate cortex and conflict

monitoring. *Psychological Review*, *108*, 624–652.

Box, G. E. P. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Oxford: John Wiley & Sons.

Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117-128.

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

Brown, S. D., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 587-599.

Brown, T. L. (2011, Feb). The relationship between stroop interference and facilitation effects: statistical artifacts, baselines, and a reassessment. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 85-99. doi: 10.1037/a0019252

Buetti, S., & Kerzel, D. (2010). Effects of saccades and response type on the simon effect: If you look at the stimulus, the simon effect may be gone. *The Quarterly Journal of Experimental Psychology*, *63*(11), 2172–2189.

Cassey, P., Heathcote, A., & Brown, S. D. (2014). Brain and behavior in decision-making. *PLoS Computational Biology*, *10*(7), 1-8.

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.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC:

a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, *108*(1), 204-256.

Cook, E. P., & Maunsell, J. H. (2002). Dynamics of neuronal responses in macaque mt and vip during motion detection. *Nature neuroscience*, *5*(10), 985–994.

Curran, T., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 531-547.

de Hollander, G., Forstmann, B. U., & Brown, S. D. (2015). Different ways of linking behavioral and neural data via computational cognitive models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

De Jong, R., Liang, C.-C., & Lauber, E. (1994). Conditional and unconditional automaticity: a dual-process model of effects of spatial stimulus-response correspondence. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 731-750.

Desmond, A., & Yang, Z. (2011). Score tests for inverse gaussian mixtures. *Applied Stochastic Models in Business and Industry*, *27*(6), 633–648.

Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S., Sudweeks, J., ... Gupta, S. (2006). *The 100-car naturalistic driving study, phase ii-results of the 100-car field experiment* (Tech. Rep.). National Highway and Traffic Safety Administration.

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.

Ditterich, J. (2010). A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Frontiers in neuroscience*, *4*, 184.

Dittrich, K., Kellen, D., & Stahl, C. (2014). Analyzing distributional properties of interference effects across modalities: chances and challenges. *Psychological research*, *78*(3), 387–399.

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*(6), 1129–1135.

Donkin, C., Brown, S. D., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review*, *18*(1), 61–69.

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, *14*(4), 392 - 401.

Dutilh, G., Krypotos, A.-M., & Wagenmakers, E.-J. (2011). Task-related versus stimulus-specific practice. *Experimental Psychology*, *58*(6), 434-442.

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics*, *74*(2), 454–465.

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin &*

Review, *16*(6), 1026–1036.

Eidels, A. (2012). Independent race of colour and word can predict the stroop effect. *Australian Journal of Psychology*, *64*(4), 189–198.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review*, *17*(6), 763–771.

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.

Eidels, A., Townsend, J. T., & Algom, D. (2010). Comparing perception of stroop stimuli in focused versus divided attention paradigms: Evidence for dramatic processing differences. *Cognition*, *114*(2), 129–150.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and l2 perceptual cue weighting for dutch vowels: The case of dutch, german, and spanish listeners. *Journal of Phonetics*, *37*(4), 452–465.

Finkbeiner, M., Coltheart, M., & Coltheart, V. (2014). Pointing the way to new constraints on the dynamical claims of computational models. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 172-85.

Finkbeiner, M., & Heathcote, A. (2016). Distinguishing the time-and magnitude-difference accounts of the simon effect: Evidence from the reach-to-touch paradigm. *Attention, Perception, & Psychophysics*, *78*(3), 848-867.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of english vowels. *Journal of phonetics*, *25*(4), 437–470.

Forstmann, B. U., Brown, S. D., Dutilh, G., Neumann, J., & Wagenmakers, E.-J. (2010). The neural substrate of prior information in perceptual decision making: a model-based analysis. *Frontiers in Human Neuroscience*, *4*(40), 1-12.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-sma facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542.

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: a structural model-based approach. *The Journal of Neuroscience*, *31*(47), 17242–17249.

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An introduction to model-based cognitive neuroscience*. New York, NY: Springer.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153–160.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis* (Vol. 2). London: Chapman and Hall.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-511.

