
**Doctoral thesis submitted to
the Faculty of Behavioural and Cultural Studies
Heidelberg University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy (Dr. phil.)
in Psychology**

Title of the publication-based thesis
Unraveling the Dynamics in Cognition with Cognitive Process Models

presented by
Lukas Schumacher, M.Sc.

year of submission
2024

Dean: Prof. Dr. Guido Sprenger
Advisors: Prof. Dr. Andreas Voss
Prof. Dr. Rolf Ulrich

CONTENTS

1	INTRODUCTION	3
1.1	Contributions	4
1.2	List of Scientific Articles of the Cumulative Dissertation	5
2	COGNITIVE PROCESS MODELS	7
2.1	Simulating the Mind	8
2.2	Inferring the Mind	9
2.3	Diffusion Decision Model	12
3	ACCOUNTS FOR DYNAMICS IN COGNITION	15
3.1	Stationary variability	17
3.2	Trial Binning	17
3.3	Regression Approach	18
3.4	Frontend-Backend Model	19
4	DYNAMICS IN DURATION DISCRIMINATION (MANUSCRIPT I)	21
5	SUPERSTATISTICS	27
5.1	Neural Superstatistics (Manuscript II)	29
5.1.1	Amortized Bayesian Inference	30
5.1.2	Benchmark Studies	32
5.1.3	Simulation Study	32
5.1.4	Human Data Application	33
5.1.5	Intermezzo	35
5.2	Validation and Comparison of Dynamic Cognitive Models (Manuscript III)	35
6	GENERAL DISCUSSION	41
6.1	Outlook	47
6.2	Concluding Remarks	52
	BIBLIOGRAPHY	55

Contents

DECLARATION IN ACCORDANCE TO § 8 (1) C) AND D) OF THE DOCTORAL DEGREE REG- ULATION OF THE FACULTY	69
APPENDIX A1 - MANUSCRIPT I	71
APPENDIX B1 - MANUSCRIPT II	91
APPENDIX C1 - MANUSCRIPT III	155

ACKNOWLEDGEMENTS

First, I express my gratitude to my PhD supervisor, Andreas Voss, for providing me with an optimal environment for my academic growth. You always gave me the freedom, trust, and autonomy to explore my interests freely and at the same time advice whenever needed.

On my scientific journey, I had the fortune to meet remarkable individuals who not only influenced my passion for science but also played pivotal roles in helping me achieve my goals. When I started studying psychology, I did not know much about science or about the potential academic career paths. Andrew Ellis greatly inspired me to pursue a PhD and, more specifically, introduced me to cognitive modeling and Bayesian statistics.

The challenging period of arriving in Heidelberg amidst the COVID-19 pandemic was alleviated by the presence of one of the most brilliant and delightful individuals I have had the pleasure to know – Stefan Radev. I will never forget the magical times we spent together in our office and around Heidelberg. You helped me grow both intellectually and personally. Thank you!

Living in Heidelberg has blessed me with many friendships, including Kathrin Sadus, Milena Marx, Wiebke Hemmin, Nils Brandenstein, Fabian Gerstner, Thorben Loring, and Thomas Gerhardt. I extend my heartfelt thanks to all of you for enriching the past few years. Gratitude also goes out to the members of our lab – Lasse Elsemüller, Mischa von Krause, Nicolas Schneider, Annika Stump, Tugba Hato, Shanqing Gao, and Marie Hunsmann.

Being a part of the graduate school *Statistical Modeling in Psychology* (SMiP) was a privilege, offering enriching workshops and retreats and providing opportunities to connect with esteemed researchers. Special thanks to Anna Neumer, Julia Liss, Johanna Höhs, Martin Schnuerch, Julius Fenn, Tobias Rebholz, and Emre Alagöz for all the great moments during and after workshops or retreats. I would also like to thank Anke Söllner for always providing help.

I extend my appreciation to my closest friends – Livio Hardegger, Noé Weigl, Vera-Sophia Pelozzi, Elisa Rostin, Andri Tuor, and Lukas Huber – for consistently having my back despite the distance that separates us. My gratitude also extends to my parents, whose unwavering belief in me and constant support have been instrumental. Lastly, I want to express my deepest thanks to my partner, Stephanie Zintel, for all the incredible moments we have shared so far and for everything you have done for me. Without you, this dissertation would not be what it is.

1 INTRODUCTION

I have yet to see any problem, however complicated, which, when looked at it the right way, did not become still more complicated.

— Poul Anderson

Picture yourself immersed in the compilation of a review article on a research topic close to your heart. You are faced with the cumbersome task of screening over a thousand articles that might end up in your review. Beginning with the process of going through them one by one, you are deciding whether a specific study matches the topic you are writing about or not.

In the field of cognitive science, researchers are interested in gaining a deeper understanding of cognitive processes like the one outlined above. On their mission to make sense of behavior and the underlying cognitive processes, mathematical models have become an important tool. One of the primary aims of these models is to describe how underlying *cognitive constructs* connect to model parameters and to specify how these generate behavior.

In the decision-making process initially described, the behavior consists of two key measurable variables: was your categorization “correct” and how long did you take to reach a decision? Among the cognitive factors that influence these behavioral variables are, for instance, processing speed, reading time, decision caution, and potential biases. In a cognitive process model for such a decision process, one maps the aforementioned cognitive constructs to model parameters and describes how these generate the two behavioral variables.

There are hundreds if not thousands of studies that have focused on developing and applying such *cognitive process models*. However, what most of them have in common is that they neglect the dynamic nature of the cognitive factors affecting the behavior. Let us again consider the initially described scenario. Based on the repeated decisions on the inclusion of an article, we would like to infer your cognitive factors with a cognitive process model. Traditionally, cognitive scientists would simply fit such a model to all your decisions and obtain a single estimate for each of these factors, for example for your information processing speed and your reading time.

However, in our scenario, we can think of many factors that would lead to changes in the cognitive constructs during your task of repeatedly screening articles. Initially, you meticulously examine each study, resulting in many correct categorizations of the articles. As time progresses, the repetitive nature of the task tempts you to skim some articles, leading to decreased decision cau-

tion. Fatigue sets in, making the reading process more taxing and slowing down your progress. Maybe you even get a sudden insight into how to screen the articles more efficiently, leading to a heightened information processing speed.

As evident from the described scenario, the underlying cognitive constructs do not remain constant over time. Assuming stability of these constructs throughout all categorization decisions fails to reflect the reality of cognitive processes. Instead, it is crucial to recognize and account for the dynamic nature of these constructs. In our simplified example, imagining a scenario where we acknowledge and address these dynamics reveals valuable insights. For instance, we may discover that typically, just thirty minutes into article screening, there is a noticeable decline in information processing. Such insight could prove immensely beneficial in enhancing both the efficiency and quality of the article screening process.

The present dissertation focuses on the dynamics inherent in cognitive processes. The central argument of this thesis is that dynamics in cognitive process model parameters do matter. After exploring the dynamics in a specific cognitive process to underscore this importance, the dissertation is dedicated to overcoming significant limitations of stationary cognitive models.

I propose a novel, innovative approach called *neural superstatistics*, which not only addresses dynamics within cognitive parameters but also does it highly efficiently. By providing reproducible open-source code and by discussing important practical aspects, I provide other researchers with a tool to account for dynamics across a broad spectrum of cognitive processes. Through various applications to *in silico* and *in vivo* experiments, I demonstrate its feasibility and the inherently dynamic nature of cognitive constructs.

This dissertation marks a step in advancing cognitive process models to a new level. Its contributions are invaluable in deepening our comprehension of cognitive processes and in building more realistic models of cognitive processes.

1.1 CONTRIBUTIONS

The present dissertation is structured as follows:

- [Chapter 2](#) introduces cognitive process models, with a special focus on the diffusion decision model.
- [Chapter 3](#) discusses the occurrence of dynamics in cognition. It then provides an overview to common approaches to include dynamics in cognitive process models.
- [Chapter 4](#) delves deeper into a specific cognitive process, namely duration discrimination. Here, I will demonstrate that accounting for dynamics can be important.

- [Chapter 5](#) dives deeper into a novel approach to account for dynamics in cognitive process models, called neural *superstatistics*, which solves limitations of the previously reviewed methods.
- [Chapter 6](#) discusses the main contributions of this thesis in a larger context and provides practical recommendations and an outlook for future research endeavors.

1.2 LIST OF SCIENTIFIC ARTICLES OF THE CUMULATIVE DISSERTATION

This dissertation is based on three scientific articles. Two of them have been published in peer-reviewed journals and one is currently under review. Copies of the articles can be found in the Appendix.

- Schumacher, L., & Voss, A. (2023). Duration discrimination: A diffusion decision modeling approach. *Attention, Perception, & Psychophysics*, 85(2), 560–577. <https://doi.org/10.3758/s13414-022-02604-1>
- Schumacher, L., Bürkner, P.-C., Voss, A., Köthe, U., & Radev, S. T. (2023). Neural superstatistics for Bayesian estimation of dynamic cognitive models. *Scientific Reports*, 13(1), Article 13778. <https://doi.org/10.1038/s41598-023-40278-3>
- Schumacher, L., Schnuerch, M., Voss, A., & Radev, S. T. (2023). Validation and comparison of non-stationary cognitive models: A diffusion model application. *arxiv*. <https://doi.org/10.48550/arXiv.2401.08626> (submitted at *PLoS Computational Biology*)

2 COGNITIVE PROCESS MODELS

Essentially, all models are wrong, but some are useful.

— G. E. P. Box

Research in psychology and cognitive science revolves around deriving meaning from empirical data and providing an explanation for systematic patterns within these data (Hempel & Oppenheim, 1948). This endeavor typically starts with the development of a *verbal* theory about how psychological phenomena influence behavior within a specific environment, such as an experiment. From these theories, we derive hypotheses which can be empirically tested. To this end, components of the verbal theories are operationalized, measured, and then linked to each other by means of a statistical model. In other words, generic *statistical models*, such as generalized linear models (GLM), are used to describe the relationships between observed variables.

While these models provide a quantitative framework for understanding complex interactions, their interpretation suffers from the ambiguity of the underlying verbal theories. Many of these theories exhibit weak logical connections to the hypotheses they are empirically evaluated against (Oberauer & Lewandowsky, 2019). Furthermore, statistical models rest on distributional and causal assumptions that may not align with the substantive theories under investigation, creating a phenomenon known as the *theory-description gap* (Haines et al., 2023). When such assumptions diverge, theories become disconnected from the statistical tests intended to validate or refute them, hindering scientific progress (Szollosi & Donkin, 2019; Yarkoni, 2022).

Many cognitive scientists argue that for a deeper understanding of how the mind works, considerations of data and verbal theorizing alone are insufficient. They propose relying on *cognitive process models*¹, akin to physicists studying gravity (Farrell & Lewandowsky, 2018; Kriegeskorte & Douglas, 2018). Unlike traditional statistical models, these models mathematically formalize underlying latent cognitive components and specify how these components generate behavior. This involves translating *verbal* theories into *formal* mathematical models (van Rooij & Blokpoel, 2020).

By decomposing cognition into several functional components and specifying how behavior is generated, these models eliminate the aforementioned ambiguity of interpretation inherent in verbal theories. This approach goes beyond estimates of “effects” or relationships. The primary

¹also known as computational cognitive models, process-based models, or cognitive models

aim of cognitive models is to explain and understand how a particular cognitive process, such as attention, memory, or belief updating unfolds (Farrell & Lewandowsky, 2018).

A principal advantage of computational models is that we are forced to specify all parts of our theory. If we had a full understanding of a cognitive process, then we should be able to engineer it – or in the words of Richard Feynman “What I cannot create, I do not understand”. Computational models thus check whether our intuitions about the behavior of a theorized system match what actually arises from its realization.

Unpacking the latent cognitive process that shapes behavior has a long history in cognitive science. Over time, various classes of computational cognitive models have emerged, each offering unique insights into the intricacies of human cognition. For instance, cognitive architectures such as ACT-R (Anderson et al., 2004) or CLARION (Sun, 2016) provide formalized frameworks that encapsulate general principles of human information processing. Other classes of models concentrate on specific cognitive processes, such as decision-making (Batchelder & Riefer, 1999; Ratcliff et al., 2016), learning (Eckstein et al., 2021), or memory processes (Burgess & Hitch, 2005), to name just a few. Even other classes put emphasis on the relation between cognition and neurological processes (Palmeri et al., 2017)

In conclusion, cognitive process modeling is an invaluable tool for advancing our understanding across various domains, including cognitive science, psychology, and neuroscience. Its departure from traditional statistical modeling approaches can enrich our theoretical understanding of cognition. As cognitive models continue to evolve, researchers unlock new avenues for exploration. In the following, I will describe the idea of a cognitive model in general before I narrow the focus to a specific cognitive model relevant to my work.

2.1 SIMULATING THE MIND

Formally, a cognitive process model can be expressed as a function g that generates data x based on a set of parameters θ :

$$x_n = g(\theta, \xi_n) \quad \text{with} \quad \xi_n \sim p(\xi), \quad (2.1)$$

where the subscript n denotes the ability of the generative function to produce a sequence of data points $\{x_n\}_{n=1}^N$, for instance, multiple trials in a psychological experiment. Unlike machines that can repeat the same actions mechanically, humans exhibit considerable variability in their behavior. To accommodate such variability cognitive process models are typically formulated as *stochastic* generators. In Equation 2.1, this stochasticity is introduced through an independent source of noise ξ_n , sampled from an appropriate noise distribution $p(\xi)$. Consequently, even

with the same parameter set, such a generative model yields diverse data, making it inherently *non-deterministic* (Radev et al., 2020).

The described generative process is *memoryless*, implying that data is generated without reliance on the history of previously simulated data points. Consequently, these simulators produce *independent and identically distributed* (IID) data. This assumption is fundamental to many cognitive process models (Batchelder & Riefer, 1999; Ratcliff & Murdock, 1976; Swets & Green, 1978) but may not always hold. In the next chapter of this dissertation, I will argue that this assumption should be questioned, discussing various approaches to challenge it.

These models can function as powerful exploratory tools. We do not necessarily need human data to learn from our models. Instead, they play a crucial role as checks on our reasoning, helping us assess whether their behavior aligns with our expectations. It is noteworthy that models often exhibit unexpected or counterintuitive behaviors, underscoring their value in uncovering novel insights.

By expressing a cognitive process mathematically as a stochastic simulator, we empower ourselves to translate hypotheses about the mind into actionable experiments *in silico*. Through simulations, we can systematically observe the outcomes of our considerations. We can also make quantitative predictions about individual behavior based on latent cognitive constructs and underlying model assumptions (McClelland, 2009). This enables us to scrutinize the implications of our ideas, leading to the derivation of testable hypotheses.

Any model that explains data is itself unobservable. Instead, a model serves as an abstract tool that exists primarily in the minds of researchers and aims to describe, predict, and explain data. It is essential to recognize that models are always simplifications of reality. We are rarely able to account for all aspects that go into a specific process. These models often necessitate sacrificing certain details to maintain feasibility (Farrell & Lewandowsky, 2018; McClelland, 2009).

2.2 INFERRING THE MIND

The *forward* equation (cf. Equation 2.1) proves versatile in various applications. Nevertheless, our primary interest often lies in *inverse* inference – determining the most plausible hidden parameters responsible for generating a given set of observations. This challenge is commonly referred to as the *inverse inference* problem (Poldrack, 2006).

It is crucial to recognize the asymmetry between forward and inverse inference regarding their computational and epistemic complexities. In essence, forward inference proves relatively straightforward, requiring only the ability to articulate the model as a simulator program and execute it with a specific parameter set. On the contrary, the task of estimating plausible parameters from a given data set involves a substantial degree of uncertainty.

As mentioned earlier, cognitive process models are typically non-deterministic and lack information preservation. This leads to a scenario where simulating data with a generative model results in the loss of information regarding the data-generating parameters, which is not directly embedded in the data. Moreover, the information contained in observed data may prove insufficient to fully and unambiguously reconstruct the data-generating process. Frequently, multiple model and parameter configurations present themselves as plausible explanations for the observed data. Consequently, intrinsic uncertainty surrounds the accurate determination of the true value of θ when relying on a finite set of observations $\{x_n\}_{n=1}^N$.

To tackle this challenge, a range of methods is available. Common approaches for deriving plausible parameters from observed data include maximum likelihood estimation in a frequentist context (Myung, 2003) or Markov chain Monte Carlo (MCMC; van Ravenzwaaij et al., 2018), and simulation-based inference in a Bayesian setting (Cranmer et al., 2020). In this discussion, the focus is on the Bayesian inference perspective, as this is the approach that was used in all three manuscripts included in the present dissertation.

In Bayesian modeling, emphasis is placed on the likelihood of the data, denoted as $p(x | \theta)$. The likelihood quantifies the probability of observing data x given a specific set of parameters θ . To estimate parameters based on observed data (i.e., inverse inference), it is necessary to invert the likelihood. This inversion requires a prior distribution $p(\theta)$, representing initial beliefs about plausible parameter values. Bayes' rule, expressed as

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}, \quad (2.2)$$

is then utilized to update these beliefs in light of the observed data. The result is a distribution of plausible parameter values given the observed data, known as the *posterior distribution* $p(\theta | x)$. In Equation 2.2, $p(x)$ denotes the *marginal likelihood*, representing the probability of the data under all possible parameter settings. It is worth noting that calculating the marginal likelihood for parameter estimation is unnecessary, as it simply ensures the area under the posterior distribution equals one.

Before estimating a cognitive model it is essential to assess its feasibility and validity. In a Bayesian framework, adhering to a principled Bayesian workflow, as outlined by Schad et al. (2021), becomes crucial. This involves three key steps: (i) conducting *prior push forward checks* to scrutinize the model's assumptions, (ii) assessing *computational faithfulness*, for instance through simulation-based calibration (SBC; Säilynoja et al., 2021; Talts et al., 2018), and (iii) evaluating *inferential calibration*.

