A parameter recovery assessment of a wide class of evidence accumulation models of decision-making.

Matthew Murrow

Department of Physics and Astronomy, Vanderbilt University, PMB 401807, 2301 Vanderbilt Place, Nashville, TN 37240

William R. Holmes

Cognitive Science Program and Department of Mathematics, Indiana University Bloomington, 1101 E. Tenth Street, Bloomington, IN 47405, USA

Author Note

Statements and Declarations:

This work was funded by the US National Science Foundation grant SES2242962. We report no conflicts of interest. Correspondence should be made to William R. Holmes, Cognitive Science Program, 1001 E. 10th St. Bloomington, IN 47405. Email: wrholmes@iu.edu. Phone: +1 615-322-2599.

Abstract

Computational modeling has become indispensable in investigating the dynamics of decision making processes. A prominent category of models in this domain are Evidence Accumulation Models (EAMs), which model both the decisions people make and the time they take. Many variations have been proposed which modify the drift rate, diffusion rate, and decision thresholds, encoding increasingly complex dynamics into the EAM framework. However, adding model features complicates parameter recovery, making model interpretation more difficult. In this work, we perform a parameter recovery study to a variety of common binary choice EAMs, identify the specific challenges for each, and explore how to improve their parameter recoverability. Though previous studies have addressed this question, they have been piecemeal in nature, with different groups applying different computational methods to study different models. We aim to unify this body of literature using the best currently available computational methods. Further, we present the first, to our knowledge, Bayesian analysis of diffusion conflict models. Our purpose here is to be thorough, not exhaustive or comprehensive. With this in mind, this article catalogues a number of results, some previously shown and some new. Further, it illustrates different approaches to model analysis. This article is intended to be a resource for researchers interested in utilizing EAMs for studying decision-making processes, providing insights into the challenges associated these models, how to analyze them in light of those challenges, and examples of how to address those challenges.

Keywords: Decision-making, Drift-diffusion, Evidence Accumulation, Changing thresholds, Urgency gating, Conflict model, Parameter recovery, Bayesian parameter estimation

A parameter recovery assessment of a wide class of evidence accumulation models of decision-making.

Introduction

For over half a century, computational modeling has been instrumental in exploring decision-making dynamics. With the challenges of directly manipulating and observing the brain, models play a crucial role in formalizing and evaluating mechanistic hypotheses regarding decision-making processes. They enable us to explore questions such as how individuals process information over time, adapt their caution levels across diverse contexts, or navigate through complex choice landscapes with multiple alternatives and attributes. Such questions often prove elusive through observation or statistical analysis alone. Models offer a robust method to indirectly investigate these questions by comparing the anticipated data patterns predicted by models to experimental findings.

One prominent category of models in this realm are Evidence Accumulation Models (EAMs). EAMs model decision outcomes by simulating the processes involved in making a choice. For instance, the commonly used Diffusion Decision Model (DDM) posits that individuals probabilistically sample information over time, incrementally accumulate evidence based on this information, and reach a decision upon reaching a critical evidence threshold. Such models, including variants, are often represented mathematically through stochastic differential equations (SDEs). A notable advantage of this model family lies in its ability to model both the decisions made and the time taken to reach them, commonly referred to as Choice–Response Time (choice-RT) data. Crucially, the duration of decision-making offers valuable insights into the properties of underlying cognitive processes.

Evidence Accumulation Models (EAMs) are widely recognized in cognitive psychology for their effectiveness. They accurately capture various aspects of decision-making and response times, including the trade-off between speed and accuracy, the right skew in human response time distributions, the relationship between average response times and their variability, and distinctions between quick and slow

errors (Brown & Heathcote, 2008; Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff, Zandt, & McKoon, 1999; Usher & McClelland, 2001). Additionally, they find application in diverse areas such as learning processes (Evans, Brown, Mewhort, & Heathcote, 2018; Fontanesi, Gluth, Spektor, & Rieskamp, 2019), categorization tasks (Nosofsky, Little, Donkin, & Fific, 2011; Nosofsky & Palmeri, 1997), memory mechanisms (Osth & Farrell, 2019; Ratcliff, 1978), language comprehension (Lerche, Christmann, & Voss, 2018; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), consumer decision-making (Busemeyer, Gluth, Rieskamp, & Turner, 2019; Evans, Holmes, & Trueblood, 2019), development and aging (Ratcliff, Thapar, & McKoon, 2001; van Wouwe et al., 2016; Wieschen, Makani, Radev, Voss, & Spaniol, 2023), and personality and mood (White, Ratcliff, Vasey, & McKoon, 2010). Researchers have also begun to integrate EAMs with psychophysiological data, such as neural recordings (Turner et al., 2013; Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016; Turner, van Maanen, & Forstmann, 2015), motor performance data (Servant, White, Montagnini, & Burle, 2016), eye movements (Krajbich, Armel, & Rangel, 2010), functional magnetic resonance imaging (Forstmann, van den Wildenberg, & Ridderinkhof, 2008; White et al., 2014), electromyography (Servant, White, Montagnini, & Burle, 2015), and electroencephalography (Kelly & O'Connell, 2013; Philiastides, Heekeren, & Sajda, 2014; Servant et al., 2016).

Their popularity and success has led to the proposal of numerous modifications to the original DDM concept. Such proposals have modified the drift rate (Cisek, Puskas, & El-Murr, 2009; Dendauw et al., 2024; Ditterich, 2006; Smith, 1995; Ulrich, Schröter, Leuthold, & Birngruber, 2015), the diffusion rate (Cisek et al., 2009; Trueblood, Heathcote, Evans, & Holmes, 2021), and the shape of the decision thresholds (Churchland, Kiani, & Shadlen, 2008; Ditterich, 2006; Evans & Hawkins, 2019; Evans, Hawkins, & Brown, 2020; Hanks, Mazurek, Kiani, Hopp, & Shadlen, 2011; Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015; Palestro, Weichart, Sederberg, & Turner, 2018; Voskuilen, Ratcliff, & Smith, 2016), with the intention of exploring decision mechanisms beyond those encoded in the standard DDM. Though the

introduction of these additional features into the EAM framework substantially increases its investigative breadth, it also presents challenges. One of the prime values of EAMs over the decades has been their interpretability. They encode interpretable mechanistic features and by fitting them to data researchers can understand how factors of interest impact those features. These features are encoded in the model's structure and parameterization and interpretation often requires analysis of the parametric behavior of the model. Encoding more complex model features is however known to make such parametric analyses more challenging and less robust (Boehm et al., 2018; Dutilh et al., 2019; Evans et al., 2019; Evans, Trueblood, & Holmes, 2020; Holmes & Trueblood, 2018; White, Servant, & Logan, 2018). Thus there is a tension. More complex models may facilitate investigation of wider ranges of questions and phenomena, but that expansiveness may come at the expense of interpretability.

A number of works have sought to address this question of estimation and model interpretability through parameter recovery studies. The general approach taken is to 1) choose a subset of models to explore, 2) simulate data from the models using a set of known parameters, 3) generate a mathematical representation of the model, then 4) fit the model representation to the simulated data to determine if the input model parameters can be recovered. Among others, Lerche and Voss (2016) and Boehm et al. (2018) explored the parameter recoverability of the DDM; Evans, Trueblood, and Holmes (2020) explored the parameter recovery of the DDM in addition to models containing time changing thresholds and drift rates; Trueblood et al. (2021) focused specifically on the recovery of the Urgency Gating Model, an EAM with non-constant drift and diffusion rate; White et al. (2018) explored the recoverability of EAMs modeling conflict tasks; van Ravenzwaaij and Oberauer (2009) explored the DDM, the Linear Ballistic Accumulator Model, and the Leaky Competing Accumulator Model; and Evans et al. (2019) explored the recoverability of EAMs of multi alternative multi attribute choice.

Though comprehensive, these studies have been piecemeal in nature. Different groups investigate different models using different computational methods. Some

studies use point estimate methods (maximize some cost function) while others use Bayesian methods. Some rely on brute force stocahstic simulations of models (e.g. Probability Density Approximation (Holmes, 2015; Turner & Sederberg, 2014) and Quantile Maximation (Heathcote, Brown, & Mewhort, 2002; Ratcliff & Tuerlinckx, 2002)) while others rely on probabilistic methods (e.g. PyDDM (Shinn, Lam, & Murray, 2020), fast-dm (Voss & Voss, 2007) and PyBEAM (Murrow & Holmes, 2024)) Some analyses are preformed using full distributional representations of the predictions of models (likelihood functions) while others compress those representations into summary statistics (RT quantiles, for example). Further, as methods evolve over time, it is unclear how those findings translate using current best practices and methods. One of the goals of this article is to unify this body of literature, and to do so using the best currently available computational methods.

In this work, we seek to address this gap in the literature. We explore the parameter recoverability of a wide number of commonly used EAMs that vary in their implementation of the drift rate, diffusion rate, and decision thresholds. To do so, we use the recently developed Python package PyBEAM, a tool which uses Bayesian methods to fit these models to full choice-RT data (Murrow & Holmes, 2024). We are using the best available modeling approach and applying it uniformly to these models to assess their qualities. We identify the specific challenges associated with fitting each model to data, then A) explore how to improve their parameter recoverability and B) provide recommendations for how to use them. The intent of this paper is to act as a single resource for the analysis of a suite of common binary choice EAMs, and to provide practical recommendations for best use to researchers who may be interested in studying them. That said, this is not a comprehensive assessment of such models. The analysis of a model should always be tied to the structure of the data available and the scientific purpose of using that model. The results and approaches here-in can however serve as a starting point for such analyses.

Models

In this section, we first introduce the general structure of the models explored in this work. Next, we introduce the specific implementation details for each model. Lastly, we discuss our approach to performing numerical experiments to test the parameter recovery of each model and provide recommendations for practical use.

Two threshold, binary evidence accumulation models

Evidence accumulation models (EAMs) hypothesize that, during a decision, information is stochastically sampled from the stimulus, then additively accumulated until a critical level of evidence is reached. In this article, we specifically focus on the class of two threshold, binary evidence accumulation models discussed in the "Introduction."

EAMs of this type take the general form shown in Panel A of Figure 1. The horizontal axis of this panel is elapsed time from stimulus presentation, while the vertical axis provides the total accumulated evidence x at time t. Two thresholds are present, an upper, positive valued function ($b_1(t)$, solid line) and lower, negative valued function ($b_2(t)$, dashed line), each corresponding to one of the two choices available. Though the thresholds in this figure are constant in value, they can also vary with time, either expanding or collapsing from their initial location. The separation between thresholds indicates the level of caution a exhibited by the decision maker. If the thresholds are far apart (near), the decision process will be slower (faster), resulting in more (less) accurate decisions.

Evidence accumulation begins at the start point z, indicated by the blue dot on the left of the panel. The start point can be located anywhere between the upper and lower thresholds and corresponds to an initial bias towards one of the two choices prior to stimulus presentation. For convenience, the start point is often written as a ratio of the threshold separation. This new parameter is referred to as the relative start point, and is given by $w = z/(2a)$. Noisy accumulation proceeds from the start point and continues until one of the two decision thresholds have been reached. If the upper

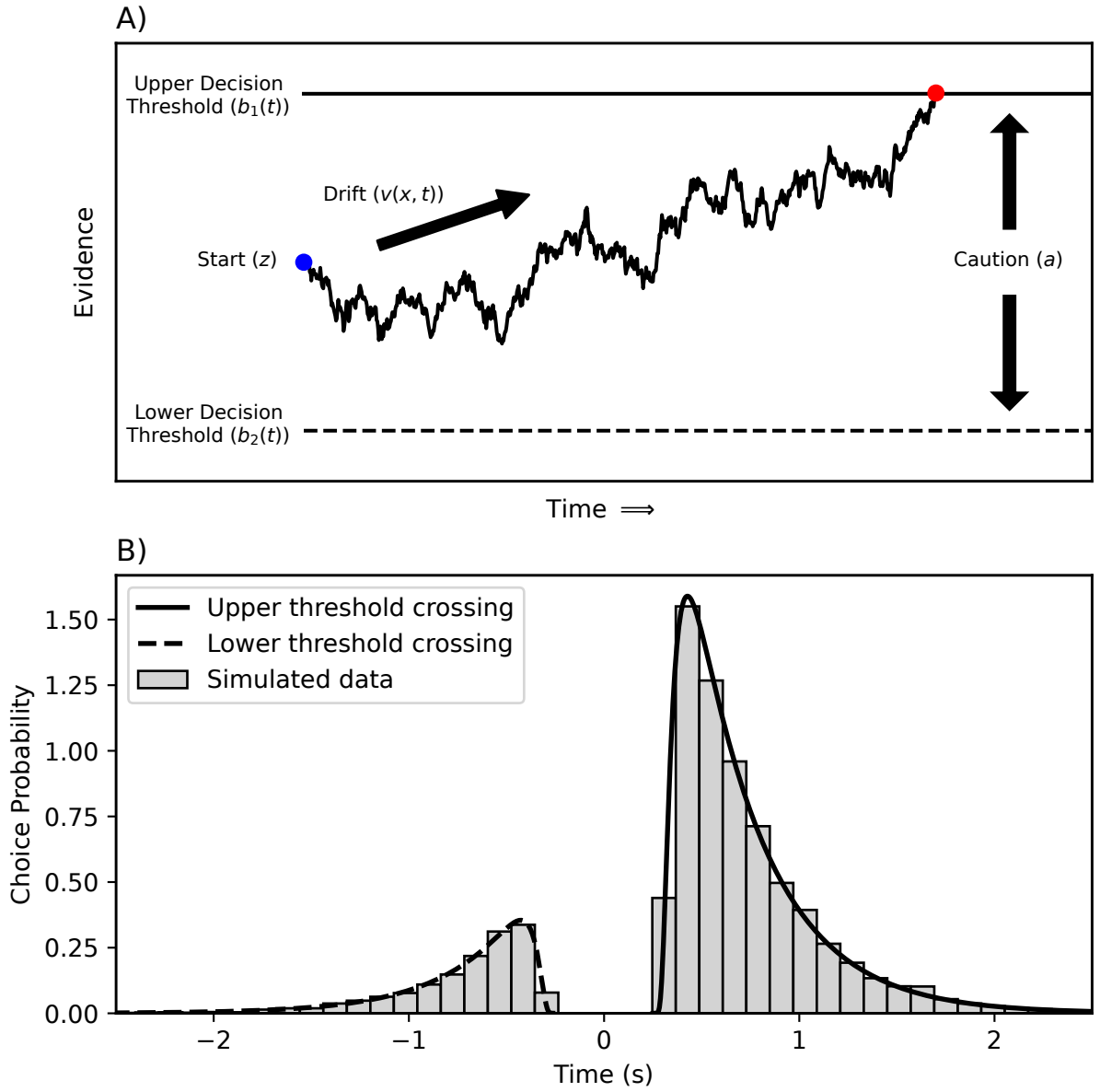


Figure 1. Figure caption on next page.

(lower) threshold is reached first, then choice one (two) is chosen and the time at which the choice occurs is the threshold crossing time. The rate at which evidence accumulates is referred to as the drift rate, indicated by the upwards pointing arrow. An additional parameter t_{nd} is added to the final choice-RT value which encodes non-decision related behaviors such as stimulus encoding and motor movement.

EAMs of this type can be written as the following Stochastic Differential Equation (SDE),

$$dx(t) = v(x,t)dt + D(x,t)dW(t), \tag{1}$$

Figure 1. A) Schematic of an EAM. Evidence is accumulated (black line) starting from an initial bias z (black dot) until one of the two decision thresholds, $b_1(t)$ or $b_2(t)$, is reached (red dot), triggering a decision. Though the thresholds are constant in this plot, their position can vary as a function of time. The distance between thresholds indicates the degree of caution (a) exhibited by the decision maker, which can remain constant or change as a function of time if the thresholds are not constant. This process is described mathematically by Equation (1). B) Example of data simulated from an EAM via Equation (1) (grey bars) with likelihood functions overlaid. Data for the upper decision threshold crossing is shown on the positive time axis, while data for the lower threshold crossing is shown on the negative time axis. The solid black line corresponds to the likelihood function for the upper decision threshold, while the dashed black line gives the likelihood function for the lower decision threshold. The probability of making a choice at any time is given on the vertical axis.

where $x(t)$ is the total evidence accumulated at time t and $v(x,t)$ is the rate of evidence accumulation, referred to as the drift rate. The function $D(x,t)$ is the diffusion rate, and it is commonly fixed for scaling purposes. Lastly, $W(t)$ is the standard Wiener process. Once evidence $x(t) \geq b_1(t)$ or $x(t) \leq b_2(t)$, a choice is triggered.

In this article, we consider a variety of models which differ in their assumptions about the drift rate, diffusion rate, and decision threshold behavior. The drift rate $v(x,t)$, diffusion rate $D(x,t)$, and thresholds $b_i(t)$ are free to vary as functions of time and/or evidence. Here, we consider several models of this type commonly applied in the literature: the Simple DDM, EAMs with leaky integration, EAMs with changing thresholds (either Weibull, exponential, or linear), the Urgency Gating Model, and the Diffusion Model of Conflict, all of which are described in detail below.

Simple DDM

The first model we examine, the Simple DDM (sDDM), is the simplest EAM we examine and the basis of all upcoming models. We use this model as the baseline by

which to gauge recovery. The sDDM is similar to the Ratcliff Diffusion Decision Model (DDM) (Ratcliff, 1978; Ratcliff & McKoon, 2008), but excludes across-trial variabilities in the start point, non-decision time, and drift rate. The sDDM assumes that the drift rate, decision thresholds, start point, and non-decision time are all fixed quantities throughout the decision process. Thus, it contains four parameters: non-decision time t_{nd}; drift rate $v(x,t) = \mu$, where μ is the stimulus strength; relative start point w; and flat, symmetric decision thresholds $b_1(t) = -b_2(t) = b$. The diffusion rate $D(x,t)$ acts as the model scaling parameter and is fixed at $D(x,t) = 1$ for all our numerical experiments. Note that all EAMs we use going forwards keep the same value for the diffusion rate, with only a small modification for the Urgency Gating Model.