Gerrits, E. (2001). *The categorisation of speech sounds by adults and children: a study of the categorical perception hypothesis and the development weighting of*

acoustic speech cues (Unpublished doctoral dissertation). Utrecht University.

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

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in cognitive sciences*, *5*(1), 10–16.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*(2), 299–308.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Goldfarb, S., Leonard, N. E., Simen, P., Caicedo-Núñez, C. H., & Holmes, P. (2014). A comparative study of drift diffusion and linear ballistic accumulator models in a reward maximization perceptual choice task. *Frontiers in neuroscience*, *8*(148).

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), 1515-1523.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*(3), 518-565.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index):

Results of empirical and theoretical research. *Advances in psychology*, *52*, 139–183.

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods*, *36*, 678–694.

Heathcote, A., Brown, S. D., & Mewhort, D. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, *9*(2), 394–401.

Heathcote, A., Brown, S. D., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review*, *7*, 185-207.

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience.* (pp. 25–48). New York, NY: Springer.

Heathcote, A., & Hayes, B. (2012). Diffusion versus linear ballistic accumulation: different models for response time with different conclusions about psychological mechanisms? *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *66*(2), 125-36.

Heathcote, A., Loft, S., & Remington, R. W. (2015). Slow down and remember to remember! a delay theory of prospective memory costs. *Psychological review*, *122*(2), 376-410.

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in psychology*, *3*, 292.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, *8*, 150.

Ho, T. C., Brown, S. D., & Serences, J. T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *The Journal of Neuroscience*, *29*(27), 8675–8687.

Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The piecewise linear ballistic accumulator model. *Cognitive psychology*, *85*, 1–29.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisitiona. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218–1227.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, *25*(1), 105–113.

International Organization for Standardization. (2015). *Road vehicles – transport information and control systems – detection-response task (drt) for assessing attentional effects of cognitive load in driving.*

Ito, M., Tsuchida, J., & Yano, M. (2001). On the effectiveness of whole spectral shape for vowel perception. *The Journal of the Acoustical Society of America*, *110*(2), 1141–1149.

Jacobs, A. M., Graf, R., & Kinder, A. (2003). Receiver operating characteristics in the lexical decision task: evidence for a simple signal-detection process

simulated by the multiple read-out model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 481-488.

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

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.

Ko, E.-S., Soderstrom, M., & Morgan, J. (2009). Development of perceptual sensitivity to extrinsic vowel duration in infants learning american english. *The Journal of the Acoustical Society of America*, *126*(5), EL134–EL139.

Kruschke, J. K. (2011). *Doing Bayesian analysis: A tutorial with R and BUGS*. Academic Press.

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.

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, *38*(1), 44–59.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. London: Academic Press.

Lee, M., & Wagenmakers, E. (2013). *Bayesian modeling for cognitive science: A practical course*. Cambridge: University Press.

Lehiste, I., & Lass, N. J. (1976). Suprasegmental features of speech. *Contemporary issues in experimental phonetics*, *225*, 239.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*, 77-105.

Lipski, S. C., Escudero, P., & Benders, T. (2012). Language experience modulates weighting of acoustic cues for vowel perception: An event-related potential study. *Psychophysiology*, *49*(5), 638–650.

Lisker, L. (1986). "voicing" in english: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and speech*, *29*(1), 3–11.

Loft, S., & Remington, R. W. (2013). Wait a second: Brief delays in responding reduce focality effects in event-based prospective memory. *The Quarterly Journal of Experimental Psychology*, *66*(7), 1432–1447.

Logan, G. D. (1980). Attention and automaticity in stroop and priming tasks: Theory and data. *Cognitive psychology*, *12*(4), 523–553.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization* (No. 8). Oxford: Oxford University Press.

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.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *Psychometrika*, *81*(2), 274–289.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-gaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental science*, *12*(3), 369–378.

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic bulletin & review*, *15*(6), 1064–1071.

McVay, J. C., & Kane, M. J. (2012). Drifting from slow to "d'oh!": Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 525-549.