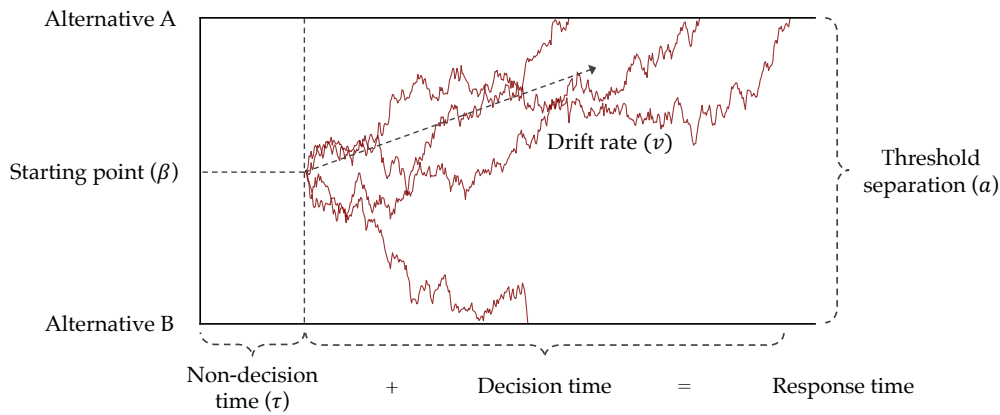


Figure 2.1: A graphical illustration of the diffusion decision model's evidence accumulation process with its four core parameters: The drift rate, which denotes the average rate of evidence accumulation; the threshold separation that sets the distance between the two choice alternatives A and B; the relative starting point of the accumulation process; and the non-decision time that accounts for the time of all decision unrelated processes.

The assessment of inferential calibration (also called *parameter recovery*) is particularly vital when seeking insights into latent cognitive constructs from estimates of cognitive process model parameters. Typically, this involves a parameter recovery study where synthetic data is simulated, and the generative model is fitted to this data to determine whether the true data-generating parameters can be accurately recovered. The successful recovery of data-generating parameters suggests model *identifiability*.

Following successful validation on simulated data, the next step involves fitting the model to human data, inferring plausible parameters, and evaluating the model's credibility through an examination of its absolute fit to the specific dataset. This assessment is conducted by re-simulating the data using the inferred parameters and comparing the simulated outcomes with the actual data – a process known as *posterior re-simulation* or *posterior retrodiction checks*. It is only when the model can accurately capture the essential patterns in the empirical data that we should proceed with interpreting and reasoning about the estimated parameters.

After outlining the general principles of cognitive modeling, I delve into a specific cognitive process model that forms the foundation of all three manuscripts in this dissertation – the diffusion decision model (DDM).

2.3 DIFFUSION DECISION MODEL

Initially developed by Ratcliff (1978), the Diffusion Decision Model (DDM) is arguably one of the most influential cognitive process models. It describes an individual's decision-making process between two alternatives. Its core assumption is that people accumulate evidence over time until a certain threshold is reached. This evidence accumulation process is mathematically formalized by the following stochastic ordinary differential equation:

$$dx_n = v dt_s + z \sqrt{dt_s} \quad \text{with} \quad z \sim \mathcal{N}(0, 1). \quad (2.3)$$

Accordingly, the evidence x_n on a trial n evolves as a random walk with some drift rate v and standard normally distributed noise z , while t_s represents time on a continuous time scale. The average rate of evidence accumulation is described with the drift rate v . Note, that the scale of the diffusion noise (z) here is set to 1.² Typically, this scaling parameter is fixed for identifiability reasons. The evidence accumulation process terminates as soon as one of two thresholds, 0 or a , is reached, and the decision corresponding to the reached boundary D_n is made:

$$D_n = \begin{cases} 1, & \text{if } x_n \geq a \\ 0, & \text{if } x_n \leq 0 \end{cases}. \quad (2.4)$$

Each boundary corresponds to one of the two choice alternatives, and their separation represents the amount of evidence required before the respective choice is made. A small(er) boundary separation means that less information is accumulated until a decision is made. Moreover, the DDM encompasses a constant τ that accounts for the duration of all decision-unrelated processes. Together with the time of evidence accumulation until one threshold is reached (i.e., decision time), this results in the response time of one specific decision:

$$x = \inf\{t_s \geq 0 \mid x(t_s) \geq a \text{ or } x(t_s) \leq 0\} + \tau \quad (2.5)$$

Additionally, the relative starting point of the evidence accumulation process β can be estimated.

The DDM, thus, has four core parameters $\theta = \{v, a, \tau, \beta\}$ (see Figure 2.1 for a graphical illustration). Each of these parameters is assumed to map on a specific cognitive construct involved in the decision-making process: the drift rate v is a proxy for mental processing speed; the threshold a is a metric for decision caution; the non-decision time τ accounts for the time spent on stimu-

²An alternative convention is setting this scaling parameter to 0.1.

lus encoding and motor action; and the relative starting point β measures a potential *a priori* bias toward a choice alternative.

Assuming such a process, one can account not only for the binary choice data but also for response times. Jointly modeling these two observable variables has proven to be advantageous because of one important phenomenon: Individuals engaged in *speeded* decision-making tasks can to some extent trade-off speed with accuracy. This so-called *speed-accuracy trade-off* (Heitz, 2014; Luce, 1986) suggests that individuals face a dilemma: they can either expedite decision-making, sacrificing accuracy or prioritize precision, necessitating more time for evidence accumulation. The DDM has no problem accounting for this phenomenon by varying the threshold parameter.

The DDM has inspired research in many fields (for a comprehensive review see, Ratcliff et al., 2016; Voss et al., 2013). Although it was initially developed to elucidate behavior in perceptual and memory-based decision-making, its application extended to a wide range of decision-making processes, including social decisions (Klauer et al., 2007) and value-based decisions (Tajima et al., 2016). Notably, a multitude of studies have unveiled connections between evidence accumulation and neural signals, such as neural firing rates, electroencephalography (EEG), and functional magnetic resonance imaging (fMRI) signals (Bode et al., 2012; Boehm et al., 2014; Churchland & Ditterich, 2012; Gold & Shadlen, 2007; Nunez et al., 2017). For example, O’Connell et al. (2012) showed in an EEG study that the drift rate parameter is highly correlated with ERP components. These investigations provided compelling evidence that the DDM serves as a robust algorithmic approximation of the decision-making processes actively executed by our brains.

Beyond its neural plausibility, a key factor contributing to the widespread adoption of the DDM is that the mapping of its parameters to the assumed latent cognitive constructs has been rigorously validated in numerous experimental studies (Arnold et al., 2015; Lerche & Voss, 2019; Voss et al., 2004). For instance, when participants prioritize accuracy over speed, resulting in more careful task execution, this is reflected in a larger separation of thresholds, indicating heightened response caution (Mormann et al., 2010). Similarly, when trials are manipulated to favor one response over another, the starting point shifts accordingly (Mulder et al., 2012). Variations in stimulus discriminability, making the task more challenging or easier, are reflected in changes to the drift rate (Ratcliff & McKoon, 2008). Moreover, mandating the use of a single finger for all responses (Lerche & Voss, 2019) or requiring multiple keypresses for each response (Voss et al., 2004) increased the value of the non-decision time parameter. Collectively, these studies strongly indicate that the DDM parameters indeed capture the cognitive constructs involved in the proposed evidence accumulation process.

More generally, the DDM belongs to a broader class of models called *evidence accumulation models* (EAM; Evans & Wagenmakers, 2020).³ EAMs can be divided into two categories: (i)

³Sometimes also called *sequential sampling models*.

accumulator-based models, where evidence for each choice alternative is accumulated in a separate accumulator. The process is terminated when one of the accumulators first reaches a pre-determined threshold. This approach allows modelers to investigate tasks with more than two choice alternatives. Notable examples include the *linear ballistic accumulator* (Brown & Heathcote, 2008), the *racing diffusion model* (Tillman et al., 2020), and the *leaky competing accumulator* (Usher & McClelland, 2001). (ii) diffusion or random walk models that track the relative evidence between two choices, with the DDM being its most prominent model. Here, evidence for a specific alternative is always also evidence against the other alternative. Despite this distinction, both types of models generally incorporate common cognitive constructs, such as decision caution, non-decision time, and *a priori* bias.

In summary, both the DDM and other models within the EAM class have significantly advanced our comprehension of various facets of decision-making and cognition more broadly. Moving forward, I will delve into the intricacies of the dynamics within cognitive constructs that these models seek to formalize.

3 ACCOUNTS FOR DYNAMICS IN COGNITION

Everything flows, nothing stands still.

— Heraclitus, 501 BC

Cognitive and behavioral components, such as those formalized by the DDM, exhibit changes over time, regardless of the time scale we look at. For instance, there is a large body of literature that investigated how cognitive aspects such as mental processing speed or decision caution change over the life span (Theisen et al., 2021; von Krause et al., 2022; von Krause et al., 2021). Such studies either investigate a population of individuals cross-sectionally or longitudinally and observe how aspects of cognition change with age. For example, von Krause et al. (2022) showed in a cross-sectional study with a sample of over one million individuals that the main cause for slowing down in tasks with increasing age is an increase in decision caution and slower decision-unrelated processes. These results challenged a widespread belief that a decreasing mental processing speed is the cause for this slowing.

Dynamic changes also appear on a much shorter time scale, for example during a session of a psychological experiment or even on a trial-by-trial basis. Imagine you are rapidly solving a speeded decision-making task. All of a sudden you realize that you made a mistake. Will you continue to perform your task rapidly on subsequent trials or will you make some adjustments? It has been found that people tend to slow down in their task-solving speed after they make an error – an effect called *post-error slowing* (Laming, 1979; Rabbitt & Rodgers, 1977). A study based on a total of over one million trials from 39 participants performing a decision task revealed that this slowdown can be attributed to an increase in the response caution (Dutilh et al., 2012). In addition, other studies found a decrease in mental processing speed and a change in non-decision time following a mistake (Damaso et al., 2022; Dutilh et al., 2013; Purcell & Kiani, 2016; Schiffler et al., 2017).

Effects such as post-error slowing occur rather abruptly and as a consequence of what happened in a previous trial. However, during an experimental session, cognitive constructs also change more slowly and gradually. Typical examples of such *time-on-task* effects are increases in performance due to practice or decrease thereof as a result of declining motivation, attention, or fatigue (Gunawan et al., 2022).

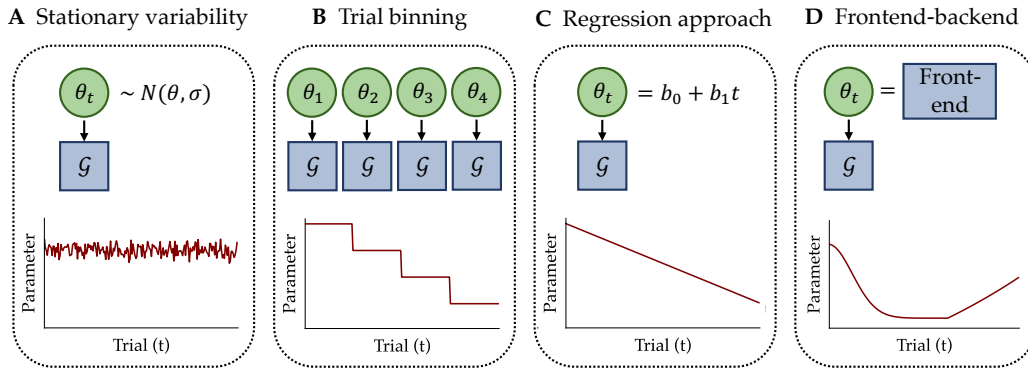


Figure 3.1: A conceptual depiction of the four common strategies for modeling temporal variations in the parameters (θ) of a cognitive process model (\mathcal{G}). **A** *Stationary variability*, also referred to as inter-trial variability, posits that parameter values fluctuate around a stable mean. **B** *Trial binning* involves categorizing data into distinct bins and fitting a cognitive process model (\mathcal{G}) to each bin individually. **C** *Regression approach* employs time (and sometimes additional contextual variables) as predictors for the parameters (θ). **D** *Frontend-backend* models employ a mechanistic model (the frontend) to elucidate the dynamics of the parameters of the cognitive process model (the backend).

In summary, we constantly adapt our cognition and behavior to external and also internal demands and circumstances. As a consequence, cognitive process models should account for such changes to accurately describe behavior.

Frequently, when applying cognitive process models, researchers tend to overlook dynamic aspects, resulting in models reliant on static parameters. These models assume stability in cognitive components over time, a presumption that, more often than not, proves to be inaccurate. Neglecting the temporal aspect of cognitive parameters can lead to numerous adverse consequences, including inflated uncertainty estimates and potentially misguided conclusions drawn from static parameters.

In essence, I argue that a comprehensive understanding of cognitive processes necessitates a more nuanced examination of their temporal dynamics. By adopting a perspective that accounts for the evolution of cognitive components, researchers can refine their models, ultimately contributing to a more robust and accurate depiction of the intricate interplay within cognitive functioning.

In the subsequent discourse, I will explore common approaches for integrating time-varying parameters into cognitive process models. I broadly classify these approaches into four categories: *stationary variability*, *trial binning*, *regression approach*, and *frontend-backend* models. Refer to [Figure 3.1](#) for an abstract illustration of these four approaches. In the following section, I will elaborate on each of these approaches.

3.1 STATIONARY VARIABILITY

The Stationary variability approach involves considering random fluctuations of one or more cognitive model parameters around a stable mean, known as stationary or inter-trial variability (see [Figure 3.1A](#)):

$$\theta_t \sim p(\theta | \eta), \quad (3.1)$$

where the parameters θ_t on a given trial are sampled from some probability distribution with a constant mean θ and some additional distributional parameters η . The rationale behind this is that certain cognitive constructs do not remain static over time; instead, they demonstrate variability while still maintaining a stable overall state. A well-known instance of this approach is the “full” diffusion decision model, which permits inter-trial variability in its fundamental parameters (i.e., drift rate, non-decision time, and starting point) (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). Stationary variability in the threshold parameter has to be omitted in order to prevent tractability issues.

Accounting for these inter-trial variabilities in the DDM has proven advantageous for two reasons: Firstly, it allows the production of different response time distributions for correct and error responses. Secondly, it makes it easier to obtain a good estimate of the non-decision time. This parameter is by definition bounded to the fastest response time in the data. When a data set contains fast responses and we assume a static non-decision time parameter then the parameter is forced to have a relatively low value although it may be higher in some trials. Employing stationary variability in this parameter solves this problem.

The inclusion of stationary inter-trial variability primarily enhances the model fit within the observed data, yet it falls short in detecting systematic changes or abrupt shifts in the model parameters since it only allows for variability around a stable mean. Systematic changes would expect the mean to increase or decrease and abrupt shifts might lead to values outside the variability of the stable mean. Furthermore, such an approach still treats behavioral data as IID and does not take information from previous trials into account. These limitations render this method inadequate for exploring systematic fluctuations in cognitive constructs.

3.2 TRIAL BINNING

An alternative method that is capable of identifying systematic changes in components of a cognitive model involves fitting a stationary model to trial bins (see [Figure 3.1B](#)):

$$\theta_t = \begin{cases} \theta_1, & \text{if } t \leq T_1 \\ \theta_2, & \text{if } T_1 < t \leq T_2 \\ \vdots & \\ \theta_m, & \text{if } T_{m-1} < t \leq T_m \end{cases}. \quad (3.2)$$

First, the data are divided into m bins, each containing a number of trials, denoted as T . Subsequently, a stationary model is fitted to each bin independently, resulting in bin-specific parameter sets $\theta_1, \dots, \theta_m$. This method assumes that cognitive constructs remain constant within a trial bin and change at a predefined time point, returning to stability thereafter. For example, Evans and Brown (2017) employed this technique to demonstrate that individuals approach statistically defined optimality concerning their speed-accuracy trade-off. This was reflected in a threshold parameter that decreased from bin to bin.

Nevertheless, choosing the number of trial bins m is often more of a pragmatic decision than one driven by theory. This leads us to a fundamental issue with the approach. If one opts for many bins with a small number of trials, the result is a relatively fine-grained parameter dynamic but also highly uncertain parameter estimates due to insufficient data within a specific bin. Thus, there is a trade-off between dynamics resolution and estimate certainty. Another drawback of trial binning is that estimates within a specific bin are not informed by data from neighboring bins. However, the attractiveness of time-varying models lies in their unique ability to utilize both past and future data to constrain estimated parameter trajectories. Also, within each bin, the model treats data as IID and is unable to detect potential changes in the model components.

3.3 REGRESSION APPROACH

The third category involves a generalized linear model (GLM) with time and possibly other contextual variables as predictors for cognitive process model parameters (see [Figure 3.1C](#)):

$$\theta_t \sim p(\mathcal{L}(\sum_{j=0}^y \beta_j t^j), \eta), \quad (3.3)$$

where the parameter θ_t at some time point t is estimated from a linear combination of polynomial terms $\mathcal{L}(\sum_{j=0}^y \beta_j t^j)$ with at least time t as a predictor. $p(\cdot, \eta)$ indicates the form of the probability distribution, which depends on some parameters denoted by η . Please note that the specific details of the distribution and the choice of the error term (if any) are not provided in [Equation 3.3](#). This depends on the assumptions made about the errors in the GLM. A common

choice, for instance, is the normal distribution for linear regression with normally distributed errors.

Cochrane et al. (2023) utilized this approach to investigate learning-related changes in the DDM’s drift rate parameter during a perceptual decision-making task. A comparison between models with different polynomial terms provided evidence for an exponential increase in drift rates due to perceptual learning.

The GLM approach presents a more attractive alternative to trial binning, offering the capability to identify both linear and non-linear changes in model parameters without sacrificing resolution. However, it is crucial to acknowledge that the underlying regression function comes with strong assumptions about the nature of the relationship between model parameters and time. Despite the common practice of fitting and comparing several plausible specifications, such as linear versus exponential models, anticipating and determining all possible specifications in advance can be challenging. Consequently, the overall flexibility of the GLM model for process characterization remains notably constrained (Gunawan et al., 2022).

3.4 FRONTEND-BACKEND MODEL

In contrast, the *frontend-backend* approach seeks to capture alterations in model parameters while providing an explanation for the dynamic nature of the parameters. In this framework, the backend model corresponds to the cognitive model that formalizes the generation process of behavioral data, such as a DDM. The frontend model comprises a mechanistic model that elucidates how the parameters of the backend model evolve over time, varying across contexts and in response to additional factors (see [Figure 3.1D](#)):

$$\theta_t = f(\theta, t, c, \eta), \quad (3.4)$$

where the cognitive model parameter θ_t at a specific time step t is determined by a function f incorporating time t , context c (e.g., experimental conditions), previous parameter values $\theta_{1:t-1}$, and additional parameters η as potential predictive variables. [Equation 3.4](#) is intentionally formulated in a general manner because different frontend-backend models vary immensely. For instance, one might instantiate a working memory capacity model that describes changes in the DDM’s drift rate parameter, while another describes how mind-wandering leads to inattention and how it changes people’s decision caution over time. Also, some use time t as an explicit factor, while others implicitly consider time through functions relying on previous parameter values $\theta_{1:t-1}$.