Leaky Integration Model

The leaky integration model, referred to as "Leakage" from here on, is an extension of the sDDM with leaky integration added to the drift rate. This model is also referred to in the literature as an Ornstein-Uhlenbeck process. Leakage is used to more realistically describe the accumulation process by modeling the decay of excitatory currents in decision neurons (Usher & McClelland, 2001). Leaky integration changes the drift rate from a constant to the following,

$$v(x,t) = \mu(t) - L\,x(t), \tag{2}$$

where $\mu(t)$ is the stimulus strength (allowed to vary with time) and L is the leakage strength. Addition of the leakage parameter causes old information to decay over a scale approximately equal to $1/L$, leading to the favorable property of evidence decaying to zero when the stimulus is removed. This property has led it to being proposed as a mechanism for preference reversals under time pressure (Busemeyer & Townsend, 1993). If leakage is large, evidence is rapidly lost from the accumulator. Conversely, small leakage values imply that accumulated evidence is retained for longer periods of time, making the accumulator less sensitive to novel stimulus information.

The leakage parameter can be difficult to recover, so we follow the lead of Evans, Trueblood, and Holmes (2020) and Trueblood et al. (2021) and run numerical

simulations with two different drift rate implementations: fixed information and changing information. In the fixed information case, the stimulus strength is a constant, given by $\mu(t) = \mu$, where μ is the strength of stimulus information. For the changing information case, Trueblood et al. (2021) proposed an experiment with time changing information implemented via a grid of pixels flashing one of two colors (blue / orange). They altered the fraction of each color on the screen at a given time while a decision was being made such that, as an example, early in the decision process, the grid may be 55 percent blue, while later it might be 55 percent orange. This is implemented by allowing the stimulus strength to change with time, given by,

$$\mu(t) = \begin{cases} \mu & \text{if } t < t_0, \\ -\mu & \text{if } t \geq t_0, \end{cases} \tag{3}$$

where, as in the fixed information case, μ_0 is the strength of stimulus information, while t_0 is the time where stimulus information changes, referred to as the flip time. In the context of the above example, at $t < t_0$, the grid would be predominately blue (corresponding to positive μ), while for $t \geq t_0$, the grid would be mainly orange (corresponding to negative μ).

Time changing decision thresholds

We next examine EAMs which have time changing decision thresholds (CT). In the case of CT, the first three parameters of the sDDM are retained unmodified: the non-decision time t_{nd}, the relative start point w, and the drift rate $v(x, t) = \mu$. However, the thresholds $b_i(t)$ now are a function of time. Though in principle they are free to increase or decrease from their starting value, most interest in the literature has been directed towards thresholds which collapse from their starting point towards zero.

Psychologically, CTs encode dynamic changes in decision strategy. They allow for the subject to optimize the degree of caution they exhibit as time progresses to best meet the demands of the decision task. Thresholds which specifically collapse were shown to provide a mechanism that maximizes the reward rate in decision tasks with unpredictable information (Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget,

2012; Thura, Beauregard-Racine, Fradet, & Cisek, 2012). In addition, collapsing thresholds provide a natural way to optimize decision behavior when experimental instructions include deadlines or emphasize speed of responding. For example, if a response time deadline is given to a subject, a fixed threshold strategy forces the subject to decrease the entire threshold location. Though this guarantees that the decision is completed by the deadline, it does so at the expense of accuracy early in the decision process. Conversely, collapsing thresholds allow the subject to keep thresholds distant early in the decision process to emphasize accuracy, then decrease the thresholds near the deadline to increase decision speed. For an in depth discussion, see Malhotra, Leslie, Ludwig, and Bogacz (2018), who developed a theoretical framework which explores in depth the conditions which favor changing thresholds.

Time changing thresholds also provide an alternate cause for the discrepancy in the average error and correct choice-RT. It is known that, in general, the average error choice-RT is not equal to the average correct choice-RT (Luce, 1991; Swensson, 1972). This behavior is most commonly accounted for via the addition of across-trial variability in the drift rate, as in the DDM (Ratcliff, 1978; Ratcliff & McKoon, 2008). Changing thresholds provide an alternate mechanism to explain this behavior, predicting error RTs that are slower than the correct RTs (Ditterich, 2006).

Lastly, it has been shown that, in certain circumstances, primates implement changing thresholds in order to optimize their decision process (Hawkins et al., 2015). Though it is still a matter of debate (beyond the scope of this article) whether humans routinely exhibit changing decision thresholds, exploration into this question is an active area of research (Evans & Hawkins, 2019; Evans, Hawkins, & Brown, 2020; Hawkins et al., 2015; Palestro et al., 2018).

The principal threshold we study here is the Weibull threshold. This threshold uses a Weibull cumulative distribution function to model the decision threshold behavior. It is a commonly used threshold due to its flexibility in behavior (Hawkins et al., 2015) and has been used in a number of studies (Evans, Hawkins, & Brown, 2020;

Hawkins et al., 2015; Palestro et al., 2018). The threshold is given by,

$$b_1(t) = -b_2(t) = b_0 - \frac{b_0(1-c)}{2}\left[1 - \exp\left(-\left(\frac{t}{\lambda}\right)^{\kappa}\right)\right]. \tag{4}$$

Here, b_0 is the initial threshold location. Parameter λ is referred to as the scale parameter, and approximately sets the time at which the threshold expands or collapses. Parameter κ is referred to as the shape parameter and indicates if the threshold is of exponential-like shape ($\kappa < 1$) or logistic-like shape ($\kappa > 1$). The remaining parameter, c, is the collapse ratio, and indicates how much the thresholds expand or collapse. If $c = -1$, the thresholds collapse to zero; if $-1 < c < 1$, the thresholds collapse to somewhere between their initial location b_0 and zero; if $c = 1$, no threshold collapse occurs; and if $c > 1$, the threshold expands away from b_0. Generally, it is assumed that decision thresholds collapse, and thus $c < 1$. For a more detailed discussion of this threshold's behavior, we refer the reader to the PyBEAM publication (Murrow & Holmes, 2024) and Hawkins et al. (2015).

In addition to the Weibull threshold, we also examine several simpler thresholds containing fewer parameters. The first of these, the Reduced Weibull threshold, is identical to the Weibull threshold in Equation (4), with the exception that the collapse parameter is fixed at $c = -1$. This causes the threshold to always collapse to zero, constraining the behavior of the model and reducing the number of parameters it adds to the sDDM to two. The Reduced Weibull threshold has been applied in this or a similar form in Hawkins et al. (2015) and Evans, Hawkins, and Brown (2020).

The next two simpler CTs we examine are the linear and exponential thresholds which each contain a single additional threshold parameter. The first, linear, defines the decision thresholds as,

$$b_1(t) = -b_2(t) = b_0 - mt, \tag{5}$$

where b_0 indicates the threshold location at time zero and m is the thresholds' slope. The second, exponential, defines the decision thresholds as,

$$b_1(t) = -b_2(t) = b_0 \exp(-t/\tau), \tag{6}$$

where b_0 is, as before, the decision threshold location at time zero and τ describes the rate of threshold collapse.

Urgency Gating Model

The Urgency Gating Model (UGM) and other similar urgency based models propose an alternate, yet related, account for decision making behavior (Cisek et al., 2009). Unlike EAMs which posit that accumulation proceeds through gradual accumulation and integration of stimulus information, urgency models suggest that a time-varying gain function is the principal means by which the decision state is updated. In the context of two threshold binary choice, the simplest implementation of urgency is given by,

$$y(t) = g \cdot E(t) \cdot u(t), \tag{7}$$

where $y(t)$ is the decision state, g is a scalar gain term, $E(t)$ is the strength of the momentary evidence, and $u(t)$ is a gain function which describes an increasing urgency to make a response. A common choice for the urgency $u(t)$ is a linear function, given by,

$$u(t) = b + mt, \tag{8}$$

where b represents a baseline urgency, and m describes the rate of urgency increase over time.

The UGM is a specific implementation of this class which hypothesizes that the decision variable $y(t)$ is affected by two main factors: the time dependent urgency function $u(t)$ discussed above, and a low pass filtered representation of integrated stimulus information $x(t)$, implemented via leaky integration. The decision variable is given by,

$$y(t) = x(t)u(t) \tag{9}$$

while the integrated stimulus information is,

$$dx(t) = (E(t) - Lx)dt + \sigma dw, \tag{10}$$

where $E(t)$ is the evidence signal, $u(t)$ is the linear urgency signal introduced above, and L is the rate of leaky integration introduced in the leakage model.

As shown by Trueblood et al. (2021), when an appropriate non-dimensional variable is used, models of this kind can be written in a form identical to that of Equation (1). Doing so modifies the drift rate of Equation (1) to,

$$v(x,t) = E(t)(1+kt) + \left(\frac{k}{1+kt} - L\right)x. \tag{11}$$

where x is, in this case, a non-dimensionalized decision variable, and $k = m/b$ and represents an urgency ratio. In short, the urgency ratio describes the strength of urgency in the system, with $k = 0$ implying no urgency is present, and large k implying an urgency dominated system. In addition to altering the drift rate, the UGM introduces a modified diffusion rate, given by,

$$D(x,t) = \sigma(1+kt), \tag{12}$$

where k is again the urgency ratio, σ is the scaling parameter from the sDDM and, as for the sDDM, it is set to one.

Similarly to the leakage model, the UGM has well known recovery problems for the leakage, urgency, and threshold parameters. Thus, as we did for the leakage model, we follow the lead of Trueblood et al. (2021) and implement two versions of $E(t)$: fixed information and changing information. In the fixed information version, we assume that the evidence signal is constant, given by $E(t) = E_0$. In the changing information case, we allow the evidence signal to change with time, given by,

$$E(t) = \begin{cases} E_0 & \text{if } t < t_0, \\ -E_0 & \text{if } t \geq t_0. \end{cases} \tag{13}$$

where E_0 is the strength of stimulus information and t_0 is, as with the leakage model, the time at which stimulus information changes.

Urgency signal models are motivated by many of the same questions as CT models. However, unlike CTs, urgency signals propose that strategic manipulation is implemented via the drift and diffusion rates instead of the threshold function. Recent work has demonstrated that these accounts of decision making are in fact equivalent (Smith & Ratcliff, 2022), an observation that we expand upon in the "Results" section.

Diffusion Model of Conflict

The Diffusion Model of Conflict (DMC) is an EAM used to describe conflict tasks (Ulrich et al., 2015). Conflict tasks are decision scenarios where conflicting evidence is present for the correct response. Classic examples of conflict tasks are the Stroop task (Stroop, 1935), where a written color word can conflict with the color it is written in; the flanker task (Eriksen & Eriksen, 1974), where flanking items conflict with the target item; and the Simon task (Simon & Small, 1969), where the stimulus location can conflict with the response. The DMC was developed to model all three.

Its structure is based off the sDDM introduced earlier, and maintains three of its parameters: the non-decision time t_{nd}, the relative start point w, and the decision thresholds $b_1(t) = -b_2(t) = b$. The original implementation of the DMC also includes across-trial variability in the non-decision time and start point, but we exclude them in this work for simplicity. The drift rate, however, deviates from the assumption of the sDDM, and posits that evidence accumulation is a combination of early automatic processing and late controlled processing. The early activation is modeled via a scaled gamma function, which provides strong manipulation of the drift rate at early times, and weakens as time progresses. In the context of the flanker task, this early activation is driven by the flanking non-target arrows and is given by,

$$v_a = Ae^{-t/\tau}\Big[\frac{te}{\tau(\alpha-1)}\Big]^{\alpha-1}\Big[\frac{\alpha-1}{t} - \frac{1}{\tau}\Big], \tag{14}$$

where A is the amplitude of the early activation, equaling a positive value for congruent tasks and negative value for incongruent tasks. Parameter τ sets the scale of the early activation, while α sets the shape of the early activation.

The controlled drift rate μ_c is assumed to be constant as in the case of the sDDM, and dominates the drift as time progresses. In the context of the flanker task, for example, this models the shift from early activation driven by the flankers to the late activation driven by the target arrow. The total drift rate is the sum of the automatic and controlled process,

$$v(x,t) = v_a + v_c. \tag{15}$$

The DMC lacks a simple analytic solution, leading to it being only analyzed using simulation studies and quantile maximization approaches. A study has been performed to study parameter recovery of the DMC (White et al., 2018), but, to our knowledge, here is the first time that the DMC will be examined using the entire likelihood function and Bayesian methods. We also seek to answer two other unexplored questions: can the DMC parameters still be recovered when the response bias parameter w is included, and does the inclusion of multiple drift rate conditions meaningfully improve parameter recovery?

Choosing parameter sets

The general procedure used in this study is 1) generate a wide range of parameter sets representative of realistic model behavior, 2) simulate data from each parameter set, and 3) fit each model to the simulated data using PyBEAM, then analyze parameter recovery reliability by comparing the input and best fit parameter sets. In this section, we discuss our approach to each of these steps. To obtain a complete assessment of each model's identifiability, we generate a large number of parameter sets across a wide range of values, displayed for each model in Table 1. Additionally, we generate parameter sets across a range of simulation set sizes N to determine the practical data set sizes needed for recovery.

The process used to generate parameter sets (for all models but the DMC) is as follows. First, for each model and N value, we randomly generate parameter sets from the ranges listed in Table 1 using a Latin Hypercube Sampling design (LHS). We choose LHS over random sampling to ensure that the entire parameter space is evenly explored, something random sampling struggles with in high dimensional parameter space. Next, we filter out parameter sets which produce atypical choice-RT distributions. Our goal with this filter is to censor out parameter sets which produce data unlike that seen in experiment. To do so, we first simulate N of data for each parameter set using the methodology discussed below in section "Simulating data." Then, we eliminate sets that do not fit certain distributional criteria (Evans, Trueblood,

& Holmes, 2020; Trueblood et al., 2021). Specifically, we keep only parameter sets whose simulated data meets the following standards: mean and median response times between 0.4 and 2.5 seconds; interquartile range between 0.1 and 2 seconds; minimum RT below 1.5 seconds and maximum RT above 0.5 seconds; and less than five RTs for either decision. We randomly generate a sufficient number of parameter sets such that, after filtering, we are left with 1000 total parameter sets for each sample size.

An additional filter is used for the models with time varying information. The information change must occur at a meaningful time in the decision making process. If it occurs too early, no accumulators will have reached threshold before the flip occurs. If too late, most or all accumulators will have reached the decision threshold prior to the flip. To address this, we constrain the flip times to be between the first and third quartiles of the choice-RT distribution and eliminate parameter sets which do not fit this.

For the DMC, we used the parameter ranges given by White et al. (2018). In their work, these parameter ranges were constrained sufficiently that filtering was unnecessary, so we do not filter out any parameter sets for this model. Additionally, since the DMC is slower to run, we only use 100 unique parameter sets per experiment.

Simulating data

Data for each parameter set is simulated using the Python package PyBEAM (Murrow & Holmes, 2024), discussed in detail in the "Introduction". PyBEAM contains pre-coded versions of each model and simulates the models by integrating Equation (1) using the Euler-Maruyama method (Kloeden & Platen, 1992),

$$x_n = x_{n-1} + v(x_{n-1}, t_{n-1})\Delta t + D(x_{n-1}, t_{n-1})\Delta W\sqrt{\Delta t}, \tag{16}$$

where x_n is the accumulated evidence at time step n, x_{n-1} is the accumulated evidence at time step $n-1$, and t_{n-1} is the time at time step $n-1$. Functions $v(x_{n-1}, t_{n-1})$ and $D(x_{n-1}, t_{n-1})$ are, as in Equation (1), the drift and diffusion rates, respectively, evaluated at accumulated evidence x_{n-1} and time step t_{n-1}. Term ΔW simulates the Wiener process $W(t)$ from Equation (1) by drawing a random number from a normal

sDDM	t_{nd}	w	μ	b			
	0.1–0.6	0.3–0.7	-5–5	0.25–1.5			
Leakage	t_{nd}	w	μ	L	t_0	b	
	0.1–0.6	0.3–0.7	-5–5	1–10	0.2–2	0.1–1	
FWeibull	t_{nd}	w	μ	b	λ	κ	c
	0.1–0.6	0.3–0.7	-5–5	0.25–1.5	1–10	0.4–10	-1.0–0.5
RWeibull	t_{nd}	w	μ	b	λ	κ	
	0.1–0.6	0.3–0.7	-5–5	0.25–1.5	1–10	0.4–10	
Linear	t_{nd}	w	μ	b	m		
	0.1–0.6	0.3–0.7	-5–5	0.25–3.0	0.1–2.5		
Exp.	t_{nd}	w	μ	b	τ		
	0.1–0.6	0.3–0.7	-5–5	0.25–3.0	0.1–10		
UGM	t_{nd}	w	μ	L	k	t_0	b
	0.1– 0.6	0.3–0.7	-5–5	0.1–10	0.1–10	0.2–2	0.1–3
DMC	t_{nd}	w	A	τ	α	μ_c	b
	0.27–0.4	0.5	0.12–0.32	0.02–0.12	1.5–4.5	1.6–6.3	0.36–0.63

Table 1

Parameter ranges used for each model: the Simple DDM (sDDM), the leaky integration model (Leakage), the Full Weibull CT (FWeibull), the Reduced Weibull CT (RWeibull), the Linear CT (Exp.), the Exponential CT (Exp.), the Urgency Gating Model (UGM), and the Diffusion Model of Conflict (DMC).

distribution with mean zero and standard deviation one at each integration step. The final term is the integration time step Δt, which sets the time at step n via $t_n = n\Delta t$. The simulation ends when a decision threshold has been crossed, given by $x_n \geq b_1(t_n)$ or $x_n \leq b_2(t_n)$.

The choice of Δt is dependent upon model, with models with shorter time scales requiring smaller time steps. For this numerical experiment, our goal is to isolate parameter identifiability, so we choose conservatively small Δt values to reduce the noise added to the system via simulation error. For all models but the DMC, we use $\Delta t = 1.0e-4$ seconds, while for the DMC we use $\Delta t = 1.0e-6$ seconds. We choose a smaller time step for the DMC since its dynamics occur on a scale nearly an order of magnitude faster than that of the other models.

Fitting models to data

Once we have simulated data for our parameter sets, we next fit each model to the simulated data. To do so, we again utilize the Python package PyBEAM (Murrow & Holmes, 2024). As discussed in the Introduction, PyBEAM is a fast, accurate method for Bayesian modeling of full choice-RT distributions. Specifically, PyBEAM fits models to data by calculating the model's first passage time distribution, commonly referred to as the likelihood function. The likelihood function gives the probability of crossing either decision threshold at time t, and thus gives the probability of a making a choice at time t.