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

Melara, R. D., & Mounts, J. R. (1993). Selective attention to stroop dimensions: Effects of baseline discriminability, response mode, and practice. *Memory & Cognition*, *21*(5), 627–645.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 1142–1160.

Miller, J. L. (2001). Mapping from acoustic signal to phonetic category: Internal category structure, context effects and speeded categorisation. *Language and Cognitive Processes*, *16*(5-6), 683–690.

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

Morrison, G. S. (2013). Theories of vowel inherent spectral change. In *Vowel inherent spectral change* (pp. 31–47). Berlin Heidelberg: Springer.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90–100.

Nooteboom, S. G., & Cohen, A. (1984). *Het proces van spreken en verstaan, een nieuwe inleidingin de experimentele fonetiek.* The Netherlands: Van Gorcum: Assen.

Nooteboom, S. G., & Doodeman, G. (1980). Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America*, *67*(1), 276–287.

Norris, D. (1986). Word recognition: Context effects without priming. *Cognition*, *22*(2), 93–136.

Norris, D. (2006). The bayesian reader: explaining word recognition as an optimal bayesian decision process. *Psychological review*, *113*(2), 327-357.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*(2), 260-311.

Osth, A. F., Dennis, S., & Heathcote, A. (in press). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*.

Palef, S. R., & Olson, D. R. (1975). Spatial and verbal rivalry in a stroop-like task. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *29*(3), 201-209.

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in english. *The Journal of the Acoustical Society of America*, *32*(6), 693–703.

Pierrehumbert, J. (2001). Why phonological constraints are so coarse-grained.

Language and Cognitive Processes, *16*(5-6), 691–698.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*(2), 285–290.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 421-435.

Pols, L. C., Tromp, H. R., & Plomp, R. (1973). Frequency analysis of dutch vowels from 50 male speakers. *The journal of the Acoustical Society of America*, *53*(4), 1093–1101.

Porcu, E., Bölling, L., Lappe, M., & Liepelt, R. (2016). Pointing out mechanisms underlying joint action. *Attention, Perception, & Psychophysics*, 1–6.

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010, Oct). Exploring the differences in distributional properties between stroop and simon effects using delta plots. *Attention, Perception, & Psychophysics*, *72*(7), 2013-25. doi: 10.3758/APP.72.7.2013

Proctor, R. W., & Shao, C. (2010). Does the contribution of stimulus-hand correspondence to the auditory simon effect increase with practice? *Experimental brain research*, *204*(1), 131–137.

Proctor, R. W., Vu, K.-P. L., & Nicoletti, R. (2003). Does right-left prevalence occur for the simon effect? *Perception & Psychophysics*, *65*(8), 1318–1329.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological review*, *117*(4), 1113-1143.

Purcell, B. A., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2012). From salience

to saccades: multiple-alternative gated stochastic accumulator model of visual search. *The Journal of Neuroscience*, *32*(10), 3433–3446.

Quek, G. L., & Finkbeiner, M. (2013). Spatial and temporal attention modulate the early stages of face processing: behavioural evidence from a reaching paradigm. *PloS one*, *8*(2), 57365–57365.

Quek, G. L., & Finkbeiner, M. (2014). Face-perception is superior in the upper visual field: Evidence from masked priming. *Visual Cognition*, *22*(8), 1038–1042.

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

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *40*(5), 1226-43.

Ranney, T. A., Mazzae, E., Garrott, R., & Goodman, M. J. (2000). Nhtsa driver distraction research: Past, present, and future. In *Driver distraction internet forum* (Vol. 2000).

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological review*, *92*(2), 212-225.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291.

Ratcliff, R. (2013). Parameter variability and distributional assumptions in the

diffusion model. *Psychological review*, *120*(1), 281-292.

Ratcliff, R. (2015). Modeling one-choice and two-choice driving tasks. *Attention, Perception, & Psychophysics*, *77*(6), 2134–2144.

Ratcliff, R., Gomez, P., & McKoon, G. M. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159-182.