An illustrative example in recent research involves leveraging reinforcement learning models as frontend models to elucidate changes in DDM parameters resulting from reward-based learning

(Fontanesi et al., 2019; McDougle & Collins, 2021; Miletic et al., 2021). In this case, the frontend model describes how the drift rate parameter of the DDM changes based on an updating function. This function employs static parameters, such as the learning rate, which maps to the individuals' tendency to take feedback into account.

Such an approach goes beyond a regression approach, which provides a mere description of parameter changes using trend functions. The frontend-backend approach provides a mechanistic explanation of their temporal variations through a set of static parameters and deterministic functions. Typically, these static frontend parameters are linked to cognitive constructs in a manner similar to how cognitive process models map to cognitive constructs. While the frontend-backend approach holds promise, it is not always straightforward to work with. Mechanistic explanations for cognitive component changes may be unknown or challenging to link to the backend model.

In summary, all four approaches represent a crucial step in addressing the inadequacy of treating cognitive parameters as static. However, each has its drawbacks and challenges. Before providing a novel approach that overcomes some of these challenges, in the next chapter, I will present a specific example of a cognitive process in which accounting for dynamic changes might be necessary. In this work, I employed, among other methods, a frontend-backend model to account for the dynamics in this process.

4 DYNAMICS IN DURATION DISCRIMINATION (MANUSCRIPT I)

*It has to start somewhere, it has to start sometime,
what better place than here? What better time than now?*
— Rage Against the Machine, “Guerrilla Radio”

One aspect of human time perception involves the ability to discriminate the duration of two sequentially presented stimuli. When investigating this process, a common methodology entails presenting two stimuli: a constant *standard* stimulus and a variable *comparison* stimulus. Participants are then tasked to determine which of the two stimuli was presented for a longer duration. A fundamental psychophysical model, known as the *difference model* (Thurstone, 1927a, 1927b), conceptualizes the decision-making process as follows: individuals, when faced with the choice of discerning the duration difference between two stimuli, S_1 and S_2 , internally compute the difference as $D = S_1 - S_2$.

In an experimental context, task difficulty is typically manipulated by varying the magnitude of this difference between stimuli. The more similar the two stimuli, the more difficult the task becomes. Participant’s performance in such a task can be described by a (*sigmoidal*) psychometric function, wherein choice accuracy is a function of varying stimulus intensity, in this case the varying stimulus duration (see Figure 4.1 for examples). When analyzing the performance in this manner, two key features of the function are often of special interest. First, the horizontal shift of the sigmoid function, representing the point of *subjective equality* of the standard and comparison stimulus (Fechner, 1860). Second, the slope of the function, often referred to as the *just noticeable difference* describes the proficiency with which an individual can differentiate between the two stimuli (Ulrich & Vorberg, 2009).

Remarkably, studies indicate that these features, and consequently, individuals’ performance in this task, are not only influenced by the duration difference between the two stimuli but also by task-unrelated factors – referred to here as *context effects*. The *Type A effect*, commonly known as the *time-order error* (TOE), describes how the sequence in which stimuli are presented influences the point of subjective equality (Jamieson & Petrusic, 1975). In essence, participants tend to either overestimate or underestimate one stimulus compared to the other, contingent on the

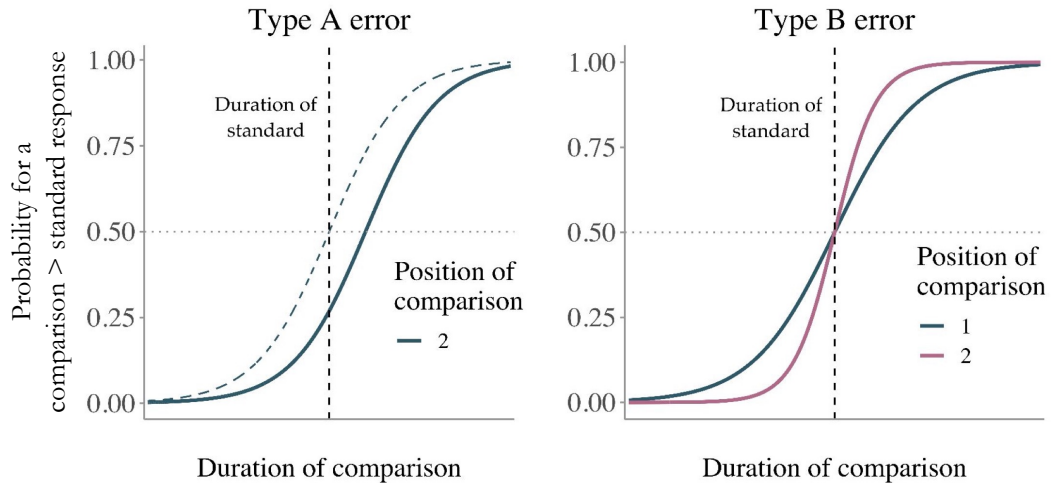


Figure 4.1: Hypothetical psychometric functions map response probabilities to varying durations of the comparison stimulus. The vertical dashed line marks the constant duration of the standard stimulus (adopted from Schumacher and Voss (2023)). The point where the psychometric functions intersect the horizontal dashed line is referred to as the *point of subjective equality*. **Left panel** The solid blue line is shifted to the right, indicating a Type A error, as the duration of the comparison is larger than the standards' duration at the point of subjective equality. **Right panel** When the comparison stimulus precedes the standard stimulus (position of comparison = 1) the slope of the psychometric function decreases, a phenomenon known as a Type B error.

order in which they are presented (Hellström, 1985). For example, individuals tend to overestimate the magnitude of the standard stimulus when it precedes the comparison stimulus but do not show such a bias when the standard stimulus is presented after the comparison (see left panel in Figure 4.1).

Moreover, it has been found that the stimulus presentation order also has an impact on the just noticeable difference. Dyjas et al. (2012) demonstrated that participants exhibit enhanced discrimination performance when discerning the duration between a constant standard stimulus and a varying comparison stimulus if the standard stimulus precedes the comparison, rather than follows it. This phenomenon is commonly termed as a *negative Type B effect* (Ulrich & Vorberg, 2009) and has been observed not only in duration discrimination but also in other domains such as weight (Ross & Gregory, 1964) and contrast discrimination (Nachmias, 2006). The Type B effect is characterized by a reduced slope of the sigmoid function for trials where the comparison stimulus precedes the standard compared to trials with the reversed order (see right panel in Figure 4.1). In contrast, the Type A effect involves a mere lateral shift of the sigmoid function, mapping response probabilities to the difference in stimulus duration.

Both Type A and Type B effects are considered *global context* effects, stemming from the history of previously encountered stimuli. The prevailing explanation for these context effects is that decisions regarding the magnitude of a stimulus feature (e.g., duration) are not solely based on the current stimulus but are also influenced by previously encountered stimuli. Consequently, information stored in the memory system shapes the perception and decisions concerning subsequently presented stimuli. This is where dynamics in cognition come into play.

Lapid et al. (2008) proposed that participants maintain an internal reference of a prototype stimulus in their memory, continually updating it over time. Dyjas et al. (2012) introduced the *internal reference model* (IRM), a mechanistic model explaining how this internal reference (I) is established and updated across trials. According to the IRM, the internal reference on a given trial (I_n) is computed as a weighted sum of the internal reference from the previous trial (I_{n-1}) and the internal representation ($S_{1,n}$) of the first stimulus in the current trial. This implies that the internal reference is dynamically updated on a trial-by-trial basis, following a geometrically moving average. The updating of the internal reference is thus expressed as follows:

$$I_n = gI_{n-1} + (1 - g)S_{1,n}, \quad (4.1)$$

where the parameter g ($0 \leq g \leq 1$) signifies the weight assigned to the internal reference. In making decisions, participants compare this internal reference (I_n) with the internal representation of the second stimulus ($S_{2,n}$), resulting in a difference ($D_n = I_n - S_{2,n}$). If this difference (D_n) is greater (smaller) than 0, participants decide that the first stimulus was longer (shorter). Multiple studies demonstrated that the IRM is capable of accounting for the Type A and Type B effects (Bausenhart et al., 2014, 2015; Dyjas et al., 2012, 2014; Ellinghaus et al., 2018).

When a variable comparison stimulus is presented before a constant standard stimulus, the stability of the internal reference across trials is compromised as the variable stimulus becomes integrated. This fluctuation in the representation of the internal reference contributes to a decline in discrimination performance. Consequently, the magnitude of the Type A and Type B effect is expected to rise with higher values of g since the perception is more profoundly influenced by the dynamically changing internal reference (Dyjas et al., 2014).

Hellström (1979) proposed an alternative account, called *sensation weighting model* (SWM). This model does not assume an updating mechanism of an internal reference but assumes that the two presented stimuli are simply weighted differently:

$$D_n = w_1S_{1,n} - w_2S_{2,n}, \quad (4.2)$$

where D_n is the subjective difference between the two stimuli $S_{1,n}$ and $S_{2,n}$ each weighted by w_1 and w_2 ($0 \leq w \leq 1$). In this model, context effects are attributed to varying weights assigned to stimuli. Hellström (1985) demonstrated that assigning a larger weight to the second stimulus leads to a Type B effect. Hellström et al. (2020) argued that the SWM but not the IRM adequately explains the full spectrum of observed Type B and Type A effects. The IRM encounters difficulties when the standard stimulus is no longer fixed, especially in *roving standard tasks*, and struggles to account for positive Type B effects observed at times (de Jong et al., 2021; Hellström et al., 2020). Dyjas et al. (2014) proposed a promising hybrid approach, suggesting that combining the generality of the SWM with the trial-by-trial updating mechanism of the IRM could be fruitful.

Both models discussed rely on the concept of stimulus comparison, employing a linear model with distinct weights for the two stimuli or integration of past stimulus experiences. These models represent significant advancements over the standard difference model in explaining diverse context effects. However, they offer limited insights into the decision process itself.

Manuscript I¹, thus, focused on enhancing the understanding of duration discrimination processes by integrating the principles of the IRM and SWM into a DDM, allowing for a comprehensive analysis of the decision-making process. This integration enabled a comprehensive analysis of the decision-making process, offering insights into crucial cognitive aspects such as potential *a priori* biases. Moreover, considering both choice and response time data can mitigate inferential biases and serve as a valuable constraint for parameter estimation, ultimately enhancing parameter recoverability (Ballard & McClure, 2019; Shahar et al., 2019).

The study pursued several objectives. Firstly, as previously discussed, we aimed to integrate the principles of the IRM and SWM into a DDM to incorporate response times and gain deeper insights into duration discrimination. Secondly, the study sought to conduct a rigorous model comparison among competing models combining different combinations of the IRM, SWM, and DDM. Thirdly, it tested a specific hypothesis regarding whether the DDM's starting point parameter is influenced by the first stimulus presented.

The underlying rationale of these models lies in the notion that the difference between the durations of two stimuli, computed based on the IRM or SWM, impacts the drift rate of the DDM. For instance, in the case of combining the IRM and the DDM model, the drift rate v_n on a given trial n is calculated as follows:

$$v_n = v_0 + v_1(I_n - S_{2,n}), \quad (4.3)$$

¹Schumacher, L., & Voss, A. (2023). Duration discrimination: A diffusion decision modeling approach. *Attention, Perception, & Psychophysics*, 85(2), 560–577. <https://doi.org/10.3758/s13414-022-02604-1>

where v_0 represents a baseline drift rate, and v_1 scales the difference between the internal reference I_n (computed according to Equation 4.1) and the second stimulus $S_{2,n}$. When this difference is large, the task of discriminating between the two stimuli durations becomes relatively easy, resulting in a higher drift rate.

This model constitutes a frontend-backend model, with the IRM as the frontend and the DDM as the backend model. It produces non-IID data since the internal reference (I_n) on a given trial n is computed using the updating mechanism (cf. Equation 4.1), which depends on previously encountered stimuli. The factor time is, thus, implicitly included. However, our primary interest lies not in the explicit dynamics of the drift rate parameter itself but rather in accounting for them to explain crucial patterns in the data, such as Type A and B errors.

Conversely, the model integrating the principles of the SWM into a DDM computes the drift rate according to:

$$v_n = v_0 + w_1 S_{1,n} + w_2 S_{2,n}, \quad (4.4)$$

where w_1 and w_2 denote weights for the first ($S_{1,n}$) and the second stimulus ($S_{2,n}$). In this model, the drift rate (v_n) on a given trial n solely depends on the presented stimuli. It does not incorporate information from past stimulus encounters or explicitly utilize the factor of time. Consequently, it produces IID data and does not fit into any of the dynamic modeling approaches described in Chapter 3.

Additionally, we tested a combination of the IRM and the SWM, expressed as:

$$v_n = v_0 + w_1 I_n + w_2 S_{2,n}, \quad (4.5)$$

where both the internal reference I_n and the second stimulus $S_{2,n}$ are weighted. This model incorporates the concepts of the IRM by including the internal reference updating mechanism and the rationale of the SWM by differently weighting both stimulus representations.

Besides a rigorous comparison between these competing models, we investigated whether the magnitude of the first stimulus $S_{1,n}$ or, in the case of the IRM, the internal reference I_n , affects the relative starting point of the DDM. This assumption stems from the idea that individuals initially encode the duration of the first stimulus, which then influences the evidence accumulation process while attending to the presentation of the second stimulus.

We re-analyzed the data from two experiments, which were previously published by Dyjas et al. (2012). In these experiments, participants had to discriminate the duration of two auditory stimuli (Experiment 1) or two visual stimuli (Experiment 2). In both instances, one stimulus was

a variable comparison and the other a constant standard. The stimulus order was manipulated, which led to a clear Type A and Type B effect for all participants. On the basis of these data, we compared different DDMs that either used the traditional difference model, the IRM, the SWM, or a combination of the two latter, for the drift rate and also whether the starting point parameter depends on the first stimulus or the internal reference.

Results concerning the relative goodness-of-fit computed with approximate leave-one-out cross-validation (Vehari et al., 2017) indicated that a model that incorporated the SWM model for the drift rate and let the starting point vary as a function of the first stimulus provided overall the best out-of-sample prediction across the two experiments. However, other models with the IRM mechanism as well as the model including a combination of the IRM and SWM performed only slightly worse. Also, it appeared that the influence of the first stimulus on the starting point of the evidence accumulation process was very small. We further evaluated the model with the best relative goodness-of-fit in terms of absolute fit to the data of the two experiments by means of posterior re-simulation. The results showed that the model is not only capable of closely fitting the choice data with its Type A and B effect patterns but also explains the full response time distribution.

In summary, this work introduced an innovative way to model how people make decisions about the discrimination of time intervals. By combining existing models of how we perceive and compare stimuli (IRM, SWM) into a DDM, we found that our approach accurately predicts how well people can judge the difference between two subsequently presented stimuli. Importantly, our model expanded existing models by considering not only choice but also response time data. Additionally, we demonstrated that the model was able to predict two well-known effects in this area of study: the Type B effect and the Type A effect. However, our analyses did not provide strong evidence in favor of the IRM or the SWM because strong model mimicry was observed.

5 SUPERSTATISTICS

Everything should be made as simple as possible, but not simpler.

— Albert Einstein

In [Chapter 3](#), I discussed prevalent strategies for addressing dynamics in cognitive model parameters, followed by a specific application of one of these methods. In the current chapter, I delve into *superstatistics* – a method to directly estimate non-stationary dynamics in cognitive process models from observed data.

Beck and Cohen (2003) introduced the term “superstatistics” to denote a combination of multiple stochastic processes operating on distinct temporal scales, providing a framework to elucidate heterogeneous temporal dynamics. Instead of assuming static model parameters, superstatistics introduces a hierarchical structure comprising at least two models. First, a low-level (i.e., observation or microscopic) model that formalizes the local behavior of a (cognitive) system. Second, a high-level (i.e., transition or macroscopic) model that characterizes the parameter dynamics of the low-level model (see [Figure 5.1](#) for a conceptual illustration).¹ This framework enables the estimation of non-stationary low-level parameter dynamics directly from observed data. In other words, such models remain largely agnostic about the dynamics of the model components, impose limited constraints, and operate in a data-driven manner.

The superstatistics approach overcomes various limitations associated with the previously presented methodologies. In contrast to stationary models, superstatistical models can generate non-stationary variations in the parameters of the low-level model, enabling gradual changes as well as abrupt transitions between different states. Another key feature is that parameter estimates are influenced by past data points, diverging from the assumption of IID data. Unlike the trial-binning approach, models within the superstatistics framework harness the entirety of available data, alleviating concerns about having insufficient data points for accurate parameter estimation. Differing from the regression approach, superstatistics places minimal constraints on potential parameter trajectories, resulting in a significantly less restrictive modeling framework.

At first glance, the superstatistics framework may seem similar to the frontend-backend model approach discussed earlier. However, in the case of superstatistics, the transition model is a stochas-

¹There is no intrinsic time scale assigned to low- and high-level processes; their interpretation is contingent on the scale relevant to the specific research question, rendering these terms relative in meaning.

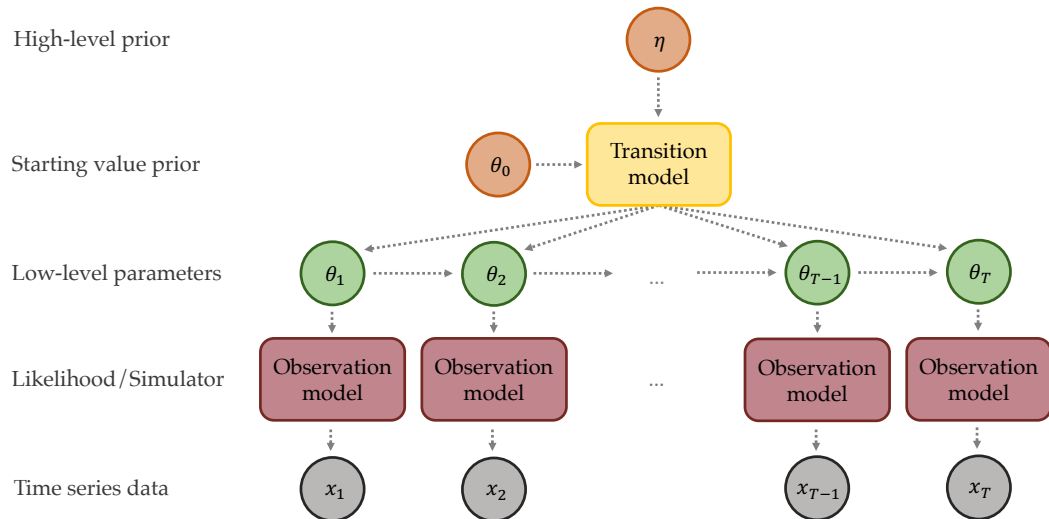


Figure 5.1: A graphical illustration of a superstatistical model. A low-level observation model (likelihood) produces time series $\{x_t\}_{t=1}^T$ with time-varying parameters $\{\theta_t\}_{t=1}^T$. These parameters follow a high-level transition model using static parameters η and sensible starting values θ_0 . Prior distributions are employed both on the high-level parameters η and the starting values of the low-level parameters θ_0 .

tic process rather than a set of deterministic functions like the frontend model. Due to the utilization of a stochastic process for the dynamics of the parameters, every parameter at any given time step in the time series is treated as a free parameter. Also, the transition function of superstatistical models does not encompass mechanistic explanations for the parameter dynamics, opting instead for enhanced flexibility in their estimation. While mechanistic explanations hold pivotal roles in cognitive science, instances arise where such explanations are either absent or are only applicable to specific parameters. Consequently, I view these two approaches as complementary. The superstatistical framework adopts a bottom-up, exploratory stance, serving as a tool for hypothesis generation. In subsequent stages, one could potentially devise plausible frontend models, drawing insights from parameter trajectories inferred using a superstatistical model. Furthermore, superstatistical models can function as benchmarks for scrutinizing and validating competing frontend-backend models, achieved through a comparative analysis of resulting parameter dynamics derived from both methodologies.