An example of the likelihood function overlaid on a simulated data set is shown in Panel B of Figure 1. The horizontal axis displays the time coordinate, while the vertical axis gives the probability of making a choice. For convenience, we place the lower threshold crossing data in negative time, and the upper threshold crossing data in positive time.

Discussed in detail in the package publication, PyBEAM generates the likelihood function in two main steps. First, it converts the SDE formalism of Equation (1) to the

probabilistic Fokker-Planck equation,

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial \left[v(x,t)p(x,t)\right]}{\partial x} + \frac{1}{2}\frac{\partial^2 \left[D(x,t)^2 p(x,t)\right]}{\partial x^2}, \tag{17}$$

where $p(x,t)$ is the probability of accumulated evidence x at time t, and $v(x,t)$ and $D(x,t)$ are the drift and diffusion rates discussed earlier. This equations provides the probability at a given time t of having accumulated evidence quantity x. Then, to determine the probability $f_i(t)$ of crossing a decision threshold b_i at time t, it calculates from Equation (17) the probability flux at the threshold, given by

$$J(x,t) = v(x,t)p(x,t) - \frac{1}{2}\frac{\partial \left[D(x,t)^2 p(x,t)\right]}{\partial x}. \tag{18}$$

This is the probability flux at point (x,t) and the likelihood of crossing threshold b_i is $f_i(t) = J(b_i(t),t)$.

PyBEAM uses the likelihood function to measure the level of agreement between a model with given parameters and data. This log-likelihood is used in a Bayesian framework to fit these models to data and obtain approximate posterior distributions for their parameters. The computed log-likelihood is then used by PyBEAM to perform Bayesian parameter estimation with the Python package PyMC (Salvatier, Wiecki, & Fonnesbeck, 2016), a robust, highly supported package built specifically for Markov chain Monte Carlo based inference. Though slower than other optimization methods like max log-likelihood or chi-squared statistics, we choose the above approach for fitting these models to data for several reasons. First, the use of PyBEAM allows access to rapidly generated, high resolution likelihood functions for all models with little to no modification. Most previous parameter recovery studies are restricted to Quantile Maximization approaches generated through simulation, which compress the information contained in the likelihood function. Second, the Bayesian inference algorithms of PyBEAM provide access to the entire distribution of parameter space rather than just the best fit parameters, giving us a more comprehensive way to analyze our model recoverability. This becomes particularly relevant in our "Results" section for our analysis of over-parameterized models.

Results

In this section, we present the findings of our numerical experiments. We examine the quality of fit for the five models discussed in "Methods": the sDDM, EAMs with leaky integration, EAMs with CTs, the UGM, and the DMC. We choose to include the sDDM since it is the basis of the more complex EAMs we examine and thus serves as a useful baseline to compare the other models to. The remaining are commonly used models (introduced in "Methods") which have, to our knowledge, not been comprehensively examined using the full likelihood function with Bayesian methods. There are of course numerous other models and variants that we do not consider here, but we hope this process and the reference scripts provided online in the PyBEAM documentaion (https://pybeam-documentation.readthedocs.io/en/latest/) will facilitate similar study of other models where useful.

Simple DDM (sDDM)

We start with our parameter recovery experiments for the sDDM. For this numerical experiment, we choose simulated data set sizes of $N = 100,\ 250,\ 500,\ 1000,$ and $10,000$ points. For each N, we use LHS to generate $1,000$ unique parameter sets, leading to a total of $5,000$ parameter sets. Data is simulated from the model using PyBEAM, which itself implements Equation (16) discussed in "Methods".

We then fit the sDDM to each generated data set. We display the results of this in Figure 2. Each column corresponds to one of the four Simple DDM parameters: the non-decision time (t_{nd}), the relative start point (w), the drift rate (μ), and the threshold location (b) (where $a = 2b$ is the threshold separation / caution). The column which corresponds to each parameter is indicated along the bottom of the figure. Each row corresponds to the number of simulated data points N in each numerical experiment, ranging from $N = 100$ to $N = 10,000$. The horizontal axis corresponds to the true parameters input to each simulation, while the vertical axis displays the best fit parameter sets, given by the parameter set with the maximum sum log-likelihood. The

R^2 values give the quality of fit, with points which fall along the red lines indicating perfect parameter recovery.

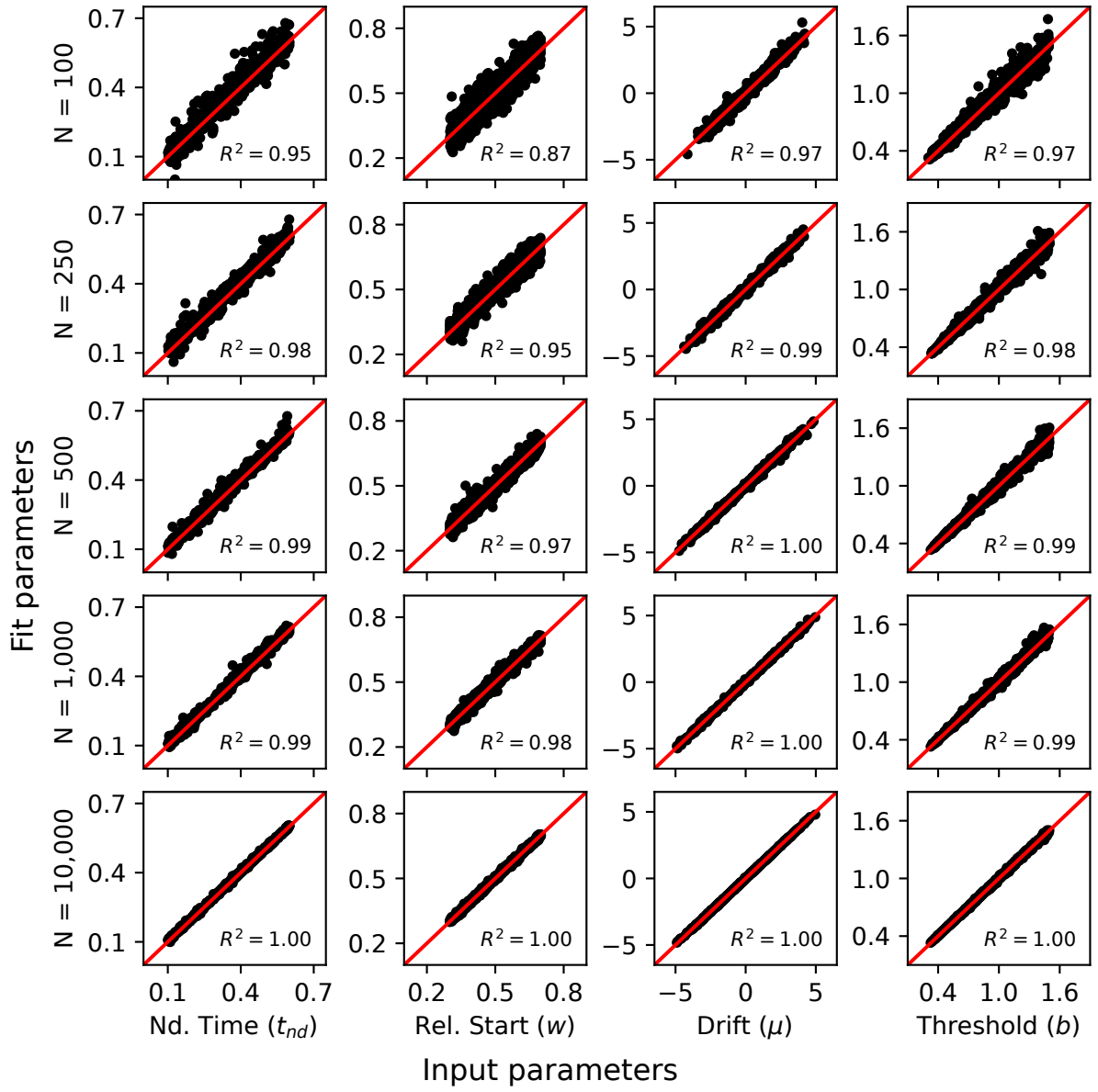


Figure 2. Quality of fit for the sDDM for $N = 100,\ 250,\ 500,\ 1000,$ and $10,000$ simulated data points. Horizontal axes indicate the true simulated parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot denote the location where the simulation and fit parameters are equal. The correlation coefficients R^2 indicate the quality of fit to the red line.

We find that, regardless of N, recovery for the Simple DDM parameters is very

good. Recovery is stable for as few as $N = 100$ samples, with near perfect recovery achieved once $N = 1,000$ is reached. The most difficult parameter to recover is the relative start point, which struggles relative to the other parameters for small N. We also find that b is relatively difficult to recover when it is large (approximately $b > 1$). This has several likely causes. First, when threshold starts are large, choice-RT distributions are dominated by the drift rate, resulting in fewer error RTs. This can make it more difficult for the precise value of the threshold to be narrowed down. Second, when thresholds are large, small deviations from the simulated thresholds result in only tiny errors since they are a smaller percentage of the threshold location. This model is well studied, many of these observations have been made, and we include here mainly as a starting point and baseline to compare future models with.

Leaky integration (Leakage)

In this section, we report the results of our parameter recovery experiments for the Leakage model. We display the first set of results in Figure 3. Each column corresponds to a different parameter, noted beneath the bottom row of panels. The horizontal axes provide the simulated input parameters, while the vertical axes display the fit parameters determined from the max log-likelihood of the posteriors. The red line on each panel indicates where the input and fit parameters are equal, with the R^2 values indicating the quality of fit of the data to the line. Each row displays the results for $N = 1,000$ data points.

In row A1, we display results for fixed information (FI) with a single drift rate condition, while in row A2, we display results for FI with two drift rate conditions. In row B1, we display results for changing information (CI) with a single drift rate condition, while in row B2, we display results for CI with two drift rate conditions. The two drift rate conditions are used for a similar reason to that of the CI, being that leakage is a difficult parameter to recover. Thus, we simulate a slightly more complicated data set in rows A2 and B2 to see if it has a meaningful impact on recovery: a smaller μ and a larger μ. This is used to model an experiment where both a

low quality and high quality stimulus are shown to the participant. In the context of the random dot kinematogram discussed earlier, this corresponds to a low coherence and a high coherence stimulus, respectively. The fits of both drift rate conditions are still displayed in the "Drift μ" column of Figure 3.

We find that, in the FI case, it is very difficult to recover the leakage parameter. Both in the one and two drift condition cases, the leakage parameter is mostly unidentifiable, with a small improvement present when two drift conditions are used. In the single condition case (A1), all other parameters are well recovered, with the exception of the the threshold b which struggles to be recovered for large values. Addition of the second drift rate condition slightly improves recovery of w and μ, and significantly improves the recovery of b, especially for large values.

Conversely, in the CI case, we find that the leakage parameter is well recovered, with substantially better R^2 values than that of FI. The remaining parameters are also well recovered, with the exception of large drift rates. Small drift rates ($\mu < 3$) are well recovered for CI, but larger drift rates exhibit substantially more variance, particularly in the single condition case (B1). Lastly, moving from a single to two drift rate conditions (B2) provided a large improvement in the recovery of b, with R^2 jumping from 0.92 to 1 when a second drift condition is added.

We next examine how parameter recovery scales with the number of samples. We perform the same experiment presented in row B1 of Figure 3 with fixed information and a single drift rate condition. However, we report results for $N = 100,\ 250,\ 500,\ 1000,$ and $10,000$ data points to demonstrate how recovery scales with simulation set size. We report the results of this in Figure 4. For $N = 10,000$ we find that recovery is excellent for all parameters. The $N = 1,000$ results are the same as Figure 3 which shows good recovery for each parameter. However, as N decreases, recovery becomes increasingly worse, with L unrecoverable for $N = 500$, and the remaining parameters difficult to recover for $N = 250$ and $N = 100$ data points. Thus, when attempting to recover leakage, it is necessary to use as many data points as possible to ensure good recovery of all parameters. Additionally, it is preferred to use

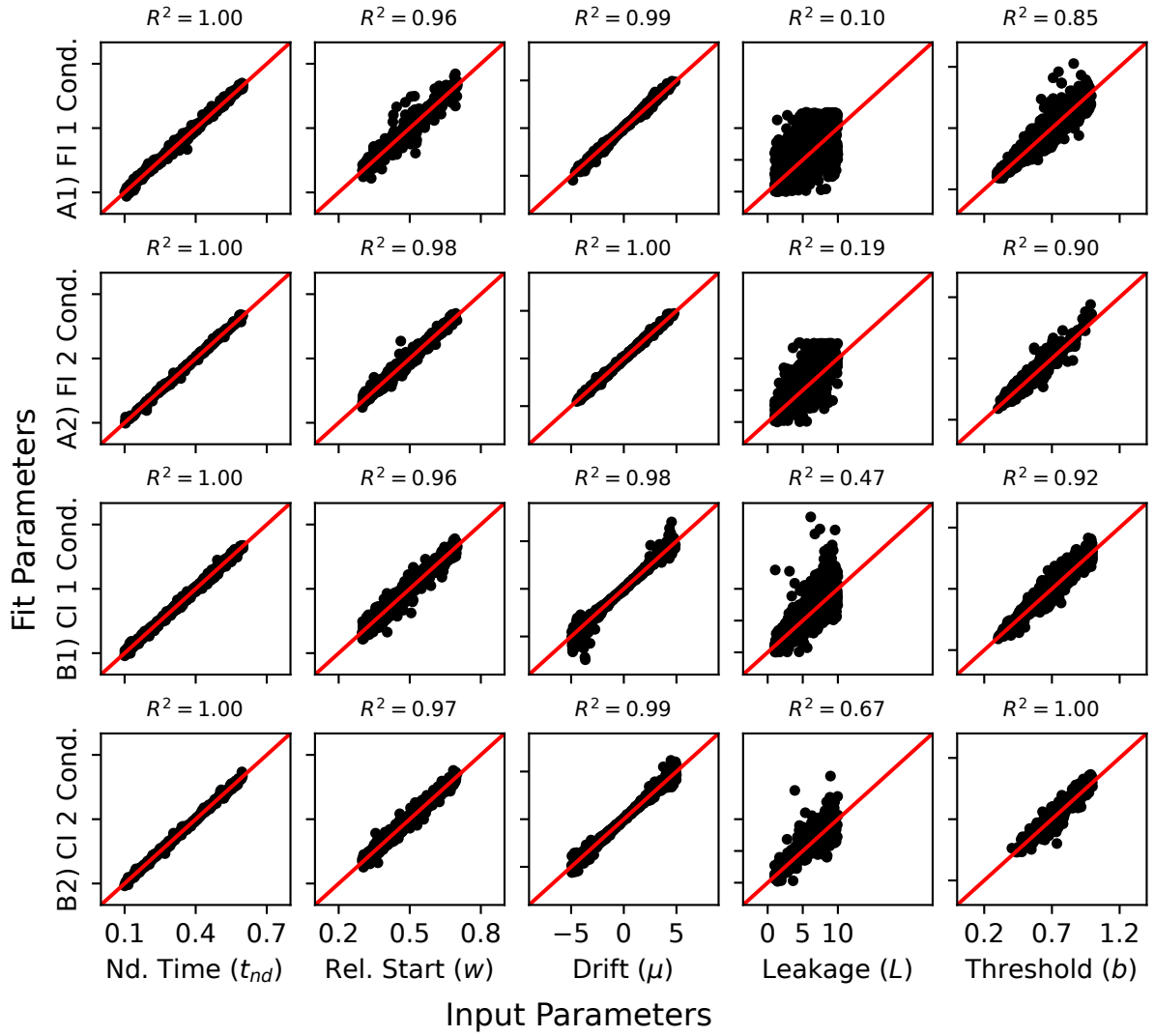


Figure 3. Quality of fit for the leakage model with $N = 1,000$ simulated data points for each numerical experiment. Horizontal axes indicate the true parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot are where the simulation and fit parameters are equal. The correlation coefficient R^2 indicates the quality of fit to the red line. Rows A1 and A2 show the results for the fixed information (FI) experiments for one and two drift rate conditions, respectively. Rows B1 and B2 show the results for the changing information (CI) experiments for one and two drift rate conditions, respectively.

more than one experimental condition as well. Though in Figure 3 we used two drift rate conditions, other condition types, such as caution manipulations, could also prove highly useful.

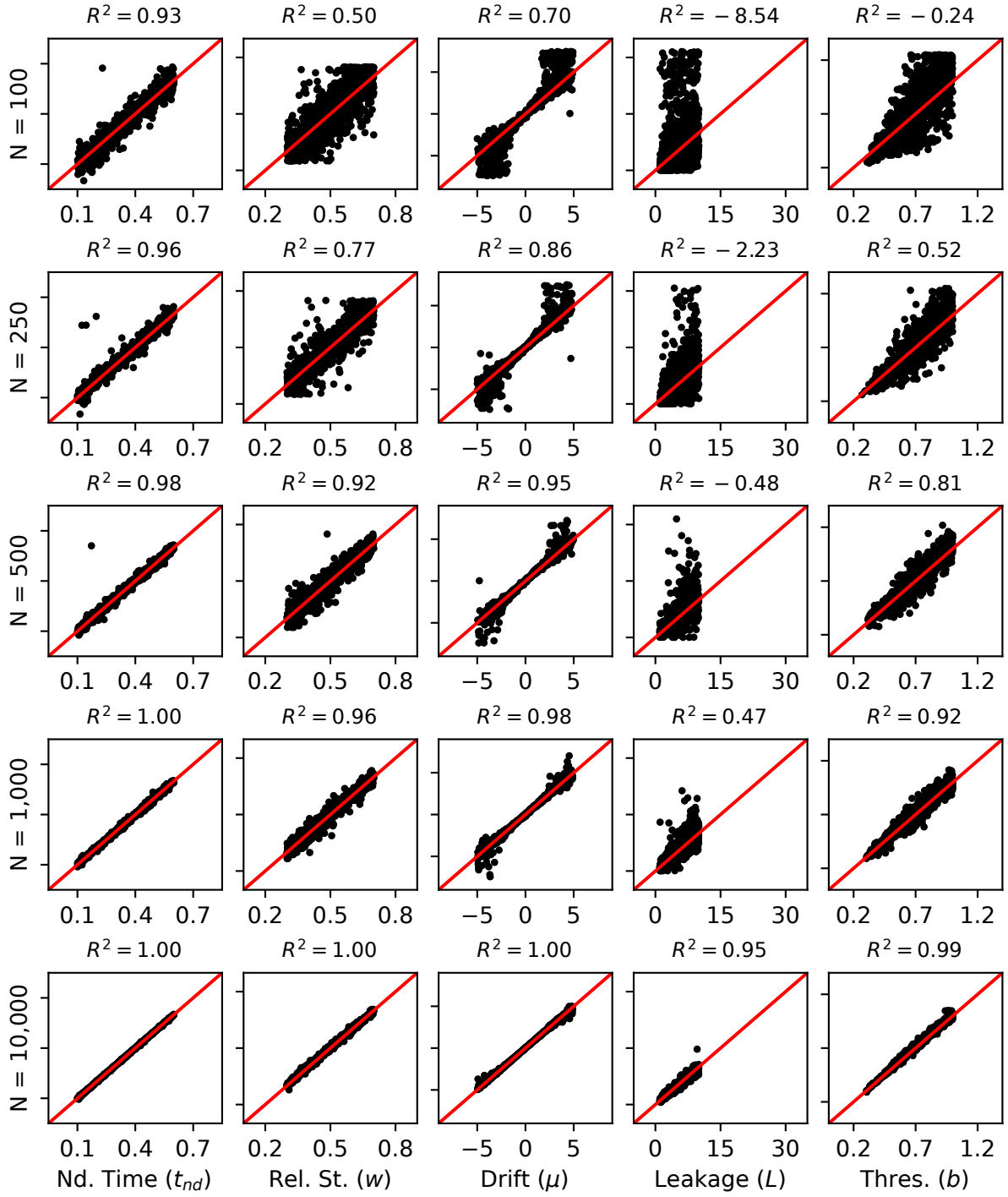


Figure 4. Figure caption on next page.