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

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, *111*(2), 333-67.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological review*, *116*(1), 59-83.

Ratcliff, R., & Strayer, D. (2014). Modeling simple driving tasks with a one-boundary diffusion model. *Psychonomic Bulletin & Review*, *21*(3), 577–589.

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), 278-289.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and aging*, *16*(2), 323-341.

Ratcliff, R., Thapar, A., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & psychophysics*, *65*(4), 523–535.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the

effects of aging on recognition memory. *Journal of Memory and Language*, *50*(4), 408–424.

Ratcliff, R., & Van Dongen, H. P. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences*, *108*(27), 11285–11290.

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science*, *181*(4099), 574–576.

Reed, A. V. (1976). List length and the time course of recognition in immediate memory. *Memory & Cognition*, *4*(1), 16–30.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological bulletin*, *92*(1), 81 - 110.

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.

Rueda-Domingo, T., Lardelli-Claret, P., de Dios Luna-del Castillo, J., Jiménez-Moleón, J. J., Garcı??a-Martı??n, M., & Bueno-Cavanillas, A. (2004). The influence of passengers on the risk of the driver causing a car collision in spain: Analysis of collisions from 1990 to 1999. *Accident Analysis & Prevention*, *36*(3), 481–489.

Sabri, M., Melara, R. D., & Algom, D. (2001). A confluence of contexts: Asymmetric

versus global failures of selective attention to stroop dimensions. *Journal of experimental psychology: Human perception and performance*, *27*(3), 515-537.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403-428.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414-429.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods*, *33*(4), 457–469.

Sewell, D. K., Lilburn, S. D., & Smith, P. L. (2016). Object selection costs in visual working memory: A diffusion model analysis of the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Shahar, N., Teodorescu, A. R., Usher, M., Pereg, M., & Meiran, N. (2014). Selective influence of working memory load on exceptionally slow reaction times. *Journal of Experimental Psychology: General*, *143*(5), 1837-1860.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, *32*, 1248—1284.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Rem—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Smith, P. L. (1995). Psychophysically principled models of visual simple reaction time. *Psychological review*, *102*(3), 567-593.

Song, J.-H., & Nakayama, K. (2009, Aug). Hidden cognitive states revealed in choice reaching tasks. *Trends Cogn Sci*, *13*(8), 360-6. doi: 10.1016/j.tics.2009.04.009

Speckman, P., Rouder, J., Morey, R., & Pratte, M. (2008). Delta plots and coherent distribution ordering. *The American Statistician*, *62*, 262-266.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*, 583-639.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398.

Stan Development Team. (2016). R stan: the r interface to stan, version 2.10.1. [Computer software manual]. Retrieved from `http://mc-stan.org`

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory & cognition*, *42*(8), 1357–1372.

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of roc functions: A diffusion model analysis. *Journal of memory and language*, *70*, 36–52.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zroc slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*(1), 1–34.

Sternberg, S. (1969). The discovery of processing stages: Extensions of donders'

method. *Acta psychologica*, *30*, 276–315.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2013). Measuring cognitive distraction in the automobile. *AAA Foundation for Traffic Safety*.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of experimental psychology: Applied*, *9*(1), 23-32.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological science*, *12*(6), 462–466.

Strayer, D. L., Turrill, J., Coleman, J., Ortiz, E., & Cooper, J. M. (2014). Measuring cognitive distraction in the automobile: Ii. assessing in-vehicle voice-based interactive technologies. *AAA Foundation for Traffic Safety.*. Retrieved from `https://www.aaafoundation.org/sites/default/files/MeasuringCognitiveDistractions.pdf),`

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(8), 1300–1324.

Strayer, D. L., Watson, J. M., & Drews, F. A. (2011). Cognitive distraction while multitasking in the automobile. *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 29.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of*

Experimental Psychology, *18*(6), 643-662.

Sussman, E. D., Bishop, H., Madnick, B., & Walter, R. (1985). Driver inattention and highway safety. *Transportation Research Record*, *1047*, 40–48.