Superstatistics as a modeling framework to estimate non-stationary parameter trajectories directly from data has found applications in many fields of research. Among them are the examination of train delays (Briggs & Beck, 2007), cancer survivals (Leon Chen & Beck, 2008), wind velocity fluctuations (Rizzo et al., 2004; Santhanam & Kantz, 2008; Stoevesandt & Peinke, 2010), earth surface temperature (Yalcin & Beck, 2013), sea-level fluctuations (Rabassa & Beck, 2015), air pollution (Williams et al., 2020), and economics (Denys et al., 2016; Van der Straeten & Beck,

2009), to name just a few. These applications underscore the versatility of the superstatistics concept in addressing complex phenomena across diverse domains.

Surprisingly, despite its appeal, it has found very little application in cognitive science (but see Metzner et al., 2021). Computational complexities might be one of the reasons for this absence. In fact, estimating and comparing superstatistical models presents significant challenges, particularly within a Bayesian framework. Several factors contribute to these challenges. First, both the high-level and low-level models are stochastic, introducing substantial uncertainty regarding the values of all model parameters when confronted with a finite number of observations. Second, the low-level models may be intricate and nonlinear, making it challenging to establish a closed-form analytic expression connecting model parameters to data (i.e., rendering the likelihood function intractable), or the evaluation of the likelihood might be computationally demanding. Lastly, even for stationary low-level models, the computational burden can become daunting when applied to multiple datasets. This is because standard Bayesian methods are not amortized, necessitating sequential re-running from scratch for each dataset.

In this dissertation, I argue that cognitive science can benefit from adopting a superstatistics modeling framework. As discussed in Chapter 3, cognition exhibits numerous dynamic aspects. To capture these nuances in mathematical models and thus obtain a more realistic representation of cognitive processes, it is imperative to incorporate these dynamics. However, existing methods for achieving this have notable limitations. Manuscript II² therefore explores the superstatistics framework in the context of cognitive process models, more specifically the DDM. In this work, we developed a novel Bayesian estimation method for such models using custom neural networks. Additionally, we provide authors with open-source code, which can be adjusted for individual needs (<https://github.com/bayesflow-org/Neural-Superstatistics>).

5.1 NEURAL SUPERSTATISTICS (MANUSCRIPT II)

Following Mark et al. (2018), we characterize dynamic models by a low-level observation model with time-varying parameters $\{\theta_t\}_{t=1}^T$ of length T , which dynamically evolve according to a high-level transition model with static parameters η . The low-level model is specified by a likelihood function \mathcal{L} , while the high-level model encompasses a transition function \mathcal{T} . In our study, the focus was on addressing general superstatistical models, where the likelihood function \mathcal{L} of the low-level model may not be analytically tractable. These models are implemented as randomized stateful simulators, generating observable trajectories $\{x_t\}_{t=1}^T$ through the following general recurrent system:

²Schumacher, L., Bürkner, P.-C., Voss, A., Köthe, U., & Radev, S. T. (2023). Neural superstatistics for Bayesian estimation of dynamic cognitive models. *Scientific Reports*, 13(1), Article 13778. <https://doi.org/10.1038/s41598-023-40278-3>

$$\theta_t = \mathcal{T}(\theta_{0:t-1}, \eta, \xi_t) \quad \text{with} \quad \xi_t \sim p(\xi | \eta) \quad (5.1)$$

$$x_t = \mathcal{G}(x_{1:t-1}, \theta_t, z_t) \quad \text{with} \quad z_t \sim p(z | \theta_t). \quad (5.2)$$

In these equations, \mathcal{T} represents a high-level transition function with static parameters η . \mathcal{G} denotes a transformation that encapsulates the functional assumptions of the low-level model. Moreover, the random noises, $\xi_t \sim p(\xi)$ and $z_t \sim p(z)$, introduce variability. The initial parameter values and the static high-level parameters follow prior distributions, $\theta_0 \sim p(\theta)$ and $\eta \sim p(\eta)$, respectively. The former encapsulates available information regarding plausible low-level parameter values and the latter represents initial beliefs about the behavior of the parameter trajectories. This framework allows for a comprehensive exploration of dynamic models, especially in situations where closed-form expressions for the low-level model likelihood may not be readily available.

Moreover, it is a highly flexible framework. In terms of the transition function \mathcal{T} , we have many options to choose from. For instance, we can assume that the low-level parameters follow a Gaussian random walk:

$$\mathcal{T}(\theta_{t-1}, \eta, \xi_t) = \theta_{t-1} + \eta \xi_t \quad \text{with} \quad \xi_t \sim \mathcal{N}(0, 1). \quad (5.3)$$

Another possibility could be a Gaussian process (GP) transition model where the parameters depend not only on the previous state of the process but rather on the entire history of the process:

$$\theta_{1:T} \sim \mathcal{GP}(\mu_\theta, K_\theta), \quad (5.4)$$

where μ_θ and K_θ correspond to a mean and covariance function defined by a vector of time step indices. In this case, the transition model parameter η denotes the parameters of a Gaussian kernel.

5.1.1 AMORTIZED BAYESIAN INFERENCE

When fitting superstatistical models to data, the goal in a Bayesian analysis setting is to obtain full posterior distributions for both the entire low-level parameter trajectory $\{\theta_t\}_{t=1}^T$ and the static high-level parameters η . To this end, we devised an *amortized Bayesian inference* method (Radev et al., 2020) based on recurrent probabilistic neural networks, specifically tailored for superstatistical models, hence termed as neural superstatistics. This method comprises two phases: (i) a relatively computationally intensive training phase of custom neural networks, and (ii) an almost

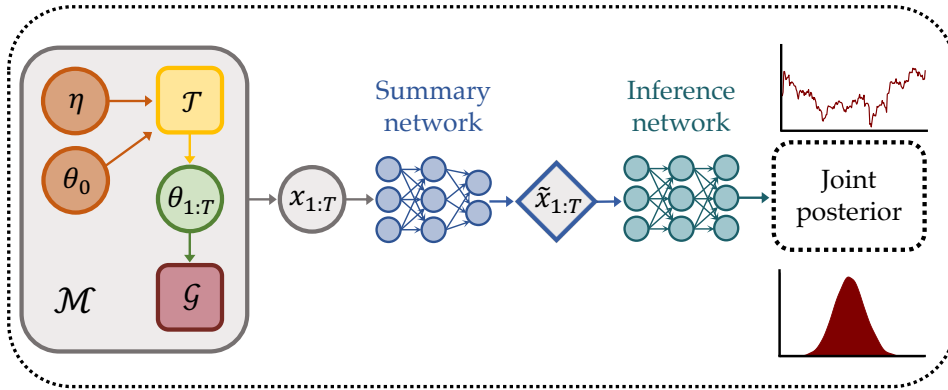


Figure 5.2: A conceptual illustration of the amortized Bayesian inference workflow for parameter estimation (adopted from Schumacher, Schnuerch, et al. (2023)). A recurrent *summary* network consumes time series $x_{1:T}$ generated with a superstatistical model \mathcal{M} and learns maximally informative summary statistics $\tilde{x}_{1:T}$. An *inference* network (i.e., normalizing flow) is trained to approximate the joint posterior distribution of time-variant low-level parameters $\theta_{1:T}$ as well as static high-level parameters η given the learned summary statistics.

instantaneous inference phase, where the full joint posterior $p(\theta_{1:T}, \eta | x_{1:T})$ is recovered from observed time series $\{x_t\}_{t=1}^T$.

The training of the neural networks relies on simulated data sets generated by a generative model (refer to Figure 5.2 for an illustration of this workflow). The simulated data is input to a *summary network*, which learns *maximally informative summary statistics*. The output of the summary network, along with the data-generating parameters (e.g., $\theta_{1:T}$ and η), are then passed into an *inference network*. This invertible neural network learns the relationship between data and parameters and, after sufficient training, becomes a *Bayesian inference expert*, capable of inferring the joint posterior distribution solely from the observed data.

Traditional Bayesian estimation methods, like *Markov chain Monte Carlo* (MCMC), necessitate repetitive computational costs when applied to new data sets. Additionally, they require the availability of a closed-form likelihood function for the low-level model. In contrast, with *amortized inference*, once the neural network is trained, it can be effortlessly applied to numerous datasets, resulting in the *amortization* of the computational burden. One notable advantage of amortization is its facilitation of model validation checks, which often require fitting the model to a large number of data sets. Furthermore, this method eliminates the necessity for closed-form likelihoods, thereby expanding its applicability to complex cognitive process models lacking an analytical solution (see for example, Usher & McClelland, 2001; Voss et al., 2019; Wieschen et al., 2020)

5.1.2 BENCHMARK STUDIES

To assess the efficacy of our novel estimation method, we conducted benchmark studies comparing it with two existing algorithms, namely `bayesloop` (Mark et al., 2018) and `stan` (Carpenter et al., 2017). While `bayesloop` relies on grid approximation for low-dimensional problems, `stan` employs Hamiltonian Monte Carlo (HMC) sampling (Neal, 2011), which is commonly considered the gold standard for Bayesian inference. Both methods operate in a non-amortized manner and are restricted to estimating superstatistical models exclusively with closed-form likelihoods.

Given that `bayesloop` is currently limited to fitting simple low-level models, we utilized a basic Poisson process to model data containing counts of coal mining accidents in the United Kingdom from 1852 to 1962. This low-level model involves a single parameter λ , representing the accident rate per year, with a Gaussian random walk (cf. Equation 5.3) serving as the high-level transitions model. Both `bayesloop` and our novel neural estimation method produced nearly identical latent trajectories for the low-level model parameter λ , demonstrating the capability of our method to estimate plausible parameter trajectories for simple low-level models.

In our second benchmark with `stan`, we fitted a non-stationary diffusion decision model (NSDDM) to data simulated with a static DDM that used constant parameters. The NSDDM, featuring a Gaussian random walk transition model, allowed the drift rate, threshold, and non-decision time to vary freely over time. Both our neural estimation and the `stan` method performed equally well in recovering the true parameters of the static DDM. These results showed that the NSDDM is capable of approaching a stable parameter value and does not result in pseudo-dynamics.

In summary, our neural estimation method demonstrated performance comparable to established methods for Bayesian inference, with a notable advantage: due to amortization, the neural estimation significantly outperformed `stan` in terms of computation time.

5.1.3 SIMULATION STUDY

Subsequently, we evaluated the recoverability of NSDDM parameters under induced misspecifications, using models different from those utilized during network training. This thorough investigation involved simulating data sets, each comprising $T = 400$ time points, across four scenarios: (i) a static DDM with constant parameters; (ii) a DDM with stationary variability, where the three DDM parameters fluctuate randomly around a constant value; (iii) a NSDDM with a Gaussian random walk transition model; and (iv) a DDM with constant parameters exhibiting abrupt and uniform jumps at three predefined time points, constituting a regime-switching model.

Importantly, the neural approximator was exclusively trained using simulations from (iii). However, during amortized inference, we applied the network to 200 data sets from each of the four

scenarios. This approach enabled us to explore the network’s response in an open-world setting where the true data generator might deviate from the reference model used in the training phase.

The results indicated excellent parameter recoverability in the first three cases, with estimated parameter trajectories quickly approaching the true data-generating parameters. In the third scenario, the neural estimation closely followed the true non-stationary trajectory. In the fourth, severely misspecified case, parameter recovery remained fairly good, although the inferred parameter trajectory struggled to shift abruptly to the new true parameter value when a regime switch occurred. This aligns with expectations, considering the transition model of the NSDDM (i.e., Gaussian random walk) did not allow for such jumps. Overall, the ability to recover true data-generating parameters under various misspecifications underscores the feasibility of NSDDMs and highlights the strength of our novel neural estimation approach, setting the stage for fitting NSDDMs to actual human data.

Furthermore, we also tested the computational faithfulness of our model and estimation method with means of simulation-based calibration (SBC; Säilynoja et al., 2021; Talts et al., 2018). The analyses indicate that our neural Bayesian method demonstrates satisfactory calibration, albeit with slightly miscalibrated posteriors for the non-decision time parameter.

5.1.4 HUMAN DATA APPLICATION

Following the successful evaluation of the NSDDM and our novel neural estimation method *in silico*, we applied variants of the NSDDM to data sets collected from two separate speeded two-alternative choice tasks. The first application served as an initial exploration using data from a well-known task in experimental psychology, which was previously examined by Evans and Brown (2017). In their study, they explored dynamic changes in the threshold parameter over a maximum of 1 320 trials in different between-subject conditions, utilizing the previously discussed trial binning approach.

It allowed us to compare the inferred parameter trajectories from our neural superstatistics method with those obtained in the original study using their trial binning approach. Upon comparing our estimates with those obtained by Evans and Brown (2017), it was apparent that both approaches yield comparable qualitative and quantitative patterns. This consistency not only aligned with our promising results in simulated scenarios but also emphasized the convergent validity of our superstatistics approach in real-world data applications.

In the second application, we analyzed relatively long time series ($T = 3\,200$ trials per individual) stemming from a lexical decision-making task. We aimed to demonstrate the effectiveness of our method in estimating a more complex NSDDM featuring a Gaussian process (cf. Equation 5.4) transition model and multiple drift rate parameters for various task difficulty conditions.

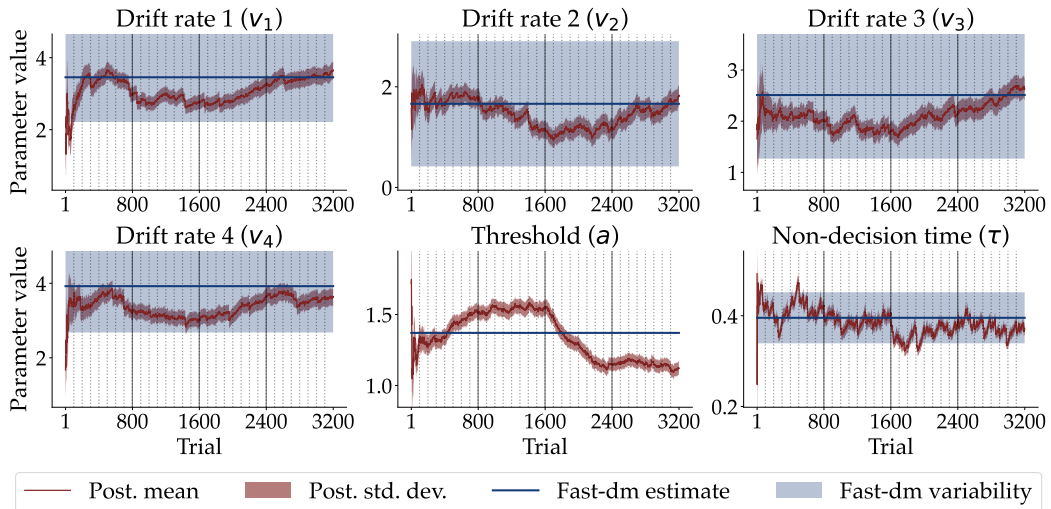


Figure 5.3: Low-level DDM parameter trajectories inferred from one individual’s data (adopted from Schumacher, Bürkner, et al. (2023)). Across-trial posterior means and standard deviations for all six parameters, namely four drift rates $v_1 - v_4$ (one per difficulty condition), the threshold a , and the non-decision time τ are shown in red. The point estimates of the stationary DDM parameters estimated with `fast-dm` along with the inter-trial variabilities (not available for the threshold parameter) are depicted in blue.

Additionally, a DDM with stationary variability was fit with the `fast-dm` software (Voss & Voss, 2007) to the same data for comparison.

Data re-simulation with the inferred parameters showed an excellent fit to the marginal empiric response time (RT) distribution for both the NSDDM and the stationary DDM. Analysis of the entire RT time series revealed that for many individuals RTs decreased over time. Also, the overall variability of the RTs was less pronounced in the later stages of the experiment. Only the NSDDM was capable of accurately reproducing these two patterns, as the stationary DDM only produces IID data. Being able to account for the time series posits a noteworthy advantage of the NSDDM over its stationary counterpart.

However, the most significant advantage of the neural superstatistic approach is that the full parameter (non-stationary) trajectory can now be inspected (see Figure 5.3). In fact, all parameter trajectories displayed significant fluctuations and pronounced oscillations throughout the experiment. Notably, due to the assumption of homogeneous variation, the inter-trial variabilities inferred with `fast-dm` tend to overestimate the uncertainty in parameter estimates. The dynamic drift rates exhibited fluctuations generally within the uncertainty corridors delineated by the homogeneous inter-trial variabilities but displayed considerably narrower error bars. Consequently, the local parameters appeared much less uncertain than the homogeneous variability parameters suggested. Conversely, the dynamic non-decision time demonstrated more extensive fluctuations

compared to the corresponding flat inter-trial variability. Also, while $f_{\text{fast-dm}}$ estimated a single threshold value, which remained constant over time, the dynamic threshold parameter indicated a substantial decrease in decision caution throughout the experiment for most individuals. In summary, there was a remarkable incongruity between heterogeneous and homogeneous dynamics observed in nearly all individuals.

5.1.5 INTERMEZZO

Manuscript II explored superstatistics as a novel modeling framework to capture non-stationary dynamics within cognitive processes and explored the applicability of a neural Bayesian method for estimating superstatistical models. Through extensive simulations and two benchmark studies, we established the computational fidelity and adequacy of our method. Subsequently, we applied our approach to a NSDDM, estimating temporal trajectories for key parameters, namely the drift rate, threshold, and non-decision time, using data from two experiments. Our findings demonstrated that such a non-stationary model (i) can be accurately fitted to lengthy empirical time series and (ii) unveils nuanced patterns obscured by traditional stationary models.