In summary, the leakage parameter cannot be recovered when FI is used, but can

Figure 4. Quality of fit for the single condition, CI leakage model for $N = 100,\ 250,\ 500,\ 1000,$ and $10,000$ simulated data points. Horizontal axes indicate the true simulated parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot denote the location where the simulation and fit parameters are equal. The correlation coefficients R^2 indicate the quality of fit to the red line.

when CI is used. Additionally, recovery of non-leakage parameters improved when a second drift condition was added to the experiment, with the largest increase in the threshold b. Thus, we recommend that if the leakage parameter is your target, always use CI. It may also be helpful to include multiple flip times into your experiment or another novel modification of the stimulus strength. Further, it is highly recommended to use multiple drift rate conditions (or other types of experimental conditions) to help to mitigate this effect, especially with small sample size N.

Changing thresholds (CT)

We now explore the effect of time changing thresholds and urgency signals on parameter recovery. We seek to answer three main questions. 1) First, are the parameters of the Full Weibull model recoverable. If so, how many data points are necessary? 2) Are alternative threshold models with fewer parameters equally capable of describing data while achieving better parameter recovery? 3) Is it possible to distinguish between time changing threshold models when applied to human data?

Full Weibull recovery. We begin by testing parameter recovery of the Full Weibull model discussed in "Methods". To do so, we follow a similar procedure to that discussed in sections "Methods" and "sDDM". We first generate $1,000$ parameter sets using LHS, filtering out parameter sets which do not produce reasonable choice-RT distributions. Then, we simulate $N = 1,000$ choice-RT data points for each parameter set. Lastly, we use PyBEAM to fit the Full Weibull model to the simulated data. The effect of data set size on this model will be discussed in a subsequent simulation

experiment.

The result of this numerical experiment is shown in Figure 5. Panels A1-A7 display scatter plots of the true versus best fit Full Weibull parameters. Note that we report the $\log_{10}$ of the shape (λ) and scale (κ) parameters. Both λ and κ have large functional parameter ranges, making them difficult to sample directly using Bayesian methods. To address this, PyBEAM samples from the log base 10 of these parameters, so we report their log here as well. Panel A8 displays the threshold error (TE), calculated as,

$$\text{TE} = \frac{\int_{RT_{min}}^{RT_{max}} \mid b_{1,true}(t) - b_{1,fit}(t) \mid dt}{\int_{RT_{min}}^{RT_{max}} b_{1,true}(t)\ dt} \times 100\%, \tag{19}$$

where RT_{min} and RT_{max} are the minimum and maximum choice-RTs from the simulated data set, respectively, and $b_{1,true}(t)$ and $b_{1,fit}(t)$ are the true and best fit upper thresholds, respectively. This metric calculates the area between the true and best fit upper thresholds, then divides it by the area under the true upper threshold to approximate provide a normalized measure of error in the threshold. If TE is high (low), then there is a large (small) distance between the true and best fit thresholds, implying a poor (good) fit. Note that we only integrate between RT_{min} and RT_{max} since the threshold location is relevant only when data is present.

We find that, for $N = 1,000$ data points, recovery is very good for the non-decision time (t_{nd}), relative start point (w), and drift rate (μ). The threshold start parameter (b) is poorly recovered, while the shape (λ), scale (κ), and collapse (c) parameters are completely unrecoverable. The Full Weibull threshold is over-parameterized, causing its parameters to be highly correlated and thus poorly recoverable individually (Gutenkunst et al., 2007; Holmes & Trueblood, 2018). However, even though the threshold parameters are unrecoverable, the threshold itself is highly recoverable, with a TE = 5.2%. For reference, applying this metric to the sDDM threshold gives a median TE = 1.6% for $N = 1,000$ data points.

To more clearly illustrate the results presented in Figure 5, we provide an example fit for the Full Weibull model in Figure 6. This example is chosen to represent an "average" fit, one which contains features common to many simulated data sets. Panels

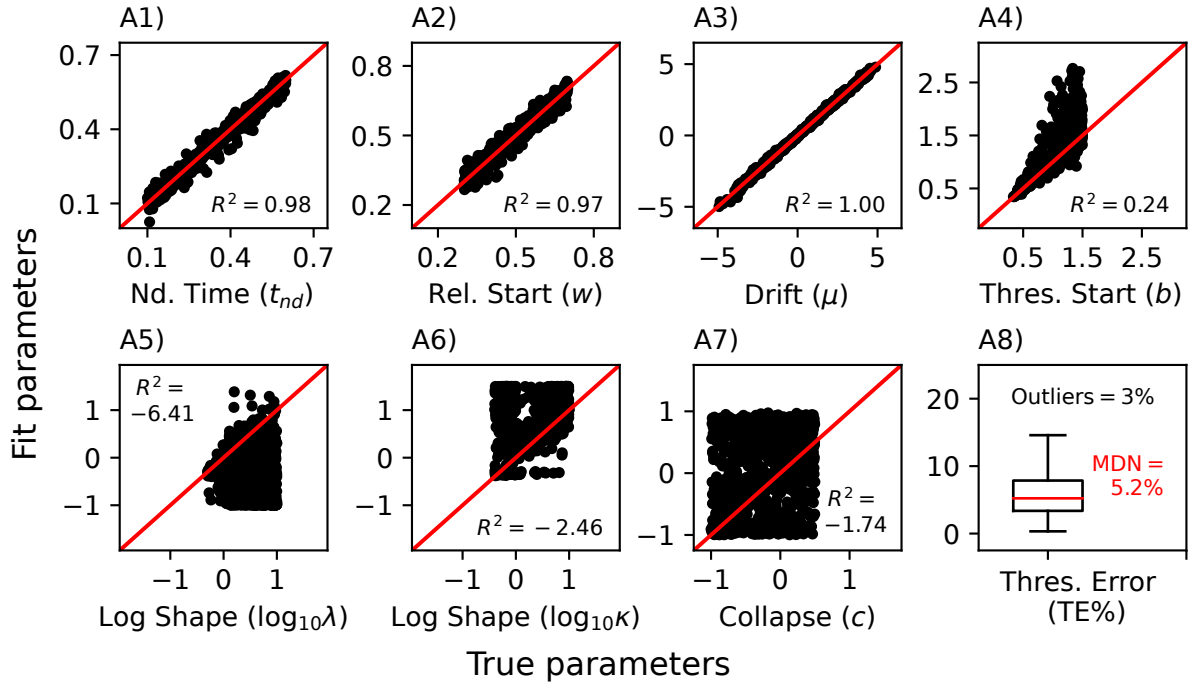


Figure 5. Quality of fit for the Full Weibull model with $N = 1,000$ simulated data points each. For scatter plots, horizontal axes indicate the true parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot are where the simulation and fit parameters are equal. The correlation coefficient R^2 indicates the quality of fit to the red line. The boxplot in panel A8 indicates the percent error between the true and fit threshold (TE) calculated using Equation 19. The percentage of outliers (those parameter sets that do not fit within the box and whisker) are listed above the boxplot, while the median value is noted on the right side of the boxplot in red.

A1-A7 report the posteriors for each Full Weibull parameter. Panel B compares the true threshold to the best fit threshold and reports its TE = 4.5% (approximately the median TE from Figure 5). The dotted black line is RT_{min} for this data set, while RT_{max} is the upper x-limit of the graph. Panel C is a histogram of the choice-RT data, with upper threshold crossings in positive time and lower threshold crossings in negative time. The likelihood functions for the true and best fit parameter sets are plotted, displayed in a solid blue and dotted red line, respectively. The panel also contains the log-likelihood values for the true and best fit parameter sets.

Figure 6 illustrates several features of the Full Weibull model. 1) Even in scenarios where parameters are not recovered, the fit quality is still high and in this case the recovered parameters produce a more favourable log-likelihood than even the generating parameters (Panel C). 2) The threshold parameter indeterminacy does not interfere with recovery of the relative start point or the drift rate. Both fits align closely with their true values, and the posteriors are Normally distributed. 3) However, the non-decision time parameter is underestimated and its posterior is skewed left. While in this case the mis-fit is relatively small ($\sim 10\%$ of true value), this is a systematic problem with the Full Weibull model. While the true and fit thresholds agree well on the range of observed RTs (Panel B), they diverge substantially for $t < RT_{min}$. This results in a non-decision time that is smaller than that of the true parameter set. Thus, caution must be applied when interpreting the non-decision time predicted by the Full Weibull model. If this is of significant concern, recent work has presented alternate ways to constrain the non-decision time through the use of electro-myographical activity (Weindel, Gajdos, Burle, & Alario, 2021) or non-parameterized non-decision functions (Verdonck & Tuerlinckx, 2016).

We last explore the effect of simulated data set size N on the Full Weibull model's recovery quality. For this experiment, we follow the same procedure discussed in "sDDM". We first generate $1,000$ parameter sets using Latin Hypercube Sampling for $N = 100,\ 250,\ 500,\ 1000,$ and $10,000$ points, leading to a total of $5,000$ parameter sets. As with the sDDM, we generate unique parameter sets for each N to eliminate

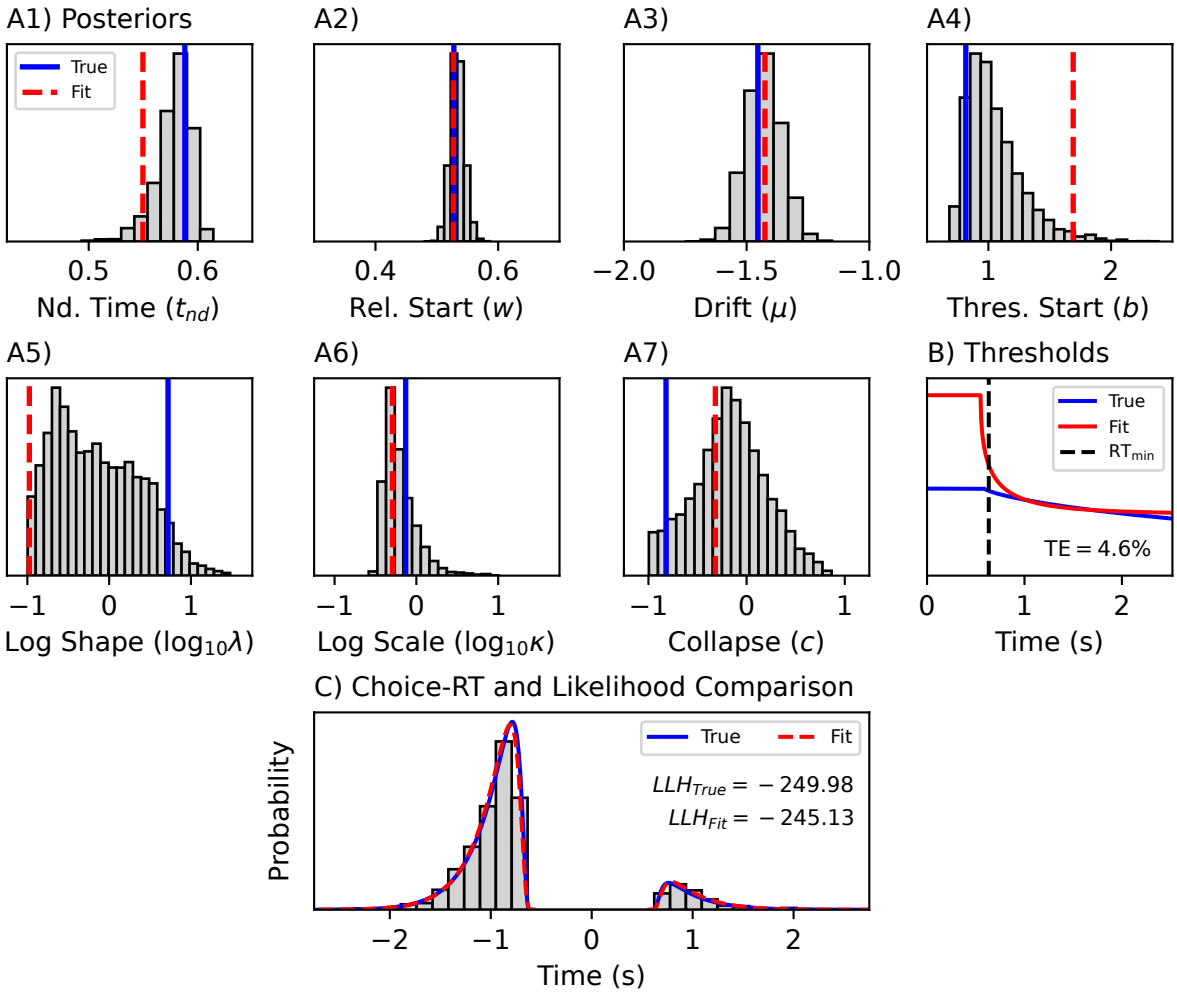
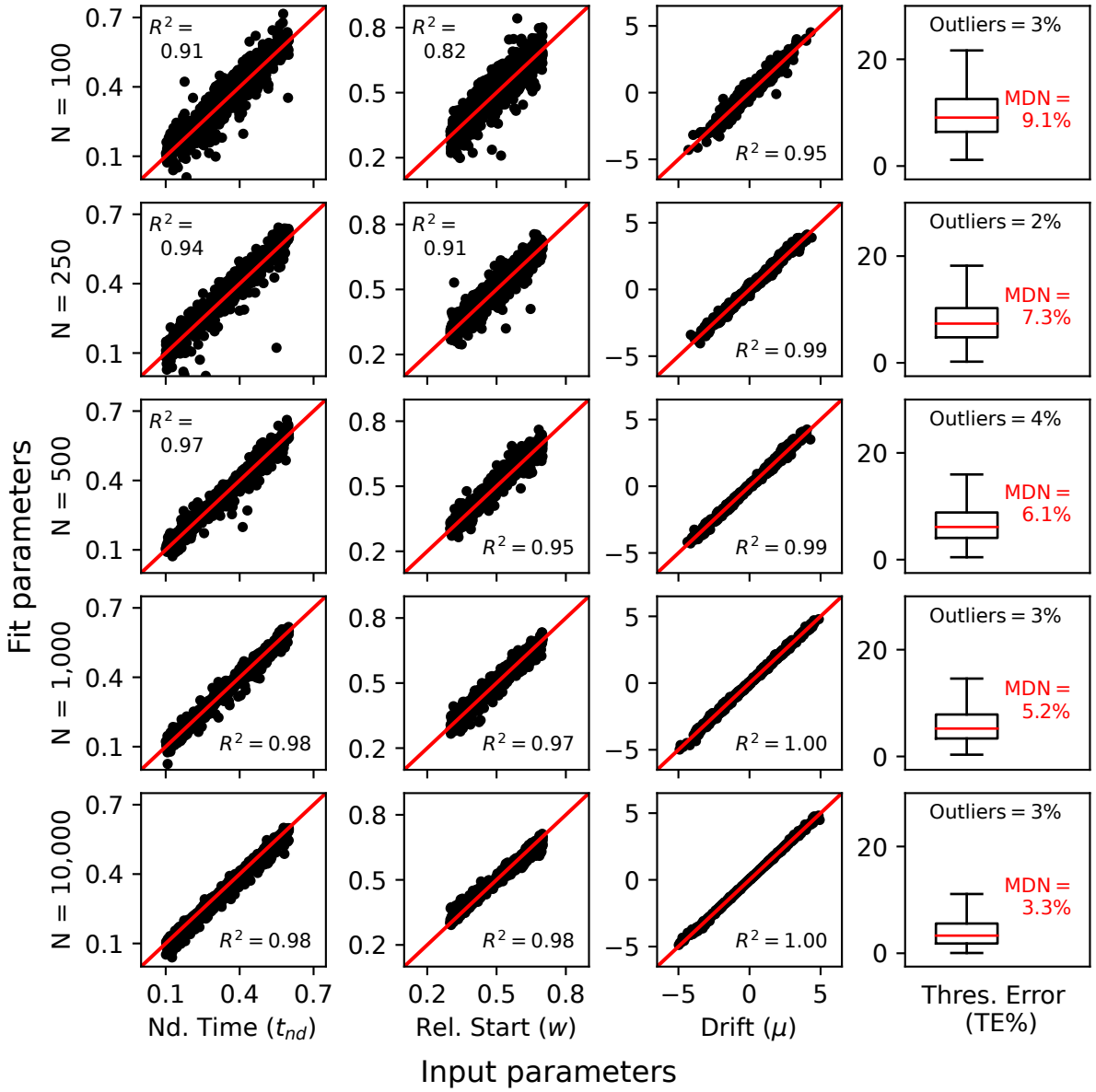


Figure 6. Example fit for a Full Weibull model. Parameter set chosen to be representative of the median behavior. Panel group A show the posteriors for the model parameters, with red lines indicating the best fit parameter based on max log-likelihood and blue line corresponding to the true parameter value. Panel B displays the threshold recovery, with the true threshold in blue, fit threshold in red, and minimum RT value the vertical dashed black line. The TE is displayed on the bottom of the panel. Panel C displays the fit to data, with data displayed in the grey histogram, the true likelihood in the blue line, and the best fit likelihood in the red dashed line. The log-likelihood values of the true and best fit parameters are displayed as well in LLH_{True} and LLH_{Fit}, respectively.

parameter sets which produce few or no error RTs. We then fit the Full Weibull model to each simulated data set. The results of this numerical experiment are shown in Figure 7. Since threshold parameters are not recoverable themselves, we only report the threshold error (TE) and not the parameters themselves.