Tandonnet, C., Burle, B., Hasbroucq, T., & Vidal, F. (2005). Spatial enhancement of eeg traces by surface laplacian estimation: comparison between local and global methods. *Clinical Neurophysiology*, *116*(1), 18–24.

Teodorescu, A. R., Moran, R., & Usher, M. (2015). Absolutely relative or relatively absolute: violations of value invariance in human decision making. *Psychonomic bulletin & review*, *23*(1), 22–38.

Teodorescu, A. R., & Usher, M. (2013, Jan). Disentangling decision models: from independence to competition. *Psychological review*, *120*(1), 1-38. doi: 10.1037/a0030776

Ter Braak, C. J. (2006). A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces. *Statistics and Computing*, *16*(3), 239–249.

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging*, *18*(3), 415-429.

The JASP Team. (2016). *Jasp (version 0.7.5)[computer software].* https://jasp-stats.org/.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434–464.

Townsend, J. T., & Eidels, A. (2011). Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin & Review*, *18*(4), 659–681.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*(4), 321–359.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, *121*(2), 179 -205.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods*, *18*(3), 368-84.

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological review*, *122*(2), 312 - 336.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550-592.

van der Feest, S. V. H., & Swingley, D. (2011, Mar). Dutch and english listeners' interpretation of vowel duration. *Journal of the Acoustical Society of America*, *129*(3), 57-63.

van Heuven, V., Van Houten, J., & De Vries, J. (1986). De perceptie van nederlandse

klinkers door turken. *Spektator*, *15*, 225–238.

van Ravenzwaaij, D., Brown, S. D., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, *119*(3), 381–393.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2015). A simple introduction to markov chain monte–carlo. *Psychonomic Bulletin & Review.*.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology*, *219*(4), 1017–1025.

van Ravenzwaaij, D., Provost, A., & Brown, S. D. (2016). A confirmatory approach for integrating neural and behavioral data into a single model. *Journal of Mathematical Psychology*.

Vehtari, A., Gelman, A., & Gabry, J. (2016). *Practical bayesian model evaluation using leave-one-out cross-validation and waic.* Retrieved from `http://arxiv.org/abs/1507.04544`

Verdonck, S., & Tuerlinckx, F. (2014). The ising decision maker: A binary stochastic network for choice response time. *Psychological review*, *121*(3), 422-462.

Verdonck, S., & Tuerlinckx, F. (2015). Factoring out non-decision time in choice rt data: Theory and implications. *Psychological review*, *123*(2), 208-218.

Vidal, F., Burle, B., Grapperon, J., & Hasbroucq, T. (2011). An erp study of cognitive architecture and the insertion of mental processes: Donders revisited. *Psychophysiology*, *48*(9), 1242–1251.

Vollrath, M., Meilinger, T., & Krüger, H.-P. (2002). How the presence of passengers influences the risk of a collision with another vehicle. *Accident Analysis &*

Prevention, *34*(5), 649–654.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.

Wagenmakers, E.-J., & Brown, S. D. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological review*, *114*(3), 830-841.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159.

Wagenmakers, E.-J., Steyvers, M., Raaijmakers, J. G., Shiffrin, R. M., Van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*(3), 332–367.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wang, J.-S., Knipling, R. R., & Goodman, M. J. (1996). The role of driver inattention in crashes: New statistics from the 1995 crashworthiness data system. In *40th annual proceedings of the association for the advancement of automotive medicine* (Vol. 377, p. 392).

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of

simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 385 - 398.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica*, *41*(1), 67–85.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2012). The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing. *The Journal of the Acoustical Society of America*, *131*(2), 1465–1479.

Winn, M. B., Rhone, A. E., Chatterjee, M., & Idsardi, W. J. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in psychology*, *4*.

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.

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the english lexicon project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 597.

Zandbelt, B., Purcell, B. A., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2014). Response times from ensembles of accumulators. *Proceedings of the National Academy of Sciences*, *111*(7), 2848–2853.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human

auditory cortex. *Cerebral cortex*, *11*(10), 946–953.