Notably, we demonstrated how to enhance stationary cognitive models through the integration of a superstatistics framework. However, a critical question lingered: Do the parameter trajectories inferred with a superstatistical cognitive process model genuinely capture shifts in the latent constructs they intend to represent, or are they merely modeling artifacts? To explore this question, Manuscript III³ was initiated, focusing on an experimental validation study.

5.2 VALIDATION AND COMPARISON OF DYNAMIC COGNITIVE MODELS (MANUSCRIPT III)

In this work, our focus once again centered on the DDM as a low-level observation model. One of the reasons for this choice is that the stationary DDM has been rigorously validated before (Arnold et al., 2015; Lerche & Voss, 2019; Voss et al., 2004). These studies demonstrated that the components underlying the DDM (as detailed in Chapter 2) indeed correspond to the intended cognitive constructs. Consequently, it becomes straightforward to experimentally manipulate specific model parameters. For instance, the drift rate, reflecting mental processing speed, is influenced by task difficulty, with high difficulty generally resulting in lower drift rate values, and *vice versa*. Similarly, the threshold parameter, indicative of decision caution, can be manipulated, for instance, by providing verbal instructions that emphasize speed over accuracy.

³Schumacher, L., Schnuerch, M., Voss, A., & Radev, S. T. (2023). Validation and comparison of non-stationary cognitive models: A diffusion model application. *arxiv*. <https://doi.org/10.48550/arXiv.2401.08626>

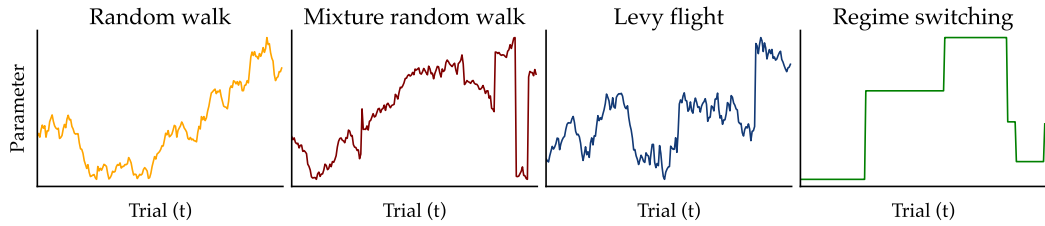


Figure 5.4: An exemplification of the four high-level (transition) models, central to our investigation, which govern the temporal dynamics of a hypothetical low-level model parameter (adopted from Schumacher, Schnuerch, et al. (2023)).

In our validation study of the NSDDM, we implemented a color discrimination task (previously employed in the validation study by Voss et al. (2004)). In this task, we incorporated the two aforementioned manipulations in a specific temporal pattern across 768 trials. The task difficulty changed on every 8th or 16th trial, transitioning between four difficulty levels. Additionally, verbal instructions emphasizing either speed or accuracy changed less frequently after every 48th trial, resulting in distinct temporal patterns for the two manipulations. Task difficulty changed frequently and gradually, while task instructions underwent abrupt switches (termed *regime switching*).

The primary objective of our experiment was to examine whether the parameter trajectories inferred with the NSDDM align with the changing patterns of the experimental conditions. Specifically, we anticipate the drift rate parameter to mirror the gradual changes in task difficulty. Simultaneously, we expect the threshold parameter to exhibit sudden shifts when the priority switches between speed and accuracy. It's crucial to note that in this application, the NSDDM lacked information about the experimental conditions, relying solely on the behavioral data for parameter trajectory inference.

We also explored what kind of transition model is most suitable for the expected dynamics. As I discussed in Chapter 3, there are various transition functions to choose from. Given the described manipulation sequences, we tested transition functions capable of capturing gradual changes, abrupt shifts, or both. To this end, a total of four NSDDMs distinguished only by their transition function were compared (see Figure 5.4 for exemplar trajectories): (i) a Gaussian random walk; (ii) a mixture between a Gaussian random walk and uniformly distributed jumps; (iii) a Lévy flight; and (iv) a regime switching function, where parameters either remain the same as in the previous time step or shift uniformly. These models differ in their complexity (i.e., number of high-level parameters) and their ability to accommodate various types of temporal dynamics. The Gaussian random walk primarily captures relatively small gradual changes, while (ii) and (iii) have the capacity to model both gradual changes and sudden shifts. In contrast, (iv) exclusively allows for sudden shifts without accounting for small gradual changes.

To conduct Bayesian model comparison between these competing models, we introduced a novel amortized method based on neural methods from Elsemüller et al. (2023) and Radev et al. (2021). The procedure involved training an ensemble of ten neural approximators based on simulations from all four NSDDMs. Similar to the method we used for parameter estimation, these approximators comprise a summary network and an inference network. The former once again performs data compression by learning maximally informative summary statistics for the simulated time series $x_{1:T}$. The inference network approximates the posterior model probability (PMP) for the candidate models, $p(\mathcal{M} | x_{1:T})$, given the outputs of the summary network.

Before we applied our Bayesian model comparison method to the empirical data, we rigorously validated the method *in silico*. To this end, we performed simulation-based calibration as well as a model recovery study based on 10 000 simulated data sets per model. The results indicated that our method is well-calibrated, suggesting that the inferred PMPs are trustworthy. The model recovery analysis showed that some of the models were frequently confused with another model. For example, the Lévy flight DDM was confused with the random walk DDM 30% of the time. Also, the mixture random walk DDM got confused with the regime switching DDM about 40% of the time. This result showed that some of the models produce fairly similar data and are thus not always easy to distinguish.

Our model comparison procedure indicated the Lévy flight DDM as the most plausible model, with an average PMP of approximately 60% across all individuals. It stood out as the most plausible model for 9 out of the 14 participants. In contrast, the mixture random walk model garnered an average PMP of less than 30% and was identified as the most plausible model for 5 participants. The random walk DDM and regime-switching DDM consistently demonstrated lower plausibility compared to the other models and did not emerge as superior for any of the participants. From this, we concluded that transition models allowing for small gradual changes as well as abrupt shifts are more appropriate for this data set compared to transition functions that only allow for either of the two dynamic types. This makes sense as the experimental manipulation sequences resemble both of these dynamic types.

To check whether the experimental manipulations had the expected effects on the behavior of the participants, response times, as well as choice accuracy, were aggregated over the two experimental conditions (difficulty; instruction) and individuals. In both, the *accuracy* and the *speed* condition RTs increased as a function of task difficulty. The effect of difficulty on response times was more pronounced in the accuracy condition compared to the speed condition and response times were generally faster in the speed condition. Furthermore, the choice accuracy decreased with increasing task difficulty. Individuals responded slightly less accurately in the speed condition.

All four NSDDMs were fitted to the empiric data to evaluate their absolute fit to these critical data patterns. Posterior re-simulation results indicated that all models successfully captured essential patterns in the data. Only the choice accuracy in the most difficult condition was consistently overestimated. This misfit is likely due to the model's inability to appropriately decrease the value of the drift rate parameter when a switch to this difficulty level occurred. However, this minor misfit could likely be rectified by incorporating experimental context information into the model. Despite this, the results underscored the NSDDMs' remarkable capability to fit empirical data solely based on behavioral information, without any information concerning the experimental context.

At the core of the validation study were the inferred parameters, prompting the pivotal question: Do these parameter dynamics mirror the sequence of the experimental manipulations? To address this, both time-averaged and time-varying estimates were examined. The former provides a global overview of average parameter differences between experimental conditions, while the latter offers a more detailed analysis of parameter evolution throughout the experiment.

Time-averaged analyses revealed that both drift rates and thresholds exhibited the expected patterns. On average, the drift rate decreased with increasing task difficulty, regardless of the instructional condition. For the threshold parameter, assumed to be influenced by the manipulation of verbal instructions, increased values were observed in the accuracy condition compared to the speed condition.

In the analysis of time-varying estimates, parameter trajectories from the model with the highest PMP for each participant were inspected (see [Figure 5.5](#) for an example). The dynamic of the drift rate aligned with the global trend of the task difficulty sequence for all participants. Examining the trajectory of the threshold parameter, the hypothesized pattern of a decrease during a shift from an accuracy to a speed instruction, and *vice versa*, was evident in the estimated trajectories.

In summary, both at an aggregate level and on a time series level, the drift rate and threshold parameters demonstrated alignment with the expected patterns from experimental manipulations. Posterior re-simulations with NSDDMs showcased excellent absolute model fit to behavioral data, even without explicit information about the experimental context. Consequently, the validation study supports the notion that NSDDMs can effectively detect genuine changes in cognitive constructs.

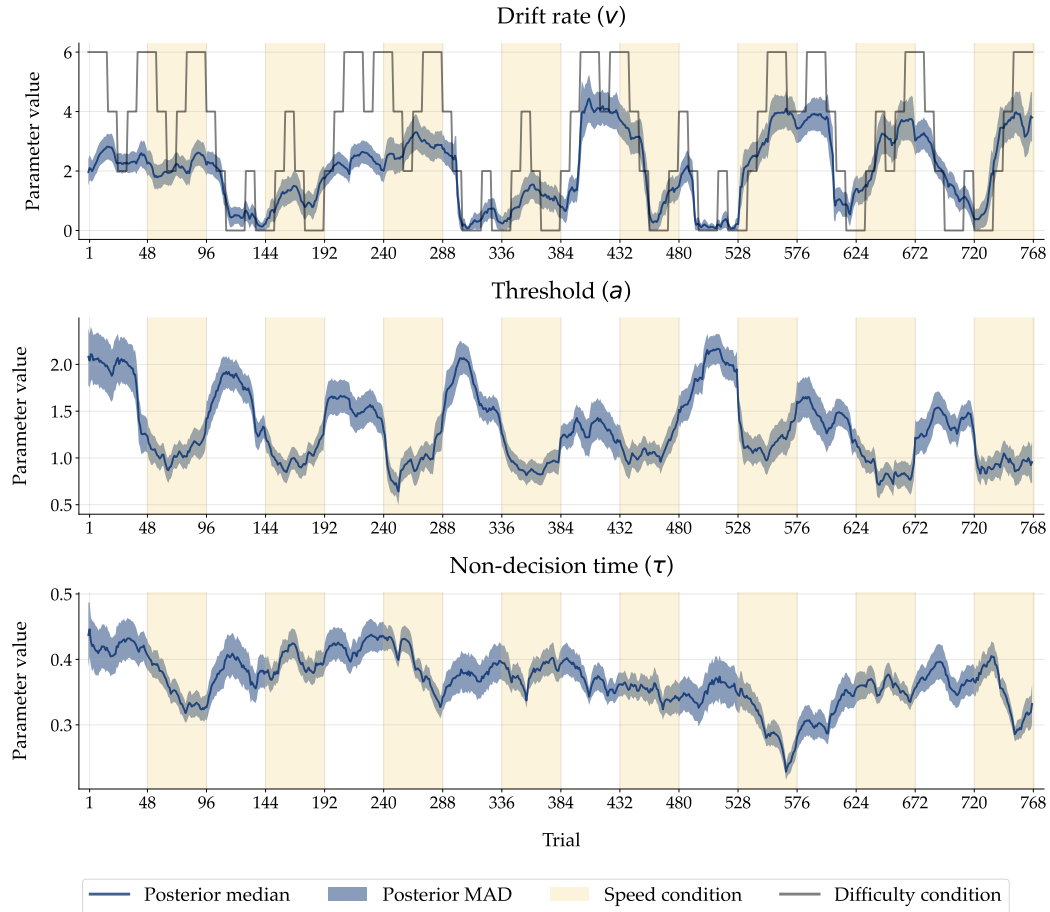


Figure 5.5: Parameter trajectories of an exemplar individual corresponding to the respective best-fitting non-stationary diffusion decision model (adopted from Schumacher, Schnuerch, et al. (2023)). In this instance, they resulted from a Lévy flight DDM. The low-level DDM parameters (i.e., drift rate, threshold, and non-decision time) are displayed on separate rows. The solid lines depict the posterior medians, whereas the shaded regions correspond to the median absolute deviation. Trials, where speed was emphasized, are shaded in yellow color, while white regions indicate trials with an emphasis on accuracy. In the top panel, the sequence of the task difficulty levels is depicted in black.

6 GENERAL DISCUSSION

Randomness is indistinguishable from complicated, undetected, and undetectable order; but order itself is indistinguishable from artful randomness.

— Nassim Nicholas Taleb

This dissertation commenced with a brief introduction to cognitive process models and why such models are indispensable tools for scientific reasoning in the domains of cognitive science. I discussed the tendency of these models to overlook the dynamic nature inherent in human cognition. Subsequently, various strategies were explored to address this limitation, including stationary variability, trial binning, regression approach, and frontend-backend modeling.

The subsequent sections of the dissertation were dedicated to three manuscripts, divided into two overarching themes. In the first part, I delved into the dynamics of a specific cognitive process, namely duration discrimination. The exploration involved showcasing the utility of mechanistic models, specifically the internal reference model and the sensation weighting model, as a frontend to inform parameters of a backend model – in this case, the DDM.

The second theme showcased two manuscripts presenting a novel approach to incorporate dynamics into the latent components of cognitive process models, termed neural superstatistics. Manuscript II introduced the superstatistics framework, designed for estimating non-stationary variability of cognitive process model parameters directly from behavioral data. Addressing computational challenges, the study developed neural methods for Bayesian inference of such models, rigorously validated through benchmarks and simulation studies. Furthermore, the applicability of this approach was demonstrated using two data sets containing human behavioral data from well-known decision-making tasks.

Building upon this foundation, Manuscript III extended the investigation, empirically validating the dynamics of latent cognitive constructs estimated with superstatistical models. Additionally, a comparative analysis of competing model implementations was conducted. Collectively, these manuscripts significantly contribute to advancing the field of cognitive science by providing tools to incorporate dynamic aspects into cognitive process models.

The commonality across the three manuscripts is their shared focus on the development of novel models to account for dynamic changes in cognitive processes throughout the duration of an experimental task. Manuscript I involved a frontend-backend approach, incorporating a spe-

cific theory to delineate changes in a cognitive construct. In contrast, Manuscript II and Manuscript III employed a superstatistics approach, which does not provide mechanistic explanations for changes and instead imposes minimal assumptions to allow the actual data to dictate the most plausible trajectories of individual components. For the scientific practitioner seeking a comprehensive understanding of dynamic cognitive processes, a pivotal question emerges: is one of the approaches superior to the other?

APPROACHES TO MODEL DYNAMICS Before delving deeper into weighing frontend-backend models against the superstatistics approach, I would like to discuss the status of the other methods outlined in [Chapter 3](#), namely stationary variability, trial binning, and regression approach. I argue that estimating non-stationary models within a superstatistics framework should always be preferred over these three approaches.

Stationary variability is commonly employed to enhance the in-sample fit to the data by increasing model flexibility. However, in this approach, parameters only fluctuate around a stable mean, and the history of prior task trials is ignored. Superstatistics addresses both of these limitations. In the simulation study of Manuscript II, data were simulated using a DDM incorporating stationary variability. Subsequently, a NSDDM was fitted to these data, revealing robust parameter recovery performance. This demonstrated that the NSDDM accurately approximates the stable mean of the data-generating stationary DDM. Thus, when the “true” data-generating process aligns with a model exhibiting stationary variability, a non-stationary model still can replicate this pattern. However, if not, the non-stationary model can discern the deviation. Consequently, there is little reason to opt for a stationary variable model when the non-stationary counterpart can also produce such patterns while offering additional flexibility.

The other two approaches, trial binning, and the regression approach, primarily address questions regarding the evolution of a cognitive process model parameter over time. Trial binning grapples with an undesirable trade-off between estimating certainty and dynamic resolution, while the regression approach often relies on strong assumptions regarding the shape of plausible parameter trajectories. In contrast, a superstatistical model leverages information from the entire data set and imposes minimal assumptions on parameter dynamics, thereby overcoming the limitations of both aforementioned approaches. Therefore, given the challenges faced by these three approaches and the solutions offered by superstatistics, it seems clear that superstatistics should be considered the default choice among them.

Next, I focus on the comparison between frontend-backend models and the superstatistics approach. There, a general superiority of one of the two models cannot be assumed. Instead, the choice between the two approaches depends on the research focus. If the explicit interest lies in understanding changes in established cognitive model components or capturing specific patterns

in the data by accounting for such changes, then the development of a plausible frontend-backend model might be the ultimate goal. Superstatistical models can still be a valuable tool in this kind of research. On the other hand, if the primary goal is not particularly centered around dynamic changes in the components and one simply wants to test whether one or more parameters differ between experimental conditions or groups, a frontend-backend model might serve no purpose. However, estimating non-stationary versions of the model through a superstatistics framework remains a useful and advantageous endeavor.

FRONTEND-BACKEND MODEL VS. SUPERSTATISTICS Let us first discuss the former case. If one already has an explanation or a hypothesis regarding the dynamics of cognitive model parameters that can be translated into a cognitive mechanistic model (i.e., a frontend), it is advisable to do so. Often that is exactly what the goal of a study is – to test possible explanations for specific changes. Note, that such frontend-backend models go beyond the previously discussed regression approach. In the former, the parameters of the frontend model can once again be mapped to latent cognitive constructs, making it an explanatory tool. In contrast, the regression approach treats additional parameters merely as weights, such as the slope of a line, making it a descriptive tool. For example, in the internal reference model, which served as a frontend for the DDM in Manuscript I, the rationale was that a prototype stimulus is internally generated over time based on an updating mechanism that integrates prior encounters with the stimulus. In this model, the factor of time was implicitly embedded – the drift rate parameter depended on the history of previously encountered stimuli. Additionally, the free parameter of the frontend model mapped to a cognitive construct, measuring how much a person relies on the prototype instead of the currently encountered stimulus.

In the case where we already have a frontend-backend model in mind, a superstatistical model can serve as a comparison and benchmarking tool. Different from the frontend-backend approach it allows the parameters to vary relatively freely. The parameter dynamics resulting from fitting a frontend-backend model to some data can then be compared to trajectories estimated with a superstatistical model. Such a comparison can potentially reveal shortcomings in the frontend-backend model and suggest potential avenues for refinements. Furthermore, in cases where a plausible frontend-backend model is not known yet, superstatistical models can be used for an initial exploration and act as a hypothesis-generating tool.