We find that parameter recovery is generally good, with excellent recovery starting at $N = 500$ data points. Similarly to the Simple DDM, the relative start point w is challenging to recover for small N and has the weakest recovery of all parameters. As discussed earlier in Figure 6 and clearly visible for $N = 10,000$ data points, the non-decision time is systematically underestimated; thus, caution should be applied

Figure 7. Quality of fit for the Full Weibull model for $N = 100,\ 250,\ 500,\ 1000,$ and $10,000$ simulated data points. For scatter plots, horizontal axes indicate the true parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot denote the location where the simulation and fit parameters are equal. The correlation coefficient R^2 indicates the quality of fit to the red line. Boxplots indicate the percent error between the true and fit thresholds TE as calculated using Equation (19) The percentage of outliers are listed above the boxplots, and the median values are listed next to the boxplots in red.

when interpreting the non-decision time in the Full Weibull model. Lastly, threshold recovery is good, with median TE's highest for $N = 100$ at TE= 9.3%, steadily decreasing to TE= 3.3% for $N = 10,000$ data points.

In summary, model recovery for the Full Weibull model is generally good, but there are a few key issues to be aware of. First, parameter recovery of the relative start point and drift rate are excellent, but require substantially more samples than the sDDM. Second, while threshold parameters are not recoverable, the threshold shapes themselves, which are the source of inference, are well recovered. Third, the non-decision time exhibits a small but systematic underestimation.

Weibull threshold versus alternative thresholds. As discussed above, the Full Weibull parameter recovery is generally good, but struggles in a few key areas. To address this, we explore the second question: can simpler thresholds be a better alternative to the Full Weibull model? Can they replicate the dynamics of the Full Weibull model while also giving improved parameter recovery? Additionally, which threshold models are most effective at determining whether a given data set provides evidence of changing versus fixed thresholds?

To address this, we first follow a similar procedure to that of the sDDM and Full Weibull models discussed above. We simulate $N = 1,000$ data points for 1000 randomly generated parameter sets for each of the three alternative models: linear, exponential,

and Reduced Weibull. Next, we fit the threshold models to their respective data sets. This allows us to compare the recoverability of simpler thresholds to that of the Full Weibull threshold. We display a summarized set of results for this in Figure 8, and the full set of results in the Supplementary Information.

We find that the simpler threshold models are more recoverable than the the Full Weibull threshold. The non-decision time t_{nd}, relative start w, and drift rate μ have approximately equal recovery to that of the Full Weibull, but the threshold error is notably better. The Linear model (Supplement Figure 1) has the most recoverable threshold with a median threshold error TE = 3%, while the exponential threshold (Supplement Figure 2) produces the worst recovery with a median TE = 4.6%. The Reduced Weibull (Supplement Figure 3) fell in the middle with a median TE of TE = 4.2%.

We next examine if the alternate threshold models are capable of replicating the Full Weibull threshold's dynamics. We start by simulating $N = 1,000$ data points for the 1000 Full Weibull parameter sets of Figure 5. Then, we fit the four alternate models discussed in "Methods" to this data: the sDDM, the Linear threshold model, the Exponential threshold model, and the Reduced Weibull threshold model. The sDDM fit is used to determine if, on average, data generated from a time changing threshold model can be accurately described using a flat threshold. The other three are the same alternate time changing threshold models used in Figure 8 which are candidates to replace the Full Weibull threshold.

The results of this numerical experiment are shown in Figure 9. In Panel A, The boxplots display the log-likelihood difference LL_D between the best fit parameters of the Full Weibull and alternate threshold models (shown on the horizontal axis). This is calculated as the log-likelihood of the Full Weibull model minus the log-likelihood of the alternate model. Thus, if LL_D is positive, the Full Weibull model is preferred, whereas if the log-likelihood difference is negative, the alternate model is preferred. These figures allow us to determine whether the Full Weibull model is distinguishable from the simpler candidates.

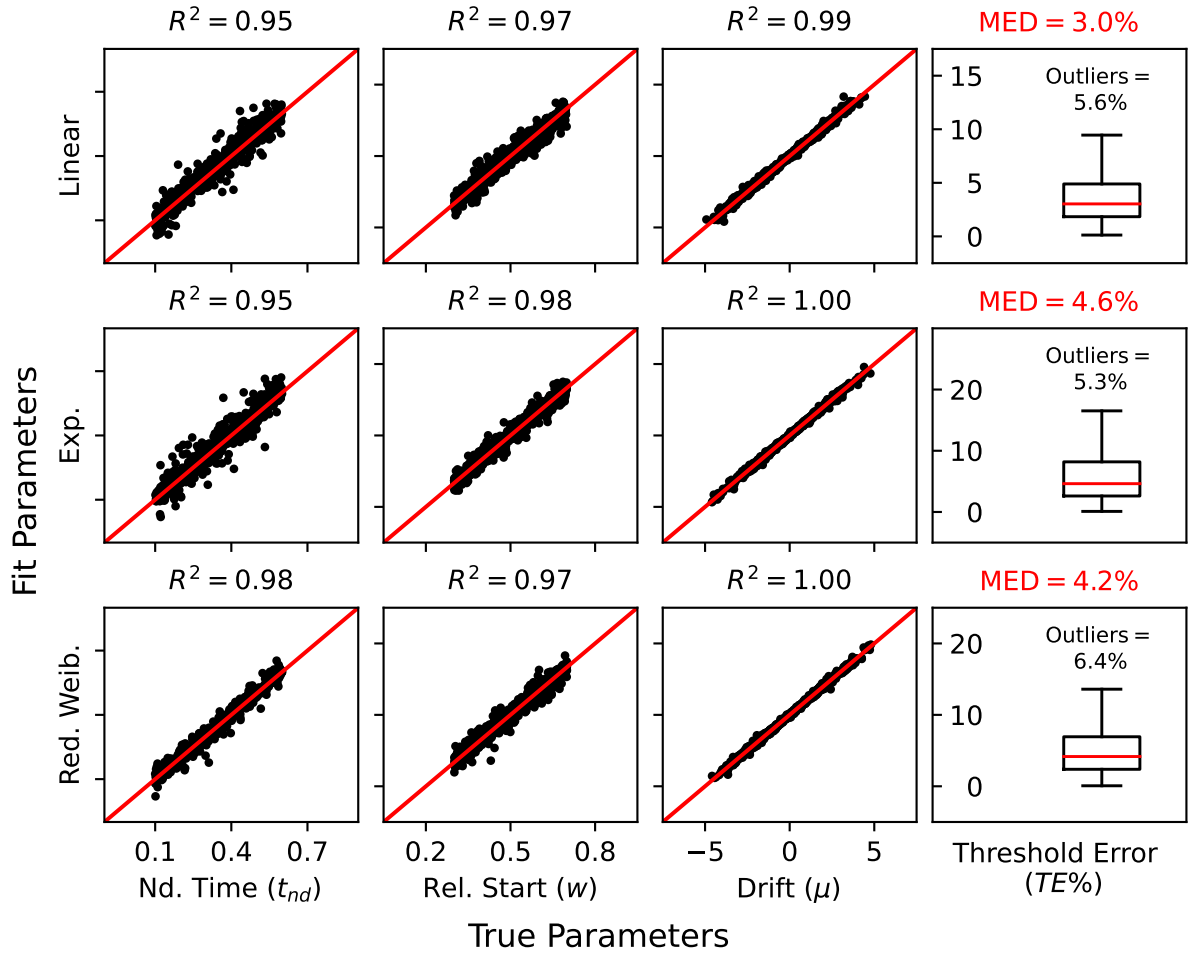


Figure 8. Quality of fit for the alternate thresholds models Linear, Exponential (Exp.), and Reduced Weibull (Red. Weib.) with $N = 1,000$ simulated data points for each. For scatter plots, horizontal axes indicate the true parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot are where the simulation and fit parameters are equal. The correlation coefficient R^2 indicates the quality of fit to the red line. The boxplots in the final rows indicate the percent error between the true and fit threshold (TE) calculated using Equation 19. The percentage of outliers (those parameter sets that do not fit within the box and whisker) are listed above the boxplot, while the median value is noted in red in the panel title.

In Panel B, the boxplots display the LL_D between the alternate model and the sDDM. Unlike Panel A, this is calculated as the log-likelihood of the alternate model minus the log-likelihood of the sDDM, meaning that positive values favor the alternate model (listed on the horizontal axis) over the sDDM. These figures allow use to determine whether each of these models can infer presence of a changing threshold when compared against the sDDM.

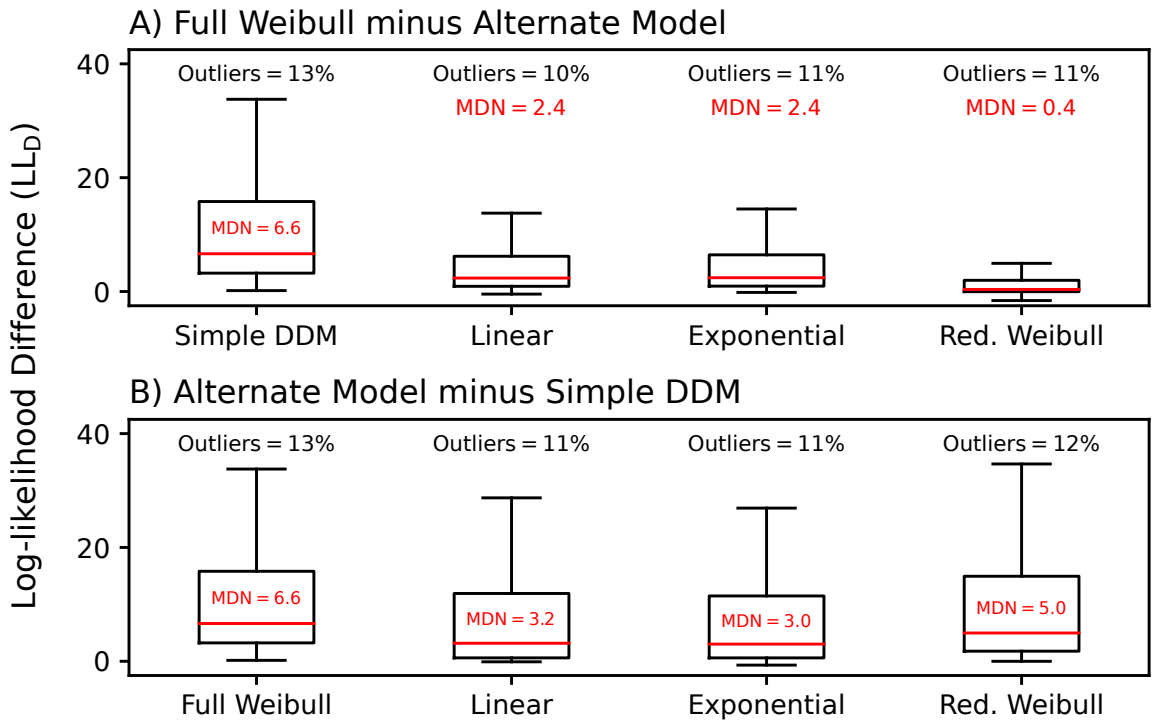


Figure 9. Boxplots of the log-likelihood difference (LL_D) models. A) LL_D between the Full Weibull and alternate models, calculated as max log-likelihood of Full Weibull minus the max log-likelihood of the alternate model, shown on the horizontal axis. If positive, Full Weibull preferred; if negative, alternate model preferred. Outlier percentage and median value displayed near each boxplot. B) LL_D between the sDDM and alternate models, calculated as max log-likelihood of sDDM minus the max log-likelihood of the alternate model, shown on the horizontal axis. If positive, sDDM preferred; if negative, alternate model preferred. Outlier percentage and median value displayed near each boxplot in red.

From Figure 9A, we see that, on average, the sDDM performs the worst of all four

alternate models, with a median $\mathrm{LL_D} = 6.6$. This suggests we cannot approximate a Full Weibull threshold with a flat threshold. The next two models, linear and exponential, perform similarly to each other, with median $\mathrm{LL_D} = 2.4$. The median provides only weak evidence for distinguishing between the alternate and Full Weibull thresholds. The last model, the Reduced Weibull, fits the Full Weibull data best. It has a median log-likelihood difference of $\mathrm{LL_D} = 0.4$, making it nearly indistinguishable from the Full Weibull. This suggests that the Reduced Weibull is a suitable replacement for the Full Weibull model, especially considering its better parameter recoverability.

Panel B tells a similar story. The Full Weibull model shows clear evidence of describing data better than the sDDM, with a median $LL_D = 6.6$. The Linear and Exponential thresholds have only weak evidence for better fitting the data than the sDDM, with medians of approximately $LL_D = 3$. Lastly, the Reduced Weibull model has strong evidence for fitting the data better than the sDDM, with a median $LL_D = 5$. As with Panel A, this suggests that, of all alternate models, the Reduced Weibull is preferred.

We make a brief note about the magnitude of the LL_D observations here, which are on the order of $1 - 10$ in size. From an inference perspective, a LL_D of $5 - 10$ will often be considered as weak or moderate evidence in support of a conclusion. Thus from this perspective, the magnitude of these log-likelihood differences are relatively small. We note however that we are simulating data sets from random parameter sets using the full Weibull threshold. This threshold is capable of producing a wide array of threshold shapes, including flat, exponential, and linear. Thus, it is likely that many of the simulated data sets we use for this experiment are well mimicked by these other thresholds and should not generate substantial LL_D. Thus, the illustrated differences between models are likely averaging results from parameter sets that do illustrate distinctive signatures of changing thresholds with sets that do not.

In summary, the Reduced Weibull is sufficiently flexible to capture most threshold behaviors of the Full Weibull while being parametrically simpler and more easily recoverable. Further, it has nearly the same capacity to detect the presence of time

varying thresholds when compared against the sDDM. While linear and exponential thresholds may have usefulness in some circumstances, based on these analyses, they do not have the same capacity as the full and reduced Weibull models if inferring the presence of time varying thresholds is intended.

Fitting changing threshold models to human data. Overall, our simulated results from Figure 7 and Figure 9 suggest that the best way to distinguish between time changing and flat thresholds is by using the Reduced Weibull threshold. However, since human data is always messier than simulated data, we next ask whether or not it is possible to distinguish between thresholds when applied to human experiments.

In Figure 10, we fit the sDDM and changing threshold models to human choice-RT data Evans, Hawkins, and Brown (2020). This data is comprised of three different random dot kinematogram experiments where participants made decisions about direction of dot motion. Four coherences are used for each: 0%, 5%, 10%, and 40%. In experiment one, 63 participants were instructed to maximize reward rate with a cutoff time of 5 seconds; in experiment two, 71 participants were given a decision deadline of 1.3 seconds; and in the third experiment, one 154 participants were instructed to emphasize decision speed with the same 5 second cutoff of experiment one. On average, Experiment 1 has $N = 381$ data points per subject, Experiment 2 has $N = 348$ per subject, and Experiment 3 has $N = 191$ per subject. Our goal in this work is to fit each data set to models with and without changing thresholds, then compare quality of fit to determine A) if it is possible to distinguish between flat and changing decision thresholds, and B) if it is possible, which thresholds is it possible for? We note that our goal here is not to retest previous conclusions of Evans, Hawkins, and Brown (2020). Rather, we are simply using this as a useful benchmark data set.

We display the results of these fits in Figure 10. Panels A, B, and C contain boxplots for Experiments 1, 2, and 3, respectively, displaying the log-likelihood difference (LL_D) between the sDDM and CT model listed beneath it for each participant. The LL_D values are calculated by taking the max log-likelihood of the changing threshold model minus the max log-likelihood of the sDDM; thus, if LL_D is

positive, the CT model is preferred, and if negative, the sDDM is preferred. Red lines display the median LL_D across all participants, with the value listed above the boxplots.

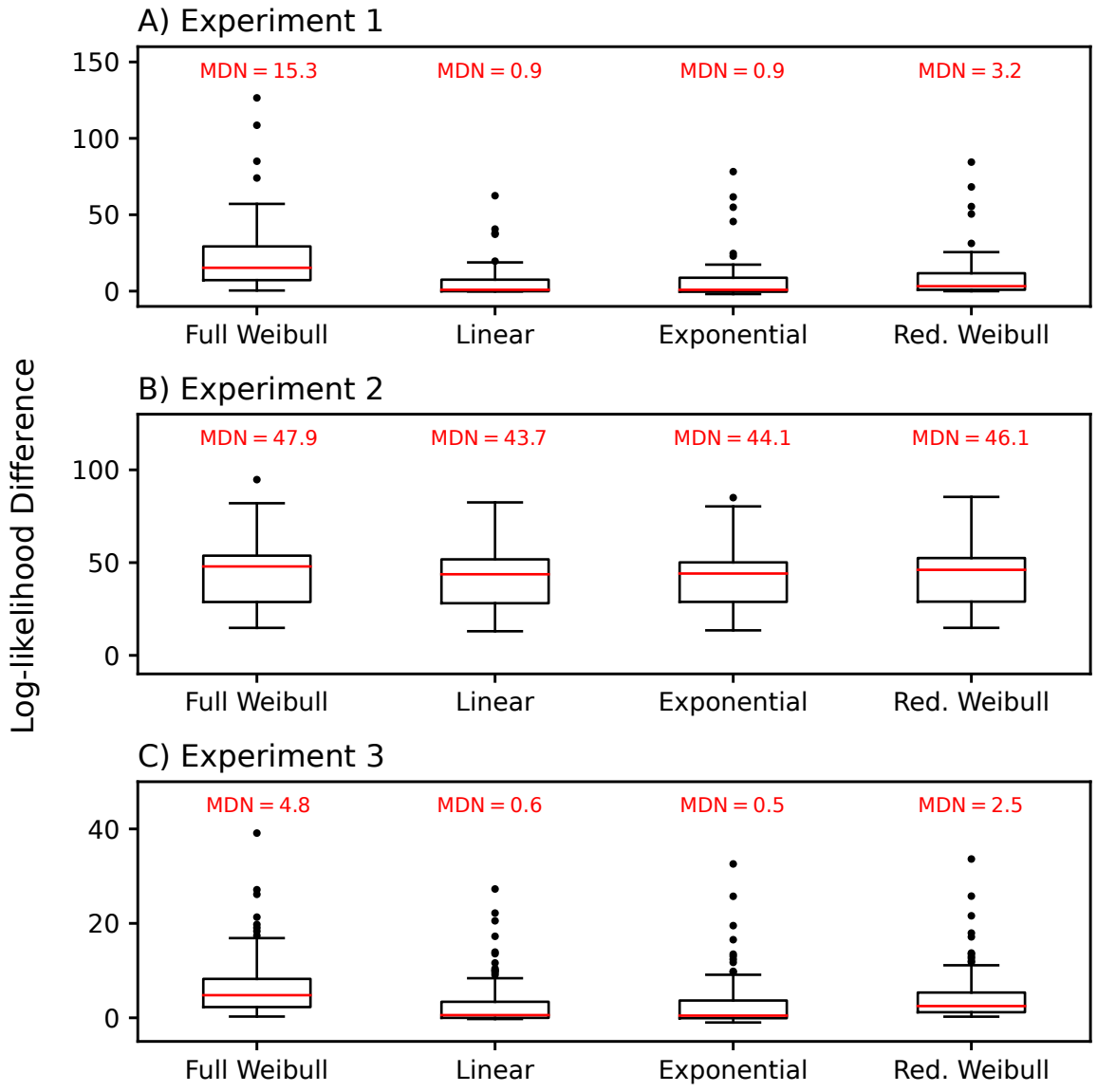


Figure 10. Boxplots of the log-likelihood difference (LL_D) between models when fit to data collected by Evans, Hawkins, and Brown (2020). The LL_D between the sDDM and alternate models, calculated as max log-likelihood of sDDM minus the max log-likelihood of the alternate model, is shown on the horizontal axis. If positive, alternate model preferred; if negative, sDDM preferred. The median value is displayed above each boxplot. Panel A, B, and C show the results for Experiments 1, 2, and 3 from Evans, respectively.

For Experiment 1, we find that, on average, only the Full Weibull threshold has strong evidence for fitting the data better than the sDDM. Both the linear and exponential thresholds show almost no evidence for a better fit, while the Reduced Weibull provides weak evidence for fitting the data better. For Experiment 2, we find that all thresholds have very strong evidence for fitting the data better than the sDDM, with the strongest evidence for the Full Weibull and Reduced Weibull models. Lastly, for Experiment 3, no model shows significant evidence for changing thresholds. The Full and Reduced Weibull thresholds shows very weak evidence for fitting the data better than the sDDM, while the linear and exponential thresholds show functionally zero evidence for it.

In Evans, Hawkins, and Brown (2020) and the refit of their data in Murrow and Holmes (2024) using PyBEAM, the Reduced Weibull threshold was used. Both found that only Experiment 2 showed significant evidence for CT, while Experiments 1 and 3 only did so for some participants. We see similar results in this work, with only Experiment 2 in panel B of Figure 10 showing clear evidence for all CT. Additionally, Experiment 3 closely matches the previous results, with most CT fitting the data similarly to the flat threshold.

Experiment 1 tells a slightly different story. Like the previous fits, the Reduced Weibull shows little evidence for fitting the data better than the sDDM. Additionally, the simpler linear and exponential CTs also fit the data roughly similar to that of the sDDM. The Reduced Weibull fits the data slightly better than the linear and exponential fits since it can model both early and late collpase, whereas the linear and exponential models can only describe early collapse.

The Full Weibull model, however, indicates a clearly better fit when compared to the sDDM. This occurs because the Full Weibull threshold allows collapse anywhere between the initial threshold location b_0 and zero, and the Experiment 1 data benefits greatly from this flexibility. When plotted, the principal threshold type that results from the Full Weibull fits is a threshold that looks similar to a step function. It starts at a constant level b_0, then later rapidly collapses to another constant level somewhere

between zero and b_0.

This suggests strongly that a simple step function will fit the Experiment 1 data as well as the more complex Full Weibull. To demonstrate this, we fit the Evans Experiment 1 data using a smoothed step function as the threshold. The smoothed step function is implemented by using the Full Weibull model and setting the shape parameter $\kappa = 10$. We report the results of this in Figure 11. We find that the step function fits nearly as well as the Full Weibull model and significantly better than the other CTs. Thus, the underlying CT behavior for this is very likely a simple step function, and it is unnecessary to use a complex Weibull threshold to accomplish this much simpler fit.

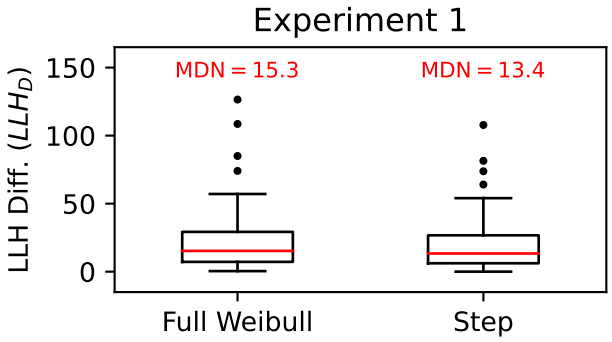


Figure 11. Boxplots of the log-likelihood difference (LL_D) between the sDDM and changing threshold model listed on the horizontal axis when fit to Experiment 1 data collected by Evans, Hawkins, and Brown (2020). The LL_D is between the sDDM and Full Weibull and stepwise models, calculated as max log-likelihood of the sDDM minus the max log-likelihood of the alternate model, and is shown on the horizontal axis. If positive, alternate model preferred; if negative, sDDM preferred. The median value is displayed above each boxplot.

To summarize, when data contains a clear signature of changing thresholds with early collapse (as was shown for Experiment 2), all threshold models are capable of detecting this. If late collapse is present, the Full and Reduced Weibull models are capable of capturing the data. When evidence is more mixed, different threshold models generate different conclusions. In light of this, we recommend that researchers fit

multiple thresholds of varying complexity to their data, then closely analyze their fit quality to best determine if CTs are present. If all threshold types provide evidence for CTs, there is a high probability they are present in that data set. If the Full Weibull and Reduced Weibull only fit the data well, then there is likely a mixture of early and late collapsing subject in your data set. If all thresholds indicate no collapse, no collapse is likely present. And lastly, if different thresholds provide different results, one should be careful not to draw strong conclusions from ambiguous supporting evidence.

Urgency Gating Model (UGM)

We next report the results of the UGM. As discussed in the methods, this model is equivalent to a changing threshold model with the addition of leaky integration. We display the results of our numerical experiments in Figure 12. Plotting conventions are similar to prior simulation studies.

Row A of Figure 12 displays the results for the FI experiment. We find that recovery of only the drift rate is acceptable, while all other parameters experience fair to poor recovery. Row B displays the results of the CI information experiment. Recovery for all parameters is substantially better than that of the FI case with the exception of the relative start point w. We find that, in some cases, the start point is completely unrecoverable in spite of the remaining parameters closely matching their true values.

This inability to recover the start point is an artifact of choosing evidence change times that are too large relative to the typical response time. To demonstrate this, we plot two different data sets in row B of Figure 12: one with large change times in black circles, and one with small change times in cyan triangles. Specifically, the black circles correspond to data sets where the flip time $t_0 > 0.35 \times M_{data}$, where M_{data} is the median of the generated data set. The cyan triangles plot data where $t_0 \leq 0.35 \times M_{data}$. When small flip times are used, recovery is better for all parameters. We display the correlation coefficient for large flip time as R^2 and small flip time as $R^2_{st_0}$. The largest improvement in recovery is present for the relative start point w. The most important improvement is in the urgency ratio (k), which is one of the primary parameters one

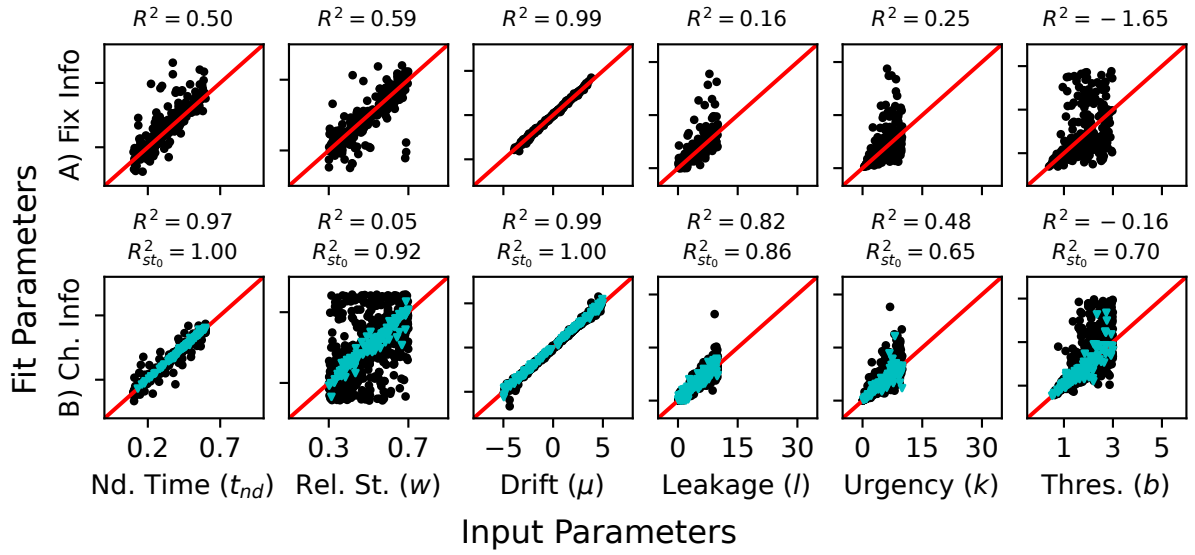


Figure 12. Quality of fit for the UGM. Horizontal axes indicate the true parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. Red lines on each scatter plot indicate where input and fit parameters are equal. Correlation coefficients for the fit of the data to the red line are displayed above each panel. **A)** Fit with fixed information for $N = 1,000$ data points. **B)** Fit with changing information for $N = 1,082$ data points. The black circles are for fits with $t_0 \geq 0.35 * M_{data}$ and have $N = 946$ points, while the cyan triangles are for fits with $t_0 \leq 0.35 * M_{data}$ and have $N = 136$ points. Correlation coefficient R^2 corresponds to the black points, while $R^2_{st_0}$ corresponds to the cyan triangles.

would be interested in for this model. Interestingly, the leakage is well recovered independent of the time at which the evidence change occurs, within the constraints of this test.

In summary, fixed information does not produce good recovery of the UGM parameters. Changing information improves recovery substantially, particularly if the change occurs relatively early relative to typical response times for the task. Though not shown in this section, following the lead of the Leakage model results, it is likely that the introduction of multiple drift rate or caution conditions will further improve recovery of the leakage and urgency parameters.

Equivalence of changing thresholds and urgency models. As discussed in "Methods", Urgency models are motivated by the same theoretical questions which led to the introduction of collapsing thresholds. However, unlike collapsing thresholds, urgency models propose that strategic manipulation exists principally in the choice of $u(t)$. It has long been known that changing thresholds and urgency signals produce similar behavior (Drugowitsch et al., 2012), with more recent work by Trueblood et al. (2021) and Smith and Ratcliff (2022) explicitly demonstrating this. Smith and Ratcliff (2022) showed that they result in mathematically identical likelihood functions under the correct conditions. Specifically, a CT is equivalent to a UGM when the threshold $b_i(t)$ is given by,

$$b_1(t) = -b_2(t) = b_0/u(t), \tag{20}$$

where b_0 is the initial threshold location. One can also do the reverse, and use a UGM to model a CT when Equation (20) is solved for $u(t)$,

$$u(t) = \pm b_0/b(t), \tag{21}$$

where it is positive if the thresholds collapses and negative if the threhsold expands (though this later case is infrequently used).

Thus, when leakage $L = 0$, we can directly compare the results of CTs and UGMs. Specifically, we can determine if the UGM's choice of urgency function is an appropriate choice. Since the UGM is equivalent to a CT, we approach this in the same way as in the previous section. We fit it to the data collected by Evans, Hawkins, and Brown (2020), then compare it to the fits of the other thresholds. So that the models are directly comparable, we set the leakage rate L to zero.

We report the results of these fits in Figure 13. As with Figure 10, the boxplots display the log-likelihood difference LL_D between the best UGM and sDDM fits. If positive, the UGM is preferred, while if negative, the sDDM is preferred. We find that for Experiments 1 and 3, as with the other thresholds there is no evidence for the UGM fitting the data better than the sDDM, fitting similarly to the linear and exponential thresholds. For Experiment 2, as with the other CT models, the UGM shows clear

preference over the sDDM; however, of all the CT models, it has the smallest LL_D, suggesting that it has the least power to determine if a collapsing threshold is present.

Overall, the UGM urgency functions appears equivalent to or worse than the linear and exponential thresholds for detecting collapsing thresholds in human data. This is unsurprising since the UGM "threshold" determined via Equation (20) has a very similar shape to that of the exponential CT, another threshold whose shape and simplicity causes it to struggle at detecting CTs in human data. This suggests that the linear urgency function employed by the UGM may in fact be a poor choice when used for detecting changing thresholds/urgency signals in data, and that more model power may be achieved through the use of different urgency models. These could include the Reduced Weibull "urgency" function from the previous section calculated via Equation (21), or the logistic urgency function of Ditterich (2006).

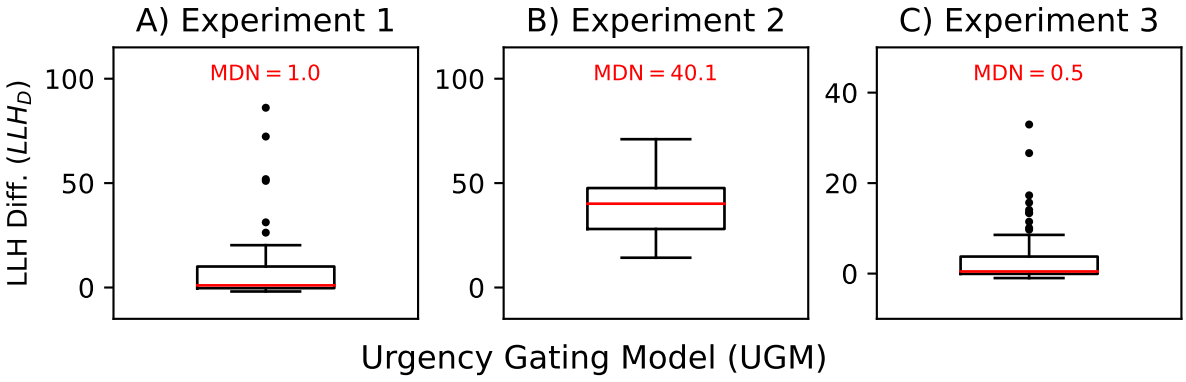


Figure 13. Boxplots of the log-likelihood difference (LL_D) between the UGM and sDDM for the data collected by Evans, Hawkins, and Brown (2020). LL_D between the UGM and sDDM is calculated as max log-likelihood of UGM minus the max log-likelihood of the sDDM. If positive, UGM preferred; if negative, sDDM preferred. Median values displayed above each boxplot. Results for Experiments 1, 2, and 3 are shown in panels A, B, and C respectively.

Diffusion Model of Conflict

We last explore the parameter recovery of the Diffusion Model of Conflict (DMC). Though parameter recovery of the DMC has been studied in the past using simulation

quantile methods (White et al., 2018), our work is the first time the DMC has been studied using Bayesian methods with the full likelihood distribution. We report our results in Figure 3. Simulation study methods and plotting conventions are similar to prior studies. Each simulated data set here included $N = 1,000$ simulated data points.

The primary model feature of interest with the DMC is the time changing drift rate function and its associated parameters. We thus follow a similar procedure used in the time changing threshold section. To assess the ability of the model to recover accurate dynamics from data, we both examine 1) parameter recovery and 2) the quality of recovery of the drift rate function as a function of time. For this second point, we quantify Drift Error, given by,

$$\text{DE} = \frac{\int_{t_{nd}}^{RT_{max}} \left| v_{true}(t) - v_{fit}(t) \right| dt}{\int_{t_{nd}}^{RT_{max}} \left| v_{true}(t) \right| dt} \times 100\%, \tag{22}$$

where v_{true} is the true drift rate and v_{fit} is the best fit drift rate. We integrate from the non-decision time t_{nd} to RT_{max} to capture the drift rate for the entire choice-RT distribution. Similar to the threshold error of Equation (19) introduced in the changing thresholds section, this metric calculates the area between the true and best fit drift rate, then divides it by the area under the true drift rate to provide a normalized measure of error in the drift. If DE is high (low), then there is a large (small) distance between the true and best fit thresholds, implying a poor (good) fit. We use this metric for much the same reason as the changing thresholds. Due to parameter degeneracy in the DMC drift rate, it is possible to fit the drift function without recovering the exact parameters. The DE provides a way to calculate recovery of the drift rate independent of the fit parameters.

Figure 14 contains four rows with different sets of experimental conditions. In row A, only congruent samples are generated, whereas in row B, only incongruent samples are present. For each, we recieve good recovery of the non-decision time t_{nd}, controlled process drift rate μ_c, and threshold b; however, recovery for remaining drift rate parameters is poor. Further, the DE is high, and though slightly better for the incongruent than congruent condition, these give little power to extract meaningful information fromt the data.