Up to this point, I have discussed the synergistic application of these two modeling frameworks and how they can be used as separate methods to address the same problem by capitalizing on their respective strengths. However, an intriguing avenue emerges when we consider the integration of these approaches into a unified model. Two possibilities come to mind: firstly, incorporating dynamics into components of the frontend model; and secondly, accommodating non-stationary

fluctuations in components of the backend model that are not yet accounted for by the frontend model. To illustrate these possibilities, I will delve into a specific frontend-backend model, namely the reinforcement learning diffusion decision model (Fontanesi et al., 2019; Miletic et al., 2020; Millner et al., 2018; Pedersen & Frank, 2020; Sewell et al., 2019).

Typically applied to data from value-based decision-making tasks, the reinforcement learning DDM explains behavioral changes resulting from error-driven learning. The reinforcement learning (RL) model posits that individuals maintain representations of the choice alternatives' values, known as *expected values*, which guide their decisions. Following a choice, the expected values are updated based on the difference between their predictions and the actual outcome – the *reward prediction error* (Sutton & Barto, 2018).

The DDM provides a comprehensive account of the decision process, elucidating how choices and response times arise from latent process components (processing speed, response caution, bias, non-decision time). Integrating these two models, with RL as the frontend and DDM as the backend, not only allows for the simultaneous explanation of choices and response times but also accommodates changes therein. In the reinforcement learning DDM, the drift rate parameter is typically assumed to be the difference between the expected values of the two choice alternatives. Additionally, the expected values of the chosen alternative are updated based on the reward prediction error, leading to dynamic changes in the drift rate parameter due to this learning mechanism.

Now, where does the superstatistical framework fit into this? As mentioned earlier, one possibility is to consider dynamics in the parameters of the frontend model. For the reinforcement learning DDM, this could involve parameters such as the *learning rate*, governing the weight assigned to the reward prediction error when updating expected values, or the *temperature* parameter, dictating sensitivity to differences between expected values. By introducing transition functions to these parameters, these cognitive components would no longer remain static over time, enabling researchers to estimate the dynamics of these frontend model parameters.

Indeed, there are compelling reasons to believe that these parameters undergo changes during the course of an experiment. For instance, individuals often begin with a relatively high learning rate, which tends to decrease as the expected values of alternatives become well-known (Jepma et al., 2020). Similarly, the temperature parameter, employed to balance the exploration of unknown alternatives and exploitation of known good ones, is also likely to vary over time (Feng et al., 2021).

Usually, frontend models do not provide mechanistic explanations for all the backend parameters. Thus, a second possibility to unify frontend-backend models with a superstatistics approach is to allow for non-stationary fluctuations of the backend parameter for which the frontend model does not account. In the context of the reinforcement learning DDM, the frontend model usually exclusively explains changes in one of the parameters of the backend model, namely the drift rate. However, as we have observed in Manuscript II and III, other parameters such as the threshold, do

not remain constant over time either. Although the experimental paradigms investigated in these two manuscripts were not value-based decision-making tasks, it is plausible that changes in other DDM parameters occur in such paradigms as well. For instance, the threshold parameter might decrease due to motivational factors. In summary, frontend-backend models and superstatistical models can not only be used synergistically as separate tools but can also be unified into one model that opens up interesting future modeling opportunities.

SUPERSTATISTICS IN THE CONTEXT OF EXPERIMENTAL CONDITIONS Now let us go back to the question of applying superstatistics in the context of discerning between experimental conditions. A significant proportion of studies involving cognitive process models are not focused on the dynamic aspects of their components but rather on understanding how experimental manipulations impact specific model parameters. In such cases, there is no necessity for a mechanistic explanation of cognitive model component dynamics through a frontend-backend model. However, such studies could still benefit from estimating non-stationary instead of stationary versions of their cognitive process model. This assertion is grounded in two arguments.

Firstly, assuming stationary parameters when, in reality, the parameters display systematic shifts leads to inflated uncertainty estimates. For instance, consider data generated with a model parameter exhibiting a U-shaped trajectory. Fitting a stationary model to such data would cause the uncertainty of the stationary parameter to align with the marginal parameter distribution, significantly overestimating it. In fact, this phenomenon was observed in one of the human data applications presented in Manuscript II.

Second, disregarding potential dynamics can lead to wrong conclusions about parameter values. For instance, consider a scenario where, over the course of an experiment, the “true” parameter value remains constant in the first half and abruptly shifts to a larger value in the second half. A stationary model would estimate a parameter value that lies in between the values of these two *regimes*. However, the “true” value did not once have this particular value across the duration of the experiment. The estimate of the stationary model, in this case, would be misleading. Such a situation is particularly detrimental when one is interested in absolute parameter values. This is often the case in comparative studies where the goal is to investigate whether a parameter value differs between experimental conditions or groups of people.

Given the imperative for researchers to mitigate inflated uncertainty in parameter estimates and exercise caution in drawing potentially erroneous conclusions from stationary parameter estimates, fitting a non-stationary cognitive model should become the default approach. However, the following question arises: How should superstatistical models be effectively employed in studies exploring the effects of experimental manipulations on specific parameter values?

Consider a simple two-alternative decision-making task where task difficulty is manipulated, and trials with different difficulty levels are randomly interleaved. When employing a NSDDM to interpret the data, we anticipate the drift rate parameter to be influenced by varying task difficulty. In such cases, assuming a single drift rate with a transition function, as employed in Manuscript III, becomes unfeasible. Accurately estimating the drift rate in various difficulty conditions becomes challenging when the parameter substantially changes from trial to trial as a result of randomly altering difficulty conditions.

Thus, we have to somehow incorporate the information of the experimental context into the model. Here we have two options. One approach is to simply assume separate drift rate parameters for each difficulty condition and update them after each trial with a transition function. In fact, this is the approach we used in Manuscript II. This has the disadvantage that it does not provide a direct estimate of the difference in parameter values between conditions. As a solution, after obtaining the parameter trajectories for the different conditions, one can in a second step analyze the difference in the posterior distributions between the conditions.

Alternatively, the context can be directly incorporated into the transition function by introducing additional weight parameters for dummy-coded condition indicator variables, functioning as on-off switches for the particular condition on a given trial. This approach has the advantage that estimates of the static weight parameters serve as a direct estimate of the general differences between conditions.

Another question that frequently arises in the context of superstatistical cognitive models is the following: How can one investigate whether the trajectories of cognitive process model parameters are changing “significantly”? A straightforward approach involves treating the posterior distributions of the entire parameter trajectory as time series data. By fitting a generalized linear model to these time series, one can test specific hypotheses about parameter changes. In this case, the full posterior distribution should be used to propagate the uncertainty of the parameter estimates. A multi-level approach with random intercepts for the different posterior samples of the joint posterior distribution can also be employed to account for the hierarchical structure of the data.

Alternatively, one can incorporate a testable hypothesis directly into the high-level transition function. For instance, to explore whether the parameters exhibit an overall increase or decrease, a transition model like the Gaussian random walk with an additional “drift” parameter could be utilized. Estimating such a parameter enables inference about a generally positive or negative trend of the trajectory. This approach avoids a “two-step approach”, where the parameter trajectory is first inferred, followed by a separate analysis treating parameters as time series data.

6.1 OUTLOOK

In the subsequent discussion, I will explore several aspects that pave the way for future research. Thus far, this dissertation has predominantly focused on the advancement of a novel method for capturing dynamics in model components, particularly in the context of superstatistical models. I demonstrated the applicability of this framework within the realm of cognitive process models, specifically the DDM. Additionally, I substantiated the validity of the inferred dynamics through superstatistical models and provided practical recommendations for their application. While this work significantly contributes to model and method development, it is essential to recognize that the development of models and methods should not be an end in itself. On the contrary, the significance of neural superstatistics will only become clear when they are applied to gain insight into substantive psychological research questions.

SUBSTANTIVE APPLICATIONS Numerous promising applications for superstatistical cognitive models exist. To exemplify this, I hone in on a specific phenomenon frequently observed in recognition memory tasks, where the utilization of a non-stationary DDM provides substantial novel insights.

Recognition memory, denoting the ability to recall previously encountered stimuli (Annis et al., 2013), is often assessed through tasks divided into two distinct phases: a study phase and a test phase. During the study phase, participants familiarize themselves with a set of items. Subsequently, in the test phase, a mixture of new and old items is presented, requiring individuals to discern whether an item was previously encountered or not.

In recognition memory tasks, a common observation is a decline in memory performance during the test phase (Criss et al., 2011; Malmberg et al., 2012; Ratcliff, 1978). Interestingly, the order of items during the testing phase emerges as a more critical predictor of the probability of correct recall than the presentation position during the study phase (Dalezman, 1976). This effect was termed *output interference*, referring to the assumption that the decline in recall performance is attributed to memory interference from output items.

An alternative explanation for the decline in performance across the test phase could be motivational factors. As the time on a task progresses, individuals' motivation may falter and, as a consequence, they may disengage and respond more recklessly (Underwood, 1978). To test these competing explanations, the DDM can be employed to disentangle processing and motivational factors. If the decline in task performance is indeed a result of interference between test items, one can hypothesize a corresponding decline in the drift rate parameter, reflective of mental processing speed. Conversely, if motivational factors are the more important driving force for the decline in performance, a decrease in response caution could be expected, which would be indicated in a decrease in the DDM's threshold parameter.

6 General Discussion

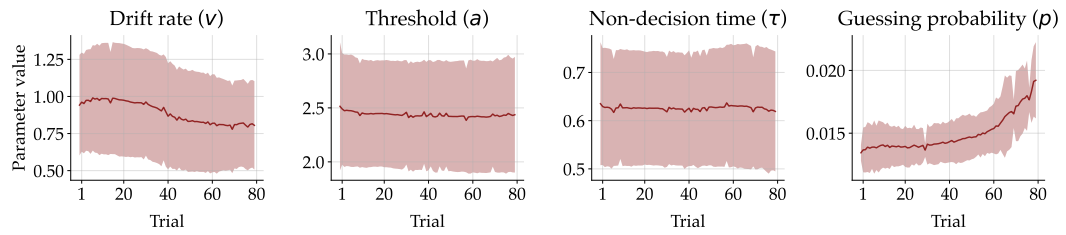


Figure 6.1: Inferred trajectories of the four mixture DDM parameters, the drift rate, threshold, non-decision time, and guessing probability. The solid red lines indicate posterior means averaged across all participants and the shaded region is the corresponding standard deviation.

A previous study, employing a frontend-backend modeling approach, provided evidence that the drift rate parameter indeed decreased over time (Osth et al., 2018). While changes in the threshold parameter greatly varied between individuals, the overall change across the sample suggested no systematic increase or decrease. This implies that the decline in performance was primarily attributed to processing factors rather than motivational influences.

However, alternative perspectives on capturing a decline in motivation exist. It is plausible that individuals do not necessarily reduce their response caution but might resort to more guessing in the later stages of the experiment. To investigate this, a modified DDM that incorporates a mixture between two states – an evidence accumulation process and a guessing process – can be employed (Ratcliff & Kang, 2021; Ratcliff & Tuerlinckx, 2002). This approach allows testing whether individuals increase their probability of performing the task in the guessing state, potentially indicating a decline in motivation.

As a preliminary exploration of this idea, I fitted a non-stationary mixture DDM to unpublished data from 32 participants who completed 80 trials of a recognition memory task. Figure 6.1 depicts the inferred parameter trajectories averaged across all individuals. The results show that the participants showed a tendency to decrease their drift rate, suggesting a decrease in processing speed. The threshold, as well as the non-decision time, remained on average constant over time. This replicates the findings from Osth et al. (2018), implying that performance decreased due to memory interference rather than a decrease in response caution.

Interestingly, the guessing probability parameter on average increased in the second half of the experiment, suggesting that individuals were more likely to randomly respond in the later stages of the task. These findings may challenge previous assertions that the decline in performance is not a result of motivational factors Osth et al. (2018).

It is crucial to approach these results with caution, recognizing the need for a more rigorous study to obtain further evidence. Re-analyzing various data sets with the non-stationary mixture DDM, alongside the previously employed frontend-backend model, and comparing fits to the

data would contribute to a more comprehensive understanding. Additionally, the implementation of a superstatistical frontend-backend model, as previously discussed, offers another avenue for exploration.

Even another possibility to model motivational factors could be explored by utilizing a Lévy-flight model (Rasanan et al., 2023; Wieschen et al., 2020), which modifies the DDM to account for sudden jumps in the evidence accumulation process. This model encompasses an additional parameter that governs the occurrence of such jumps. In this case, the rationale would be that decreasing motivation leads to more frequent premature conclusions, reflected in sudden jumps, during evidence accumulation.

All three manuscripts discussed in this dissertation, along with the previously explored potential application of the superstatistical framework, have centered around the DDM. However, the superstatistics framework we developed is highly versatile. While the DDM aligns seamlessly with this framework, the application is not limited to this specific model. Any cognitive process model, as well as other observation models, can serve as the low-level model in superstatistics. Numerous models stand to benefit from this framework. I will illustrate this point by discussing two potential candidates.

Previously, I briefly touched on reinforcement learning (RL) models in the context of frontend-backend models. In that example, the RL model served as a frontend model informing changes in DDM parameters. However, RL models are frequently used as traditional cognitive process observation models in various studies, particularly in the field of computational psychiatry (Geana et al., 2022; Liebenow et al., 2022; Maia & Frank, 2011). They play a crucial role in understanding differences in error-driven learning between healthy individuals and patients diagnosed with conditions like schizophrenia or addiction disorders (Deserno et al., 2013; Groman et al., 2022; Lim & Ersche, 2023; Montagnese et al., 2020). As previously mentioned, the core parameters of RL models are likely exhibiting non-stationary dynamics over time. Despite this, many studies assume stationarity in these parameters, which can have drastic consequences, particularly when comparing parameter values between different populations. Therefore, integrating RL models into a superstatistics framework presents promising opportunities for advancing fields that heavily depend on these models.

A second potential candidate in which the application of superstatistics enhances predictions is not a specific model but rather a general framework, namely *joint models* (de Hollander et al., 2016; Turner et al., 2017). Models within this framework jointly describe behavioral data and neurophysiological, such as EEG (Nunez et al., 2017, 2022; Schubert et al., 2019) or eye movement (Martinovici et al., 2023). The underlying assumption is that both behavioral and brain measures reflect properties of the same latent cognitive process. The joint analysis of these two data sources enables the testing of theories regarding the relationship between the two modalities.

Moreover, when employing complex cognitive models, behavioral data alone might provide insufficient constraints for parameters (Schubert et al., 2019). In such cases, leveraging information from another modality provides additional constraints to the models since the parameters then need to be able to explain both sources.

Various strategies exist within the joint modeling framework to establish links between brain and behavioral relationships (Palestro et al., 2018; Turner et al., 2017). A common approach is to use a traditional cognitive process model, such as the DDM, to describe behavioral data and utilize brain data to constrain cognitive model parameters. Alternatively, cognitive parameters can be employed to simultaneously predict both brain data and behavioral outcomes (Ghaderi-Kangavari et al., 2023).

However, a commonality among these approaches is that they usually estimate stationary cognitive model parameters. Brain data, on the other hand, often exhibit temporal dependencies and are not independent and identically distributed. Introducing superstatistics to account for non-stationary dynamics in the cognitive parameters of a joint model has the potential to enhance predictions of brain data and provide a more realistic estimate of the relationship between brain data and cognitive constructs.

METHODOLOGICAL REFINEMENT Having discussed potential avenues for the application of superstatistics, I would like to conclude the outlook with a focus on two ideas to further improve the superstatistics framework methodologically. Realizing these improvements will make the application of superstatistics even more meaningful.

Typically, in psychological studies, researchers are interested in the behavior of a population of individuals. In the realm of cognitive modeling, this often means obtaining some global estimates of model parameters and potentially comparing them with those from another group of individuals. While various methods exist to infer parameters from a population, *hierarchical modeling*¹ is currently considered state-of-the-art (Kupitz, 2020; Lee, 2011).

Hierarchical models simultaneously fit data from all individuals, estimating individual-level and group-level parameters concurrently. These models assume that individual-level parameters stem from a group-level distribution, typically modeled as a Gaussian distribution. The key advantage of this approach lies in pooling data across individuals to explicitly estimate the mean and variance of the population. The group-level estimates, in turn, constrain the parameter estimation of each individual, mitigating the risk of extreme values – a property known as *shrinkage* (McElreath, 2020). Studies, both *in vivo* and *in silico*, have demonstrated that this approach leads to more accurate uncertainty quantification and less biased inference (Boehm et al., 2018; Vandekerckhove, 2014). Moreover, by fitting data from all participants simultaneously, fewer data points

¹also known as *multi-level modeling*, *mixed modeling*, or *partial pooling*

per person are required to obtain reasonably certain parameter estimates (Lee & Wagenmakers, 2013).

In contrast, our approach in applying superstatistics so far involved fitting the model to each individual separately and subsequently aggregating the inferred parameter trajectories across individuals. A drawback of this approach, compared to the hierarchical method, is the potential for extreme individual parameter values due to the absence of shrinkage. Consequently, the population variance tends to be inflated, which is undesirable when comparing different groups of individuals (Zhang et al., 2020).

Hence, it might be worthwhile to explore hierarchical superstatistical cognitive process models in cases where researchers are interested in parameter trajectories across a population. Additionally, such a method could prove beneficial when the time series to be modeled contains relatively few time steps since it then enhances the certainty of parameter estimates. However, the benefits of hierarchical modeling (i.e., shrinkage) need to be weighed against the additional computational burden introduced, considering that non-hierarchical fitting of superstatistical cognitive process models is already challenging and computationally intensive. While it should be theoretically possible to fit such hierarchical superstatistics, further research is necessary to test its feasibility, both statistically and computationally.

Another aspect of our superstatistics estimation method that requires further attention is the factorization of the joint posterior distribution when estimating time-varying parameters. The most common factorizations are the *filtering* and *smoothing* distributions (Mark et al., 2018; Särkkä, 2013). The filtering distribution employs online analysis, where the estimation of low-level parameters at a specific time point is conditioned only on past data. Conversely, the smoothing distribution informs the low-level parameters at a given time step based on both past and future data points.

We remain uncertain about which of these two factorizations, or perhaps another one, works best for superstatistics with cognitive process models. In Manuscript II, we targeted the filtering distribution, while in Manuscript III, we switched to the smoothing distribution, a decision based more on intuition than evidence. To gain a better understanding, I conducted a preliminary simulation study.

I trained two neural approximators – one targeting the filtering distribution and the other the smoothing distribution – to estimate a NSDDM. Both models were then fitted to 500 data sets, each consisting of 800 time steps. Subsequently, I compared the two approaches across all time steps for the three low-level parameters: drift rate, threshold, and non-decision time.

Figure 6.2A illustrates the R^2 -Scores between estimated and true data-generating parameters. While both methods demonstrate excellent recovery of the non-decision time, the smoothing distribution outperforms the filtering distribution in recovering the drift rate and threshold param-

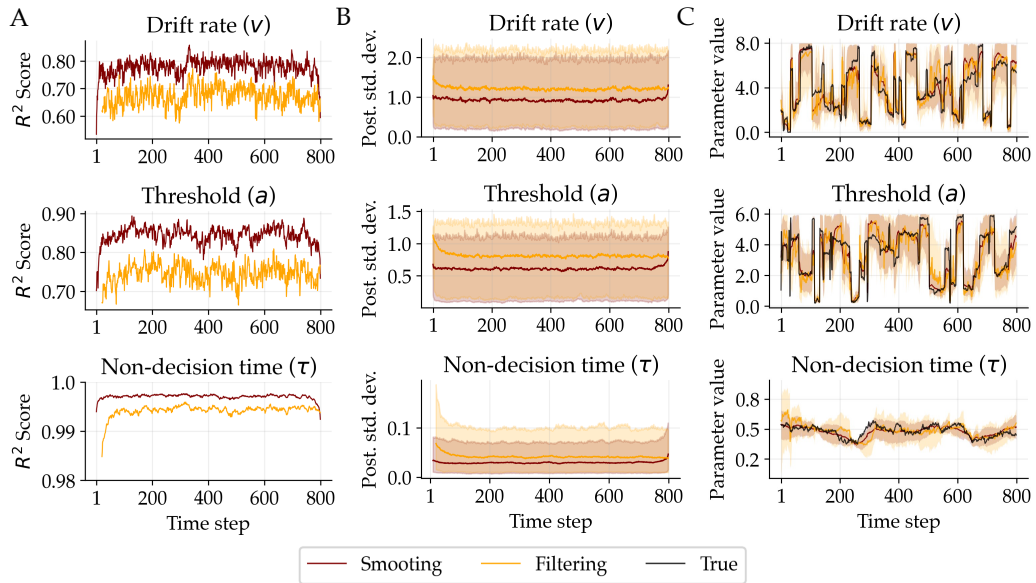


Figure 6.2: Comparison of smoothing (red) and filtering (yellow) posterior distributions across 800 time steps for three model parameters separately based on 500 simulations. **A** Parameter recovery performance as R^2 -Scores between estimated and true data-generating parameters. **B** Median and 95% credibility interval of the posterior standard deviation. **C** Exemplar data-generating parameter trajectory in black along colored recoveries.

ters. Similarly, the smoothing posterior distribution consistently exhibits a smaller posterior standard deviation, indicative of a less uncertain estimate, compared to the filtering distribution (see [Figure 6.2B](#)). For illustrative purposes, an exemplar data-generating parameter trajectory along with its recovery is depicted in [Figure 6.2C](#). Both methods can recover the overall trajectory of all three parameters, but the filtering posterior distribution shows slightly more deviations from the ground truth.

From this initial investigation, it appears that the smoothing distribution is both more accurate and more certain in estimating non-stationary parameters. Nevertheless, these results should be interpreted with caution, and a more thorough study is needed to draw a conclusion. In conclusion, it becomes evident that the methodologies developed in this dissertation hold promise for numerous future applications. At the same time, realizing the full potential of superstatistics in cognitive science necessitates additional studies that delve deeper into its methodological nuances.

6.2 CONCLUDING REMARKS

Making sense of human cognition and behavior is undeniably hard. For decades, cognitive science has tirelessly developed cognitive process models to bridge the gap between theoretical con-

structs and observable behavior. These models have continuously evolved in complexity, striving for greater realism.

The aim of this dissertation was to propel cognitive process models one step closer to reality, specifically by considering the dynamic nature of the human mind. By acknowledging and integrating dynamics into cognitive process models, we hope not only to circumvent erroneous conclusions stemming from static assumptions but also to gain deeper insights into the complexities of the human mind. I hope to have persuaded the audience that superstatistics is an incredibly versatile modeling framework with immense potential to achieve this goal.

While the concept of superstatistical models is not entirely novel, this dissertation sought to bring this method closer to the realm of cognitive science, paving the way for innovative applications and discoveries. Moreover, this dissertation addressed the inherent challenges associated with estimating parameters in dynamic cognitive models, providing practical solutions to enhance their applicability and utility.

The main focus of the present dissertation revolved around developing and validating a dynamic modeling framework. As always when something new is developed: if the aspiration is it to be useful, much detail and effort must be invested in the development phase. That is exactly what the work in this dissertation aimed at. The works presented here adhered to a state-of-the-art principled workflow within the Bayesian framework. By adhering to these standards, this dissertation aimed not only to contribute to the advancement of cognitive science but also to inspire confidence in the broader scientific community regarding the validity and utility of dynamic modeling approaches.

Looking ahead, the next phase of this research involves translating these advancements into practical applications. While the methods developed in this work hold vast promise, their usability and accessibility may present initial challenges. Therefore, efforts must be directed toward making these tools more user-friendly and readily available to researchers and practitioners in psychology, cognitive science, and related fields.

In conclusion, the methods developed in this dissertation have the potential for a wide range of applications in psychology, cognitive science, and beyond. Superstatistics, along with frontend-backend models, will play a pivotal role in these research fields, not only for explanatory and descriptive purposes but also for the prediction of future human behavior.

BIBLIOGRAPHY

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. <https://doi.org/10.1037/0033-295X.111.4.1036>
- Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1365–1376. <https://doi.org/10.1037/a0032188>
- Arnold, N. R., Bröder, A., & Bayen, U. J. (2015). Empirical validation of the diffusion model for recognition memory and a comparison of parameter-estimation methods. *Psychological Research*, *79*(5), 882–898. <https://doi.org/10.1007/s00426-014-0608-y>
- Ballard, I. C., & McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, *317*, 37–44. <https://doi.org/10.1016/j.jneumeth.2019.01.006>
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57–86. <https://doi.org/10.3758/BF03210812>
- Bausenhardt, K. M., Dyjas, O., & Ulrich, R. (2014). Temporal reproductions are influenced by an internal reference: Explaining the Vierordt effect. *Acta Psychologica*, *147*, 60–67. <https://doi.org/10.1016/j.actpsy.2013.06.011>
- Bausenhardt, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics*, *77*(4), 1033–1043. <https://doi.org/10.3758/s13414-015-0875-8>
- Beck, C., & Cohen, E. G. D. (2003). Superstatistics. *Physica A: Statistical Mechanics and its Applications*, *322*, 267–275. [https://doi.org/10.1016/S0378-4371\(03\)00019-0](https://doi.org/10.1016/S0378-4371(03)00019-0)
- Bode, S., Sewell, D. K., Lilburn, S., Forte, J. D., Smith, P. L., & Stahl, J. (2012). Predicting perceptual decision biases from early brain activity. *Journal of Neuroscience*, *32*(36), 12488–12498. <https://doi.org/10.1523/JNEUROSCI.1708-12.2012>
- Boehm, U., Marsman, M., Matzke, D., & Wagenmakers, E.-J. (2018). On the importance of avoiding shortcuts in applying cognitive models to hierarchical data. *Behavior Research Methods*, *50*(4), 1614–1631. <https://doi.org/10.3758/s13428-018-1054-3>

Bibliography

- Boehm, U., van Maanen, L., Forstmann, B., & van Rijn, H. (2014). Trial-by-trial fluctuations in CNV amplitude reflect anticipatory adjustment of response caution. *NeuroImage*, *96*, 95–105. <https://doi.org/10.1016/j.neuroimage.2014.03.063>
- Briggs, K., & Beck, C. (2007). Modelling train delays with q-exponential functions. *Physica A: Statistical Mechanics and its Applications*, *378*(2), 498–504. <https://doi.org/10.1016/j.physa.2006.11.084>
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. <https://doi.org/10.1016/j.cogpsych.2007.12.002>
- Burgess, N., & Hitch, G. (2005). Computational models of working memory: Putting long-term memory into context. *Trends in Cognitive Sciences*, *9*(11), 535–541. <https://doi.org/10.1016/j.tics.2005.09.011>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Churchland, A. K., & Ditterich, J. (2012). New advances in understanding decisions among multiple alternatives. *Current Opinion in Neurobiology*, *22*(6), 920–926. <https://doi.org/10.1016/j.conb.2012.04.009>
- Cochrane, A., Sims, C. R., Bejjanki, V. R., Green, C. S., & Bavelier, D. (2023). Multiple timescales of learning indicated by changes in evidence-accumulation processes during perceptual decision-making. *npj Science of Learning*, *8*(1), 1–10. <https://doi.org/10.1038/s41539-023-00168-9>
- Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062. <https://doi.org/10.1073/pnas.1912789117>
- Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*(4), 316–326. <https://doi.org/10.1016/j.jml.2011.02.003>
- Dalezman, J. J. (1976). Effects of output order on immediate, delayed, and final recall performance. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 597–608. <https://doi.org/10.1037/0278-7393.2.5.597>
- Damaso, K. A. M., Williams, P. G., & Heathcote, A. (2022). What happens after a fast versus slow error, and how does it relate to evidence accumulation? *Computational Brain & Behavior*, *5*(4), 527–546. <https://doi.org/10.1007/s42113-022-00137-2>
- de Hollander, G., Forstmann, B. U., & Brown, S. D. (2016). Different ways of linking behavioral and neural data via computational cognitive models. *Biological Psychiatry: Cognitive*

- tive Neuroscience and Neuroimaging*, 1(2), 101–109. <https://doi.org/10.1016/j.bpsc.2015.11.004>
- de Jong, J., Akyürek, E. G., & van Rijn, H. (2021). A common dynamic prior for time in duration discrimination. *Psychonomic Bulletin & Review*, 28(4), 1183–1190. <https://doi.org/10.3758/s13423-021-01887-z>
- Denys, M., Gubiec, T., Kutner, R., Jagielski, M., & Stanley, H. E. (2016). Universality of market superstatistics. *Physical Review E*, 94(4), Article 042305. <https://doi.org/10.1103/PhysRevE.94.042305>
- Deserno, L., Boehme, R., Heinz, A., & Schlagenhauf, F. (2013). Reinforcement learning and dopamine in schizophrenia: Dimensions of symptoms or specific features of a disease group? *Frontiers in Psychiatry*, 4, Article 172. <https://doi.org/doi.org/10.3389/fpsy.2013.00172>
- Dutilh, G., Forstmann, B. U., Vandekerckhove, J., & Wagenmakers, E.-J. (2013). A diffusion model account of age differences in posterror slowing. *Psychology and Aging*, 28(1), 64–76. <https://doi.org/10.1037/a0029875>
- Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception & Psychophysics*, 74(2), 454–465. <https://doi.org/10.3758/s13414-011-0243-2>
- Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics*, 74(8), 1819–1841. <https://doi.org/10.3758/s13414-012-0362-4>
- Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2014). Effects of stimulus order on duration discrimination sensitivity are under attentional control. *Journal of Experimental Psychology: Human Perception and Performance*, 40(1), 292–307. <https://doi.org/10.1037/a0033611>
- Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, 41, 128–137. <https://doi.org/10.1016/j.cobeha.2021.06.004>
- Ellinghaus, R., Ulrich, R., & Bausenhart, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance*, 44(1), 7–12. <https://doi.org/10.1037/xhp0000495>
- Els Müller, L., Schnuerch, M., Bürkner, P.-C., & Radev, S. T. (2023). A deep learning method for comparing bayesian hierarchical models. *arXiv*. <https://doi.org/10.48550/arXiv.2301.11873>

Bibliography

- Evans, N., & Wagenmakers, E.-J. (2020). Evidence Accumulation Models: Current Limitations and Future Directions. *The Quantitative Methods for Psychology*, *16*, 73–90. <https://doi.org/10.20982/tqmp.16.2.p073>
- Evans, N. J., & Brown, S. D. (2017). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review*, *24*(2), 597–606. <https://doi.org/10.3758/s13423-016-1135-1>
- Farrell, S., & Lewandowsky, S. (2018). *Computational Modeling of Cognition and Behavior*. Cambridge University Press. <https://doi.org/10.1017/CBO9781316272503>
- Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Härtel.
- Feng, S. F., Wang, S., Zarnescu, S., & Wilson, R. C. (2021). The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific Reports*, *11*(1), 3077. <https://doi.org/10.1038/s41598-021-82530-8>
- Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, *26*(4), 1099–1121. <https://doi.org/10.3758/s13423-018-1554-2>
- Geana, A., Barch, D. M., Gold, J. M., Carter, C. S., MacDonald, A. W., Ragland, J. D., Silverstein, S. M., & Frank, M. J. (2022). Using computational modeling to capture schizophrenia-specific reinforcement learning differences and their implications on patient classification. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *7*(10), 1035–1046. <https://doi.org/10.1016/j.bpsc.2021.03.017>
- Ghaderi-Kangavari, A., Rad, J. A., & Nunez, M. D. (2023). A general integrative neurocognitive modeling framework to jointly describe eeg and decision-making on single trials. *Computational Brain & Behavior*, *6*(3), 317–376. <https://doi.org/10.1007/s42113-023-00167-4>
- Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574. <https://doi.org/10.1146/annurev.neuro.29.051605.113038>
- Groman, S. M., Thompson, S. L., Lee, D., & Taylor, J. R. (2022). Reinforcement learning de-tuned in addiction: integrative and translational approaches. *Trends in neurosciences*, *45*(2), 96–105. <https://doi.org/10.1016/j.tins.2021.11.007>
- Gunawan, D., Hawkins, G. E., Kohn, R., Tran, M.-N., & Brown, S. D. (2022). Time-evolving psychological processes over repeated decisions. *Psychological Review*, *129*(3), 438. <https://doi.org/doi.org/10.1037/rev0000351>
- Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2023). Theoretically informed generative models can advance the psy-

- chological and brain sciences: Lessons from the reliability paradox. *PsyArXiv*. <https://doi.org/10.31234/osf.io/xr7y3>
- Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8, Article 150. <https://doi.org/10.3389/fnins.2014.00150>
- Hellström, A. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology. Human Perception and Performance*, 5(3), 460–477. <https://doi.org/10.1037/0096-1523.5.3.460>
- Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, 97(1), 35–61. <https://doi.org/10.1037/0033-2909.97.1.35>
- Hellström, Å., Patching, G. R., & Rammsayer, T. H. (2020). Sensation weighting in duration discrimination: A univariate, multivariate, and varied-design study of presentation-order effects. *Attention, Perception, & Psychophysics*, 82(6), 3196–3220. <https://doi.org/10.3758/s13414-020-01999-z>
- Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), 135–175.
- Jamieson, D. G., & Petrusic, W. M. (1975). Presentation order effects in duration discrimination. *Perception & Psychophysics*, 17(2), 197–202. <https://doi.org/10.3758/BF03203886>
- Jepma, M., Schaaf, J. V., Visser, I., & Huizenga, H. M. (2020). Uncertainty-driven regulation of learning and exploration in adolescents: A computational account. *PLoS Computational Biology*, 16(9), Article e1008276. <https://doi.org/10.1371/journal.pcbi.1008276>
- Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, 93(3), 353–368. <https://doi.org/10.1037/0022-3514.93.3.353>
- Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9), 1148–1160. <https://doi.org/10.1038/s41593-018-0210-5>
- Kupitz, C. N. (2020). *Applications of Hierarchical Bayesian Cognitive Modeling* (Doctoral dissertation). UC Irvine.
- Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica*, 43(3), 199–224. [https://doi.org/10.1016/0001-6918\(79\)90026-X](https://doi.org/10.1016/0001-6918(79)90026-X)
- Lapid, E., Ulrich, R., & Rammsayer, T. (2008). On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task. *Perception & Psychophysics*, 70(2), 291–305. <https://doi.org/10.3758/PP.70.2.291>
- Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55(1), 1–7. <https://doi.org/10.1016/j.jmp.2010.08.013>

Bibliography

- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139087759>
- Leon Chen, L., & Beck, C. (2008). A superstatistical model of metastasis and cancer survival. *Physica A: Statistical Mechanics and its Applications*, 387(13), 3162–3172. <https://doi.org/10.1016/j.physa.2008.01.116>
- Lerche, V., & Voss, A. (2019). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychological Research*, 83(6), 1194–1209. <https://doi.org/10.1007/s00426-017-0945-8>
- Liebenow, B., Jones, R., DiMarco, E., Trattner, J. D., Humphries, J., Sands, L. P., Spry, K. P., Johnson, C. K., Farkas, E. B., Jiang, A., & Kishida, K. T. (2022). Computational reinforcement learning, reward (and punishment), and dopamine in psychiatric disorders. *Frontiers in Psychiatry*, 13, Article 886297. <https://doi.org/doi.org/10.3389/fpsy.2022.886297>
- Lim, T. V., & Ersche, K. D. (2023). Theory-driven computational models of drug addiction in humans: Fruitful or futile? *Addiction Neuroscience*, 5, 100066. <https://doi.org/10.1016/j.addicn.2023.100066>
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.
- Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, 14(2), 154–162. <https://doi.org/10.1038/nn.2723>
- Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science*, 23(2), 115–119. <https://doi.org/10.1177/0956797611430692>
- Mark, C., Metzner, C., Lautscham, L., Strissel, P. L., Strick, R., & Fabry, B. (2018). Bayesian model selection for complex dynamic systems. *Nature Communications*, 9(1), 1803. <https://doi.org/10.1038/s41467-018-04241-5>
- Martinovici, A., Pieters, R., & Erdem, T. (2023). Attention trajectories capture utility accumulation and predict brand choice. *Journal of Marketing Research*, 60(4), 625–645. <https://doi.org/10.1177/00222437221141052>
- McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1), 11–38. <https://doi.org/10.1111/j.1756-8765.2008.01003.x>
- McDougle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, 28(1), 20–39. <https://doi.org/10.3758/s13423-020-01774-z>