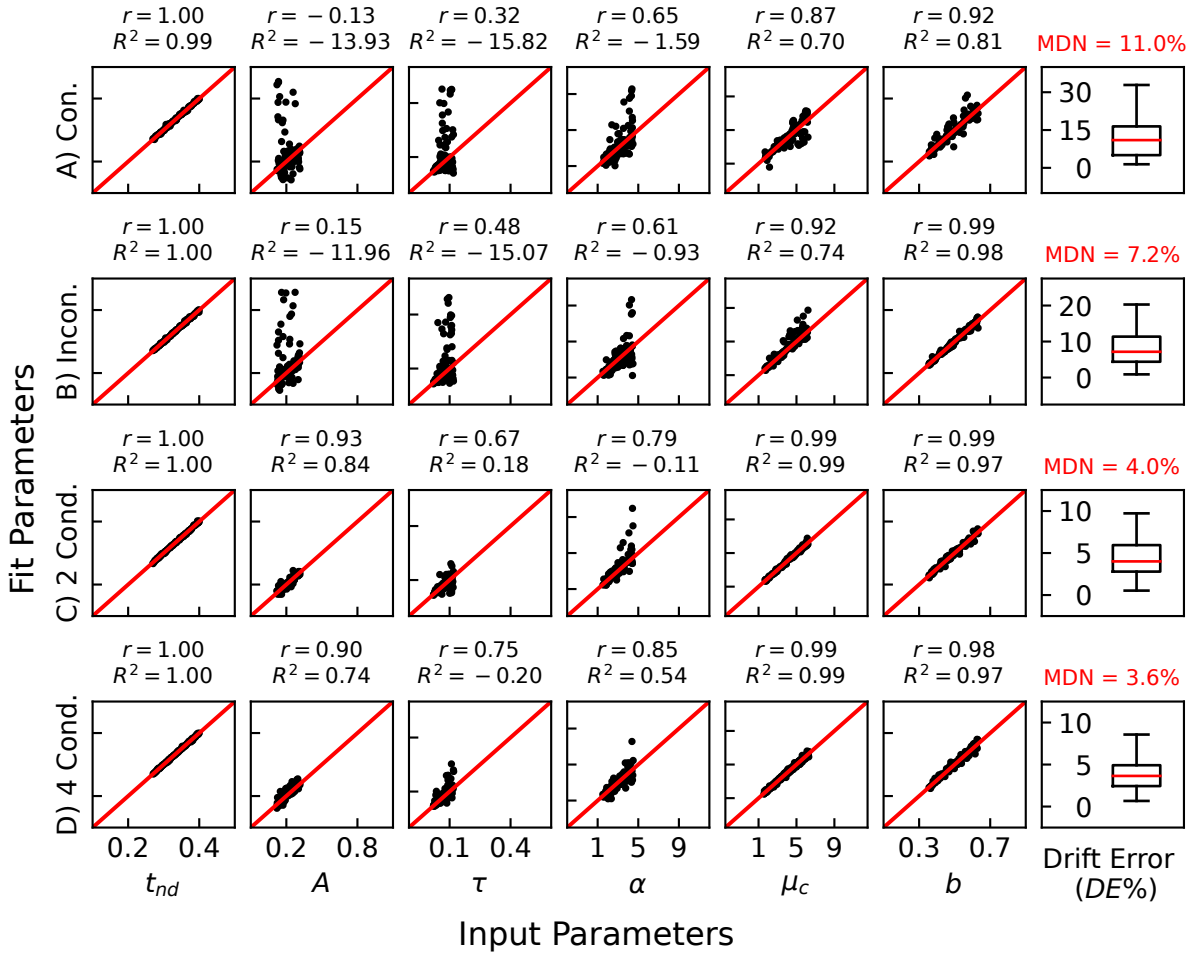


Figure 14. Quality of fit for the DMC with $N = 1,000$ simulated data points each. Horizontal axes indicate the true parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot are where the simulation and fit parameters are equal. The correlation coefficient R^2 indicates the quality of fit to the red line. The box plots in the final column indicate the Drift Error for each experiment, with the median displayed in red in the title. Row A shows results for Congruent data only. Row B shows the results for Incongruent data only. Row C shows the results for half Congruent and half Incogruent data.

Row C displays the results for a slightly more complicated experiment. Here, we have a two condition experiment, one congruent and one incongruent. This corresponds to giving the subject $N = 500$ congruent trials and $N = 500$ incogruent trials as part of the total of $1,000$ simulated data points. We find that mixing both congreuncies into a single experiment provides substantially better recovery, with all parameters receiving some degree of recoverability. Additionally, the DE is very low, suggesting that the recovery of the drift rate function is good.

The last row, D, provides results for a four condition experiment. In this case, we have the same congruent and incongruent setup as row C, but in addition, we have two conditions for the controlled drift rate μ_c. As noted by Ulrich et al. (2015), this is an easy experimental manipulation to make, and an obvious candidate for improving parameter recovery of the DMC. We find that, in general, all parameters fit nearly as well or better, with the largest gains seen in α. Parameter τ fits slightly worse in this data set, but we expect that this is likely due to random variation in the sampled parameter sets. In summary, there may be gains in recovery for this scenario when additional conditions are added, but they are marginal and the interested researcher should determine whether this manipulation provides value for experimental time.

In Figure 15, we provide an example DMC fit from the two condition experiment of row C in Figure 14. In column A, we show the drift rate recovery for the congruent (row 1) and incongruent (row 2) conditions, with the true drift rate in the solid blue line and the recovered drift rate in the dashed red line. We provide the drift error above the curves equaling DE = 4.2%. We specifically chose a fit with DE value approximately equal to the median of row C in Figure 14 to provide context for what that degree of error looks like. In row B, we provide the fit of the likelihood function to the simulated data set. The solid blue line corresponds to the true likelihood function, while the dashed red line corresponds to the best fit likelihood function. The grey histogram is the simulated data. Both the best fit and true likelihood fit the data well, demonstrating that recovery is effective at fitting the input data.

We last display results for a DMC with variable relative start point w in Figure

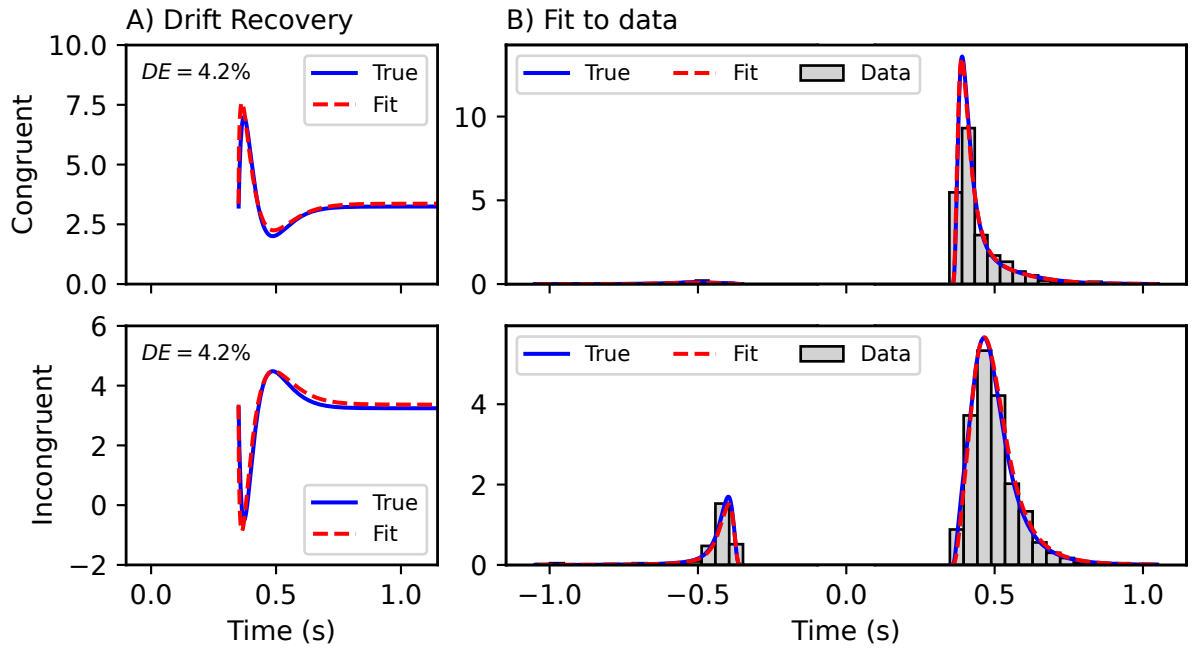


Figure 15. Example fit for the DMC taken from the two condition experiment from row C of Figure 14. Column A displays the drift rate recovery, with the DE displayed above the curves. The true drift rate is the solid blue line, while the fit drift rate is the dashed red line. Column B shows the likelihood functions fit to the choice-RT data (grey histogram). The true likelihood is the solid blue line, while the dashed red line is the fit likelihood function.

16. All work on the DMC to this point has assumed no bias, and thus the relative start parameter is fixed at $w = 0.5$. Here we determine if this parameter is recoverable. We find that the non-decision time t_{nd}, relative start point w, controlled drift rate μ_c, and threshold parameters are all recovered with high accuracy. Recovery of the remaining parameters is however impaired, and the total drift error, though not substantially higher than the fixed start point case, has a larger variance. This level of drift rate recovery error is unlikely to change conclusions drawn from its shape (Figure 15 shows a 4% error and the errors here are 7-10%). Thus, if there is reason to believe a bias may be present, the models recovery characteristics are reasonable with it included.

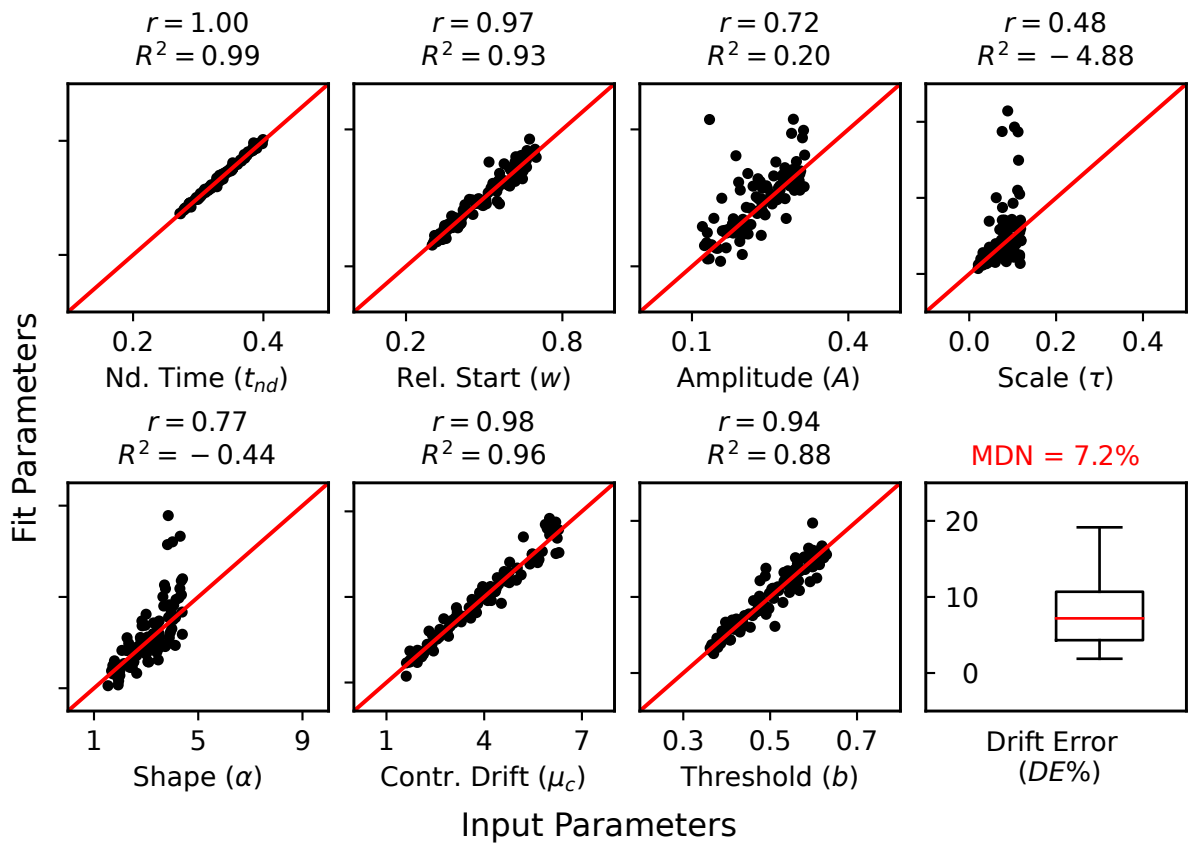


Figure 16. Fit of DMC to $N = 1,000$ data points with variable relative start w. Two conditions, one congruent and one incongruent, are used. The input parameters are displayed on the horizontal axis, while the fit parameters are on the vertical axis. The red line is the location where the true and fit parameters are equal. The r and R^2 values above each panel are the correlation and coefficient of determination for each fit to the red line. The final panel displays the drift error, with the median DE shown above the panel.

Discussion

Evidence Accumulation Models comprise an ever expanding class of models used to explore evidence accumulation and temporal aspects of decision making. In recent years, the use of more complex variants of these models has become more common. However, with this complexity comes added modeling challenges. How well posed are these models relative to the data they are being challenged with? Are model parameters recoverable and therefore interpretable? What experimental structures are sufficient to constrain these models? How can and should one judge the quality of these models mimicry of data? These questions are of paramount importance when working with these more complex models. Researchers have investigated these questions in past research (Boehm et al., 2018; Evans et al., 2019; Evans, Trueblood, & Holmes, 2020; Lerche & Voss, 2016; Trueblood et al., 2021; van Ravenzwaaij & Oberauer, 2009; White et al., 2018). However this research has been a patchwork of investigations with different groups analyzing different models with different modeling approaches. Here, we address this (partially at least) by using a state-of-the-art Bayesian modeling methodology to investigate the methodological properties of a family of commonly used EAMs.

In this work, we considered a variety of EAMs which differed in their assumptions about the drift rate, diffusion rate, and decision threshold behavior. Specifically, we examined four types of models: the Simple DDM, EAMs with leaky integration, EAMs with changing thresholds (including the Urgency Gating Model), and the Diffusion Model of Conflict. While we have investigated each of these models using a unified modeling framework, this is not a one-size fits all approach. Some of these models have different variations and each comes with their own challenges. Thus while the general approach to investigating these is similar, the specific analyses performed and approaches vary to best address each class of models.

This investigation was performed using the Python package PyBEAM, an accurate and efficient method for Bayesian choice–RT modeling of a broad class of binary choice EAMs using the full likelihood function (Murrow & Holmes, 2024). This methodology improves on simulation based methodologies - including Quantile

Maximization (Heathcote et al., 2002; Ratcliff & Tuerlinckx, 2002), Probability Density Approximation (PDA) (Holmes, 2015; Turner & Sederberg, 2014), and distribution summary statistics (Wagenmakers, Van Der Maas, & Grasman, 2007) - and is more powerful than methods that produce point parameter estimates such as PyDDM (Shinn et al., 2020). It thus allows us to provide a more comprehensive analysis of these models properties than prior studies. We also note that, to our knowledge, this is the first investigation of the DMC using Bayesian methods. We briefly summarize the results of investigations into each of these models. Some of these results have been found in prior studies and are acknowledged in the preceding text. Others are, to our knowledge, new observations.

Results from the Simple DDM (sDDM) reflect those from numerous prior investigations (included mainly for comparison) and illustrate that this model has good parameter recovery with as few as $N = 100$ data points. For the leaky integration model which adds leakage L to the sDDM's drift rate, recovery of the leakage parameter was poor when fixed information was used. However, when changing information was used, recovery of the leakage and threshold parameters improved, and improved even more when multiple drift rate conditions were used.

For the changing threshold models, we found that recovery for the relative start and drift rate parameters is always good. The non-decision time parameter is generally recoverable, though the Full Weibull model systematically underestimates it by a small amount. The threshold parameters for the Full and Reduced Weibull models are not recoverable for any simulation size N; however, the threshold shape is recoverable for both models. Thus while structural inferences can be made from these models, one needs to take care when making inferences based on parameters. For the linear and exponential models, recovery of the threshold parameters is much better, with the linear threshold providing the best recovery of all.

Use and interpretation of these different caution / threshold models when applied to human data is more murky. In some cases these different threshold models will lead to similar structural conclusions (presence or absence of time varying caution). In

others they can lead to different conclusions. More details of these results are in the main text. Given this murkiness, extra caution is recommended when using these models. For example, one could use multiple threshold models and examine consistency between results. Interpretation of results using these models will likely require problem specific approaches however and we hope the approach illustrated in this paper may provide a jumping off point to doing so.

While the Urgency Gating Model was originally discussed as a distinct model from other standard EAMs, recent work (Smith & Ratcliff, 2022) has shown that the UGM is a variation on the just discussed changing threshold model. Given this history, we analyze this model both in its own right and as a member of the family of changing threshold models. Our analysis shows that, as with the leakage model, recovery is poor when information is fixed. However, when changing information is used, recovery is excellent and improved further when multiple conditions are used. Further, parameter recovery is best when the time of evidence change is early relative to typical response times, especially if the relative start point is of interest to the experimenter.

Under an appropriate coordinate transformation, the UGM with linear urgency becomes a collapsing threshold model with a particular threshold function. We thus assess this model in comparison to the changing threshold models discussed previously. This model performs similarly to the exponentially decaying threshold model. This is expected since the changing threshold in the transformed UGM behaves like an exponential decay (they look visually very similar). Thus the strengths (parameter identify-ability) and weaknesses (inability to model late changes in thresholds) are shared.

Lastly, we found that recovery for the Diffusion Model of Conflict is good, provided the congruent and incongruent conditions are fit simultaneously. We also found that addition of a second drift rate condition slightly improved recovery of the drift rate, but not to a substantial effect. This is in contrast to prior models where the addition of distinct trial types substantially improved matters. Finally, we found that recovery of the relative start point is possible with this model, though it decreases the

recoverability of the drift rate parameters related to the automatic process.

Overall, this work is, to our knowledge, the most comprehensive analysis of complex EAMs using A) the full choice-RT distributions (using full likelihood functions), and B) Bayesian inference. That said, a model should always be analyzed in the context of the question of interest and data available. Our hope is that this article is more than just a bullet point list of observations. We intend it to illustrate different approaches of analyzing complex choice-RT models in the context in which they may be used, while also providing practical suggestions for future studies which require models of this type.

References

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., ... Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the diffusion decision model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46-75. doi: https://doi.org/10.1016/j.jmp.2018.09.004

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. doi: https://doi.org/10.1016/j.cogpsych.2007.12.002

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends Cogn Sci*, *23*(3), 251–263. doi: https://doi.org/10.1016/j.tics.2018.12.003

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432–459. doi: https://doi.org/10.1037/0033-295x.100.3.432

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*, 693–702. doi: https://doi.org/10.1038/nn.2123

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in changing conditions: The urgency-gating model. *Psychological Review*, *29*(37), 11560–11571. doi: https://doi.org/10.1523/JNEUROSCI.1844-09.2009

Dendauw, E., Evans, N. J., Logan, G. D., Haffen, E., Bennabi, D., Gajdos, T., & Servant, M. (2024). The gated cascade diffusion model: An integrated theory of decision making, motor preparation, and motor execution. *Psychological Review*. (Advance online publication) doi: https://doi.org/10.1037/rev0000464

Ditterich, J. (2006). Stochastic models of decisions about motion direction: Behavior and physiology. *Neural Networks*, *19*, 981–1012. doi: https://doi.org/10.1016/j.neunet.2006.05.042

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A.

(2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neruoscience*, *32*(11), 3612–3628. doi: https://doi.org/10.1523/JNEUROSCI.4010-11.2012

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., ... others (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic bulletin & review*, *26*(4), 1051–1069. doi: https://doi.org/10.3758/s13423-017-1417-2

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149. doi: https://doi.org/10.3758/BF03203267

Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (2018). Refining the law of practice. *Psychological review*, *125*(4), 592–605. doi: https://doi.org/10.1037/rev0000105

Evans, N. J., & Hawkins, G. E. (2019). When humans behave like monkeys: Feedback delays and extensive practice increase the efficiency of speeded decisions. *Cognition*, *184*, 11–18. doi: https://doi.org/10.1016/j.cognition.2018.11.014

Evans, N. J., Hawkins, G. E., & Brown, S. D. (2020). The role of passing time in decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 316–326. doi: https://doi.org/10.1037/xlm0000725

Evans, N. J., Holmes, W. R., & Trueblood, J. S. (2019). Response-time data provide critical constraints on dynamic models of multi-alternative, multi-attribute choice. *Psychon Bull Rev*, *26*(3), 901–933. doi: https://doi.org/10.3758/s13423-018-1557-z

Evans, N. J., Trueblood, J. S., & Holmes, W. R. (2020). A parameter recovery assessment of time-variant models of decision-making. *Behavior Research*, *52*, 193–206. doi: https://doi.org/10.3758/s13428-019-01218-0

Fontanesi, L., Gluth, S., Spektor, M., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychon Bull Rev*, *26*, 1099–1121. doi: https://doi.org/10.3758/s13423-018-1554-2

Forstmann, B. U., van den Wildenberg, W. P., & Ridderinkhof, K. R. (2008). Neural mechanisms, temporal dynamics, and individual differences in interference control. *Journal of Cognitive Neuroscience*. doi: https://doi.org/10.1162/jocn.2008.20122

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLOS Computational Biology*, *3*(10), e189. doi: https://doi.org/10.1371/journal.pcbi.0030189

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E., & Shadlen, M. N. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *The Journal of Neuroscience*, *31*(17), 6339–6352. doi: https://doi.org/10.1523/JNEUROSCI.5613-10.2011

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E. J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *The Journal of Neuroscience*, *35*(6), 2476–2484. doi: https://doi.org/10.1523/JNEUROSCI.2410-14.2015

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, *9*(3), 394–401. doi: https://doi.org/10.3758/BF03196299

Holmes, W. R. (2015). A practical guide to the probability density approximation (pda) with improved implementation and error characterization. *Journal of Mathematical Psychology*, *68*(69), 13–24. doi: https://doi.org/10.1016/j.jmp.2015.08.006

Holmes, W. R., & Trueblood, J. S. (2018). Bayesian analysis of the piecewise diffusion decision model. *Behavior Research Methods*, *50*(2), 730–743. doi: https://doi.org/10.3758/s13428-017-0901-y

Kelly, S. P., & O'Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. *Journal of Neuroscience*, *33*(50), 19434–19441. doi: https://doi.org/10.1523/jneurosci.3355-13.2013

Kloeden, P. E., & Platen, E. (1992). *Numerical solutions for stochastic differential*

equations. Berlin, Germany: Springer.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298. doi: https://doi.org/10.1038/nn.2635

Lerche, V., Christmann, U., & Voss, A. (2018). Impact of context information on metaphor elaboration a diffusion model study. *Experimental Psychology*, *65*(6), 370–384. doi: https://doi.org/10.1027/1618-3169/a000422

Lerche, V., & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in Psychology*, *7*, 1324. doi: https://doi.org/10.3389/fpsyg.2016.01324

Luce, R. D. (1991). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Malhotra, G., Leslie, D. S., Ludwig, C. J. H., & Bogacz, R. (2018). Time-varying decision boundaries: Insights from optimality analysis. *Psychonomic Bulletin & Review*, *25*(3), 971–996. doi: https://doi.org/10.3758/s13423-017-1340-6

Murrow, M., & Holmes, W. (2024). Pybeam: A bayesian approach to parameter inference for a wide class of binary evidence accumulation models. *Behavior Research*, *56*, 2636–2656. doi: https://doi.org/10.3758/s13428-023-02162-w

Nosofsky, R. M., Little, D. R., Donkin, D., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological review*, *118*(2), 280–315. doi: https://doi.org/10.1037/a0022494

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological review*, *104*(2), 266–300. doi: https://doi.org/10.1037/0033-295x.104.2.266

Osth, A. F., & Farrell, S. (2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological review*, *126*(4), 578–609. doi: https://doi.org/10.1037/rev0000149

Palestro, J. J., Weichart, E., Sederberg, P. B., & Turner, B. M. (2018). Some task demands induce collapsing bounds: Evidence from a behavioral analysis.

Psychonomic Bulletin and Review, *25*, 1225–1248. doi: https://doi.org/10.3758/s13423-018-1479-9

Philiastides, M. G., Heekeren, H. R., & Sajda, P. (2014). Human scalp potentials reflect a mixture of decision-related signals during perceptual choices. *Journal of Neuroscience*, *34*(50), 16877–16889. doi: https://doi.org/10.1523/jneurosci.3012-14.2014

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59–108. doi: https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. doi: https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. psychological science. *Psychological Science*, *9*(5), 347–356. doi: https://doi.org/10.1111/1467-9280.00067

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. doi: https://doi.org/10.1037/0882-7974.16.2.323

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychon Bull Rev*, *9*(3), 438–481. doi: https://doi.org/10.3758/BF03196302

Ratcliff, R., Zandt, T. V., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300. doi: https://doi.org/10.1037/0033-295x.106.2.261

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, *2*(55). doi: https://doi.org/10.7717/peerj-cs.55

Servant, M., White, C., Montagnini, A., & Burle, B. (2015). Using covert response activation to test latent assumptions of formal decisionmaking models in humans. *Journal of Neuroscience*, *35*(28), 10371–10385. doi:

https://doi.org/10.1523/jneurosci.0078-15.2015

Servant, M., White, C., Montagnini, A., & Burle, B. (2016). Linking theoretical decision-making mechanisms in the simon task with electrophysiological data: A model-based neuroscience study in humans. *Journal of Cognitive Neuroscience*, *28*(10), 1501–1521. doi: https://doi.org/10.1162/jocn_a_00989

Shinn, M., Lam, N. H., & Murray, J. D. (2020). A flexible framework for simulating and fitting generalized drift-diffusion models. *eLife*, *9*. doi: https://doi.org/10.7554/eLife.56938

Simon, J. R., & Small, J., A. M. (1969). Processing auditory information: Interference from an irrelevant cue. *Journal of Applied Psychology*, *53*(5), 433–435. doi: https://doi.org/10.1037/h0028034

Smith, P. L. (1995). Psychophysically principled models of visual simple reaction time. *Psychological Review*, *102*(3), 567–593. doi: https://doi.org/10.1037/0033-295X.102.3.567

Smith, P. L., & Ratcliff, R. (2022). Modeling evidence accumulation decision processes using integral equations: Urgency-gating and collapsing boundaries. *Psychological Review*, *129*(2), 235–267. doi: https://doi.org/10.1037/rev0000301

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. doi: https://doi.org/10.1037/h0054651

Swensson, R. G. (1972). The elusive tradeoff: Speed vs accuracy in visual discrimination tasks. *Perception & Psychophysics*, *12*(1), 16–32. doi: https://doi.org/10.3758/BF03212837

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: theory and experimental support. *Journal of neurophysiology*, *108*(11), 2912–2930. doi: https://doi.org/10.1152/jn.01071.2011

Trueblood, J. S., Heathcote, A., Evans, N. J., & Holmes, W. R. (2021). Urgency, leakage, and the relative nature of information processing in decision-making. *Psychological Review*, *128*(1), 160–186. doi: https://doi.org/10.1037/rev0000255

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. doi: https://doi.org/10.1016/j.neuroimage.2013.01.048

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of eeg, fmri, and behavioral data. *Neuroimage*, *128*, 96–115. doi: https://doi.org/10.1016/j.neuroimage.2015.12.030

Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin & Review*, *21*(2), 227–250. doi: https://doi.org/10.3758/s13423-013-0530-0

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. psychological review. *Psychological Review*, *122*(2), 312–336. doi: https://doi.org/10.1016/j.tics.2018.12.003

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology*, *78*, 148-174. doi: https://doi.org/10.1016/j.cogpsych.2015.02.005

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. doi: https://doi.org/10.1037/0033-295x.108.3.550

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: Ez, fast-dm, and dmat. *Journal of Mathematical Psychology*, *53*(6), 463–473. doi: https://doi.org/10.1016/j.jmp.2009.09.004

van Wouwe, N. C., Kanoff, K. E., Claassen, D. O., Spears, C. A., Neimat, J., van den Wildenberg, W. P., & Wylie, S. A. (2016). Dissociable effects of dopamine on the initial capture and the reactive inhibition of impulsive actions in parkinson's disease. *Journal of Cognitive Neuroscience*, *28*(5), 710–723. doi: https://doi.org/10.1162/jocn_a_00930

Verdonck, S., & Tuerlinckx, F. (2016). Factoring out nondecision time in choice reaction time data: Theory and implications. *Psychological Review*, *123*(2), 208–218. doi: https://doi.org/10.1037/rev0000019

Voskuilen, C., Ratcliff, R., & Smith, P. L. (2016). Comparing fixed and collapsing boundary versions of the diffusion model. *Journal of Mathematical Psychology*, *73*, 59–79. doi: https://doi.org/10.1016/j.jmp.2016.04.008

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*, 767-775. doi: https://doi.org/10.3758/BF03192967

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. doi: https://doi.org/10.1016/j.jml.2007.04.006

Wagenmakers, E.-J., Van Der Maas, H. L., & Grasman, R. P. (2007). An ez-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22. doi: https://doi.org/10.3758/BF03194023

Weindel, G., Gajdos, T., Burle, B., & Alario, F.-X. (2021). The decisive role of non-decision time for interpreting the parameters of decision making models. *PsyArXiv*. doi: https://doi.org/10.31234/osf.io/gewb3

White, C. N., Congdon, E., Mumford, J. A., Karlsgodt, K. H., Sabb, F. W., Freimer, N. B., ... Poldrack, R. A. (2014). Decomposing decision components in the stop-signal task: a model-based approach to individual differences in inhibitory control. *Journal of Cognitive Neuroscience*, *26*(8), 1601–1614. doi: https://doi.org/10.1162/jocn_a_00567

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Anxiety enhances threat processing without competition among multiple inputs: A diffusion model analysis. *Emotion*, *10*(5), 662–677. doi: https://doi.org/10.1037/a0019474

White, C. N., Servant, M., & Logan, G. (2018). Testing the validity of conflict drift-diffusion models for use in estimating cognitive processes: A parameter-recovery study. *Psychonomic Bulletin & Review*, *25*, 286–301. doi:

https://doi.org/10.3758/s13423-017-1271-2

Wieschen, E. M., Makani, A., Radev, S. T., Voss, A., & Spaniol, J. (2023). Age-related differences in decision-making: Evidence accumulation is more gradual in older age. *Experimental Aging Research*, 1–13. doi: https://doi.org/10.1080/0361073X.2023.2241333

Supplementary Information.

Matthew Murrow

Department of Physics and Astronomy, Vanderbilt University, PMB 401807, 2301 Vanderbilt Place, Nashville, TN 37240

William R. Holmes

Cognitive Science Program and Department of Mathematics, Indiana University Bloomington, 1101 E. Tenth Street, Bloomington, IN 47405, USA

Supplementary Information.

Changing thresholds

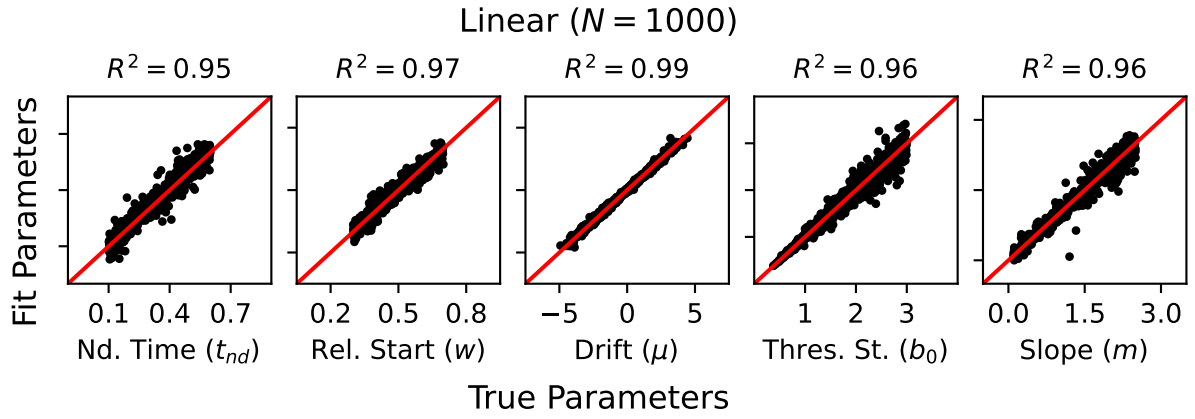


Figure 1. Quality of fit for the linear CT for $N = 1,000$ simulated data points. Horizontal axes indicate the true simulated parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot denote the location where the simulation and fit parameters are equal. The correlation coefficients R^2 above each panel indicate the quality of fit to the red line.

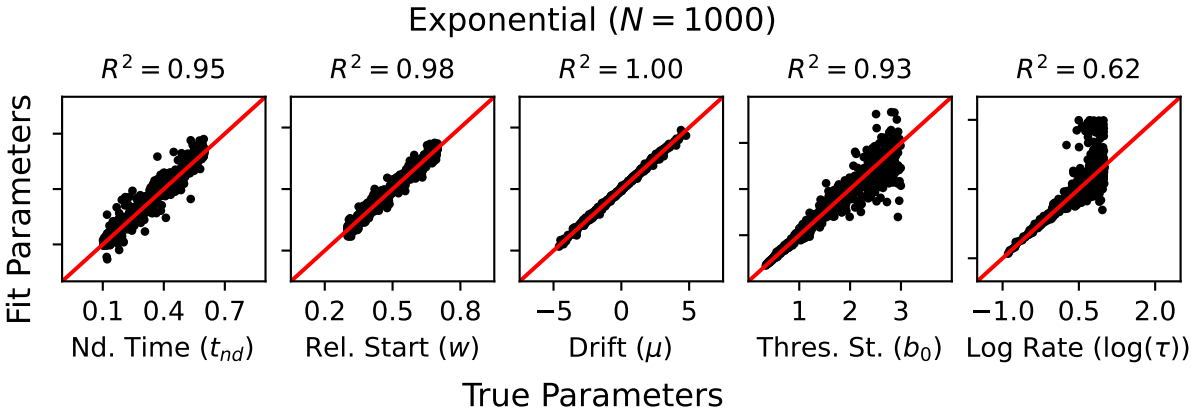


Figure 2. Quality of fit for the exponential CT for $N = 1,000$ simulated data points. Horizontal axes indicate the true simulated parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot denote the location where the simulation and fit parameters are equal. The correlation coefficients R^2 above each panel indicate the quality of fit to the red line.

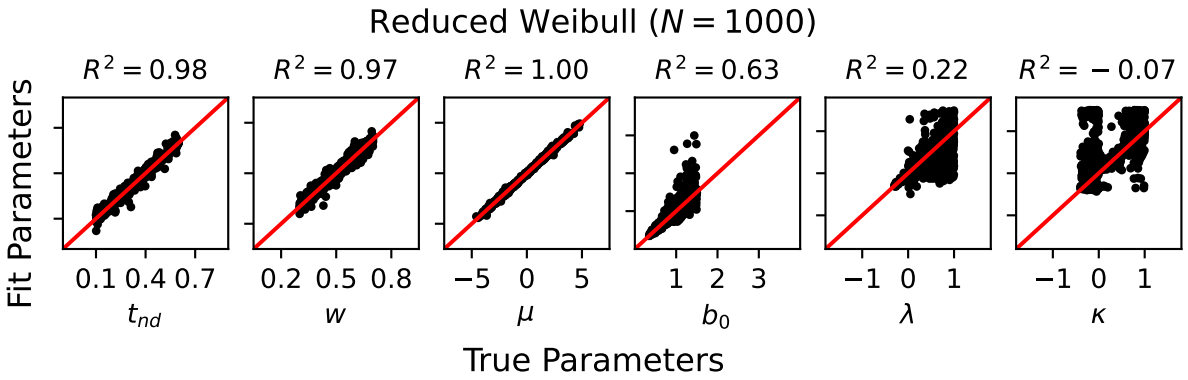


Figure 3. Quality of fit for the exponential CT for $N = 1,000$ simulated data points. Horizontal axes indicate the true simulated parameter values, while the vertical axes indicate the parameter values which best fit the simulated data. The red lines on each scatter plot denote the location where the simulation and fit parameters are equal. The correlation coefficients R^2 above each panel indicate the quality of fit to the red line.