- Mcelreath, R. (2020). *Statistical rethinking: A bayesian course with examples in R and STAN* (2. edition). Taylor & Francis.
- Metzner, C., Schilling, A., Traxdorf, M., Schulze, H., & Krauss, P. (2021). Sleep as a random walk: A super-statistical analysis of EEG data across sleep stages. *Communications Biology*, 4(1), 1–11. <https://doi.org/10.1038/s42003-021-02912-6>
- Miletić, S., Boag, R. J., & Forstmann, B. U. (2020). Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia*, 136, Article 107261. <https://doi.org/10.1016/j.neuropsychologia.2019.107261>
- Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks. *eLife*, 10, Article e63055. <https://doi.org/10.7554/eLife.63055>
- Millner, A. J., Gershman, S. J., Nock, M. K., & den Ouden, H. E. M. (2018). Pavlovian control of escape and avoidance. *Journal of Cognitive Neuroscience*, 30(10), 1379–1390. https://doi.org/10.1162/jocn_a_01224
- Montagnese, M., Knolle, F., Haarsma, J., Griffin, J. D., Richards, A., Vertes, P. E., Kiddle, B., Fletcher, P. C., Jones, P. B., Owen, M. J., Fonagy, P., Bullmore, E. T., Dolan, R. J., Moutoussis, M., Goodyer, I. M., & Murray, G. K. (2020). Reinforcement learning as an intermediate phenotype in psychosis? Deficits sensitive to illness stage but not associated with polygenic risk of schizophrenia in the general population. *Schizophrenia Research*, 222, 389–396. <https://doi.org/10.1016/j.schres.2020.04.022>
- Mormann, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, 5(6), 437–449. <https://doi.org/10.2139/ssrn.1901533>
- Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 32(7), 2335–2343. <https://doi.org/10.1523/JNEUROSCI.4156-11.2012>
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), 90–100. [https://doi.org/10.1016/S0022-2496\(02\)00028-7](https://doi.org/10.1016/S0022-2496(02)00028-7)
- Nachmias, J. (2006). The role of virtual standards in visual discrimination. *Vision Research*, 46(15), 2456–2464. <https://doi.org/10.1016/j.visres.2006.01.029>
- Neal, R. M. (2011). *MCMC using Hamiltonian dynamics* (Vol. 2). Chapman; Hall/CRC. <https://doi.org/10.1201/b10905-7>
- Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal*

Bibliography

- of mathematical psychology*, 76(Pt B), 117–130. <https://doi.org/10.1016/j.jmp.2016.03.003>
- Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2022). A tutorial on fitting joint models of M/EEG and behavior to understand cognition. *PsyArXiv*. <https://doi.org/10.31234/osf.io/vf6t5>
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596–1618. <https://doi.org/10.3758/s13423-019-01645-2>
- O’Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, 15(12), 1729–1735. <https://doi.org/10.1038/nn.3248>
- Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology*, 104, 106–142. <https://doi.org/10.1016/j.cogpsych.2018.04.002>
- Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, 84, 20–48. <https://doi.org/10.1016/j.jmp.2018.03.003>
- Palmeri, T. J., Love, B. C., & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76(Pt B), 59–64. <https://doi.org/10.1016/j.jmp.2016.10.010>
- Pedersen, M. L., & Frank, M. J. (2020). Simultaneous hierarchical bayesian parameter estimation for reinforcement learning and drift diffusion models: A tutorial and links to neural data. *Computational Brain & Behavior*, 3(4), 458–471. <https://doi.org/10.1007/s42113-020-00084-w>
- Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10(2), 59–63. <https://doi.org/10.1016/j.tics.2005.12.004>
- Purcell, B. A., & Kiani, R. (2016). Neural mechanisms of post-error adjustments of decision policy in parietal cortex. *Neuron*, 89(3), 658–671. <https://doi.org/10.1016/j.neuron.2015.12.027>
- Rabassa, P., & Beck, C. (2015). Superstatistical analysis of sea-level fluctuations. *Physica A: Statistical Mechanics and its Applications*, 417, 18–28. <https://doi.org/10.1016/j.physa.2014.08.068>
- Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? An analysis of response programming. *Quarterly Journal of Experimental Psychology*, 29(4), 727–743. <https://doi.org/10.1080/14640747708400645>
- Radev, S. T., D’Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., & Bürkner, P.-C. (2021). Amortized Bayesian model comparison with evidential deep learning. *arxiv*. <https://doi.org/10.48550/arXiv.2004.10629>

- Radev, S. T., Voss, A., Wieschen, E. M., & Bürkner, P.-C. (2020). Amortized Bayesian Inference for Models of Cognition. *arXiv*. <https://doi.org/10.48550/arXiv.2005.03899>
- Rasanan, A. H. H., Rad, J. A., & Sewell, D. K. (2023). Are there jumps in evidence accumulation, and what, if anything, do they reflect psychologically? An analysis of Lévy Flights models of decision-making. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-023-02284-4>
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. <https://doi.org/10.1037/0033-295X.85.2.59>
- Ratcliff, R., & Kang, I. (2021). Qualitative speed-accuracy tradeoff effects can be explained by a diffusion/fast-guess mixture model. *Scientific Reports*, *11*(1), Article 15169. <https://doi.org/10.1038/s41598-021-94451-7>
- Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>
- Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190–214. <https://doi.org/10.1037/0033-295X.83.3.190>
- Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. <https://doi.org/10.1111/1467-9280.00067>
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>
- Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481. <https://doi.org/10.3758/BF03196302>
- Rizzo, S., Rapisarda, A., & CACTUS Group. (2004). Environmental atmospheric turbulence at florence airport. *AIP Conference Proceedings*, *742*(1), 176–181. <https://doi.org/10.1063/1.1846475>
- Ross, H. E., & Gregory, R. L. (1964). Is the Weber fraction a Function of physical or perceived input? *Quarterly Journal of Experimental Psychology*, *16*(2), 116–122. <https://doi.org/10.1080/17470216408416356>
- Säilynoja, T., Bürkner, P.-C., & Vehtari, A. (2021). Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. *arXiv*. <https://doi.org/10.48550/arxiv.2103.10522>
- Santhanam, M. S., & Kantz, H. (2008). Return interval distribution of extreme events and long-term memory. *Physical Review E*, *78*(5), Article 051113. <https://doi.org/10.1103/PhysRevE.78.051113>

- Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139344203>
- Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, 26(1), 103–126. <https://doi.org/10.1037/met0000275>
- Schiffler, B. C., Bengtsson, S. L., & Lundqvist, D. (2017). The sustained influence of an error on future decision-making. *Frontiers in Psychology*, 8, Article 1077. <https://doi.org/doi.org/10.3389/fpsyg.2017.01077>
- Schubert, A.-L., Nunez, M. D., Hagemann, D., & Vandekerckhove, J. (2019). Individual differences in cortical processing speed predict cognitive abilities: A model-based cognitive neuroscience account. *Computational Brain & Behavior*, 2(2), 64–84. <https://doi.org/10.1007/s42113-018-0021-5>
- Schumacher, L., Bürkner, P.-C., Voss, A., Köthe, U., & Radev, S. T. (2023). Neural superstatistics for Bayesian estimation of dynamic cognitive models. *Scientific Reports*, 13(1), Article 13778. <https://doi.org/10.1038/s41598-023-40278-3>
- Schumacher, L., Schnuerch, M., Voss, A., & Radev, S. T. (2023). Validation and comparison of non-stationary cognitive models: A diffusion model application. *arxiv*. <https://doi.org/10.48550/arXiv.2401.08626>
- Schumacher, L., & Voss, A. (2023). Duration discrimination: A diffusion decision modeling approach. *Attention, Perception, & Psychophysics*, 85(2), 560–577. <https://doi.org/10.3758/s13414-022-02604-1>
- Sewell, D. K., Jach, H. K., Boag, R. J., & Van Heer, C. A. (2019). Combining error-driven models of associative learning with evidence accumulation models of decision-making. *Psychonomic Bulletin & Review*, 26(3), 868–893. <https://doi.org/10.3758/s13423-019-01570-4>
- Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS Computational Biology*, 15(2), Article e1006803. <https://doi.org/10.1371/journal.pcbi.1006803>
- Stoevesandt, B., & Peinke, J. (2010). Effects of sudden changes in inflow conditions on the angle of attack on hawt blades. *arxiv*. <https://doi.org/10.48550/arXiv.1011.5396>
- Sun, R. (2016). *Anatomy of the mind: Exploring psychological mechanisms and processes with the Clarion cognitive architecture*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199794553.001.0001>
- Sutton, R. S., & Barto, A. (2018). *Reinforcement learning an introduction* (Second edition). The MIT Press.

- Swets, J. A., & Green, D. M. (1978). *Applications of signal detection theory*. Springer US. https://doi.org/10.1007/978-1-4684-2487-4_19
- Szollosi, A., & Donkin, C. (2019). Neglected sources of flexibility in psychological theories: From replicability to good explanations. *Computational Brain & Behavior*, 2(3), 190–192. <https://doi.org/10.1007/s42113-019-00045-y>
- Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, 7(1), Article 12400. <https://doi.org/10.1038/ncomms12400>
- Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv*. <https://doi.org/10.48550/arxiv.1804.06788>
- Theisen, M., Lerche, V., von Krause, M., & Voss, A. (2021). Age differences in diffusion model parameters: A meta-analysis. *Psychological Research*, 85(5), 2012–2021. <https://doi.org/10.1007/s00426-020-01371-8>
- Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. <https://doi.org/10.1037/h0070288>
- Thurstone, L. L. (1927b). Psychophysical analysis. *The American Journal of Psychology*, 38(3), 368–389. <https://doi.org/10.2307/1415006>
- Tillman, G., Van Zandt, T., & Logan, G. D. (2020). Sequential sampling models without random between-trial variability: The racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review*, 27(5), 911–936. <https://doi.org/10.3758/s13423-020-01719-6>
- Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76, 65–79. <https://doi.org/10.1016/j.jmp.2016.01.001>
- Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators. *Attention, Perception, & Psychophysics*, 71(6), 1219–1227. <https://doi.org/10.3758/APP.71.6.1219>
- Underwood, B. J. (1978). Recognition memory as a function of length of study list. *Bulletin of the Psychonomic Society*, 12(2), 89–91. <https://doi.org/10.3758/BF03329636>
- Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3), 550–592. <https://doi.org/10.1037/0033-295X.108.3.550>
- Van der Straeten, E., & Beck, C. (2009). Superstatistical fluctuations in time series: Applications to share-price dynamics and turbulence. *Physical Review E*, 80(3), Article 036108. <https://doi.org/10.1103/PhysRevE.80.036108>

- van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review*, *25*(1), 143–154. <https://doi.org/10.3758/s13423-016-1015-8>
- van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories. *Social Psychology*, *51*(5), 285–298. <https://doi.org/10.1027/1864-9335/a000428>
- Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71. <https://doi.org/10.1016/j.jmp.2014.06.004>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*, 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- von Krause, M., Radev, S. T., & Voss, A. (2022). Mental speed is high until age 60 as revealed by analysis of over a million participants. *Nature Human Behaviour*, *6*(5), 700–708. <https://doi.org/10.1038/s41562-021-01282-7>
- von Krause, M., Radev, S. T., Voss, A., Quintus, M., Egloff, B., & Wrzus, C. (2021). Stability and Change in Diffusion Model Parameters over Two Years. *Journal of Intelligence*, *9*(2), 26. <https://doi.org/10.3390/jintelligence9020026>
- Voss, A., Lerche, V., Mertens, U., & Voss, J. (2019). Sequential sampling models with variable boundaries and non-normal noise: A comparison of six models. *Psychonomic Bulletin & Review*, *26*(3), 813–832. <https://doi.org/10.3758/s13423-018-1560-4>
- Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental Psychology*, *60*(6), 385–402. <https://doi.org/10.1027/1618-3169/a000218>
- Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220. <https://doi.org/10.3758/BF03196893>
- Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775. <https://doi.org/10.3758/BF03192967>
- Wieschen, E. M., Voss, A., & Radev, S. (2020). Jumping to conclusion? A Lévy flight model of decision making. *The Quantitative Methods for Psychology*, *16*(2), 120–132. <https://doi.org/10.20982/tqmp.16.2.p120>
- Williams, G., Schäfer, B., & Beck, C. (2020). Superstatistical approach to air pollution statistics. *Physical Review Research*, *2*(1), Article 013019. <https://doi.org/10.1103/PhysRevResearch.2.013019>
- Yalcin, G. C., & Beck, C. (2013). Environmental superstatistics. *arxiv*, (arXiv:1212.5783). <https://doi.org/10.48550/arXiv.1212.5783>

- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, Article e1. <https://doi.org/10.1017/S0140525X20001685>
- Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, 15(6), 695–707. <https://doi.org/10.1093/scan/nsaa089>




Promotionsausschuss der Fakultät für Verhaltens- und Empirische Kulturwissenschaften der Ruprecht-Karls-Universität Heidelberg / Doctoral Committee of the Faculty of Behavioural and Cultural Studies of Heidelberg University

Erklärung gemäß § 8 (1) c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften / Declaration in accordance to § 8 (1) c) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe. / I declare that I have made the submitted dissertation independently, using only the specified tools and have correctly marked all quotations.

Erklärung gemäß § 8 (1) d) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften / Declaration in accordance to § 8 (1) d) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe. / I declare that I did not use the submitted dissertation in this or any other form as an examination paper until now and that I did not submit it in another faculty.

Vorname Nachname / First name Family name	Lukas Schumacher
Datum / Date	06.03.2024
Unterschrift / Signature	

APPENDIX A1 - MANUSCRIPT I

Manuscript I: Duration Discrimination: A Diffusion Decision Modeling Approach



Duration discrimination: A diffusion decision modeling approach

Lukas Schumacher¹ · Andreas Voss¹

Accepted: 16 October 2022 / Published online: 23 January 2023
© The Author(s) 2023

Abstract

The human ability to discriminate the duration of two subsequently presented stimuli is often studied with tasks that involve a comparison between a standard stimulus (with fixed duration) and comparison stimuli (with varying durations). The performance in such tasks is influenced by the presentation order of these successively presented stimuli. The so-called Type A effect refers to the impact of presentation order on the point of subjective equality. The Type B effect describes effects of presentation order on the just-noticeable-difference. Cognitive models that account for these context effects assume that participants' duration estimation is influenced by the history of previously encountered stimuli. For example, the internal reference model assumes that the magnitude of a “typical” stimulus is represented by an internal reference. This internal reference evolves throughout an experiment and is updated on every trial. Different recent models have in common that they describe how the internal reference is computed but are agnostic to the decision process itself. In this study, we develop a new model that incorporates the mechanisms of perceptual discrimination models into a diffusion model. The diffusion model focuses on the dynamics of the decision process itself and accounts for choice and response times based on a set of latent cognitive variables. We show that our model accurately predicts the accuracy and response time distribution in a classical duration discrimination task. Further, model parameters were sensitive to the Type A and B effect. The proposed model opens up new opportunities for studying human discrimination performance (e.g., individual differences).

Keywords Diffusion decision model · Duration discrimination · Context effects

Introduction

Comparative decisions are fundamental in humans' everyday lives and have been extensively studied since the advent of psychophysics (Fechner, 1860; Hegelmaier, 1853). In a typical experiment, participants have to select one of two stimuli based on the magnitude of a specific stimulus feature. For instance, deciding which of two subsequently presented tones had a longer duration (see, e.g., Fig. 1A). A class of psychophysical models, called *difference models* (Thurstone, 1927a, b), assumes that participants compare their internal representation of the two presented stimuli and base their decision on the difference in magnitude between these internal representations, $D = X_1 - X_2$.

Usually, this difference in stimulus magnitude (e.g., duration of tones) is experimentally manipulated by varying

the intensity of one stimuli between trials. What is expected is that the difficulty of the decision depends on the difference between the two stimulus intensities. Deciding between stimuli with a relative large difference in intensity is easier than compared to stimuli that are very similar. Participants' performance in such tasks can be described with a (*sigmoidal*) psychometric function that maps varying stimulus intensities to the proportion of a certain response. The steepness of the slope of this psychometric function indicates the individual's sensitivity to differences in stimulus magnitude (see, e.g., Fig. 1B).

Many studies have shown that discrimination performance (often indexed by the *difference limen*; DL^1) in such tasks is not only influenced by the stimulus intensities but also by task-irrelevant features of the experimental context (for a recent review see, Bausenhardt et al., 2016) – effects that classical difference models cannot explain. Therefore,

✉ Lukas Schumacher
lukas.schumacher@psychologie.uni-heidelberg.de

¹ Institut of Psychology, Department of Quantitative Research Methods, Heidelberg University, Hauptstrasse 47-51, 69117 Heidelberg Germany

¹The DL , also called *just noticeable difference* (JND), is usually defined as the difference between the 75% and 25% percentile of a psychometric function divided by 2, $DL = \frac{x_{75} - x_{25}}{2}$, where x denotes the magnitude of a variable stimulus. This is a measure of the slope of the psychometric function where smaller values reflect better discrimination performance.

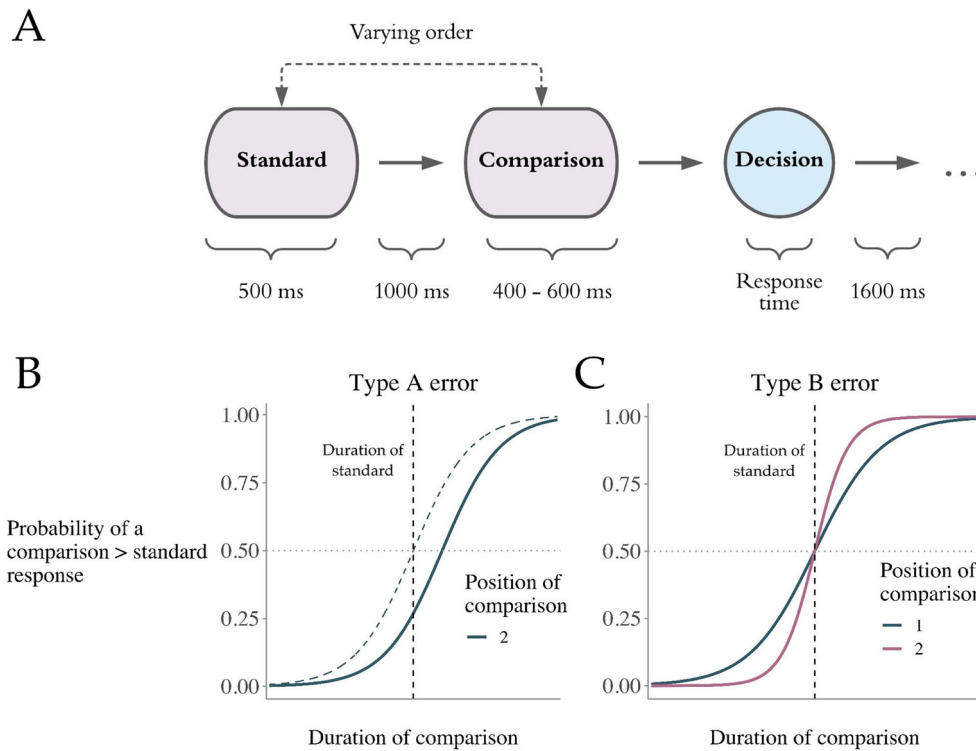


Fig. 1 **A.** Timeline of the experimental duration discrimination task used in Dyjas, Bausenhart, and Ulrich (2012). On each trial, a standard stimulus with a constant duration of 500 ms and a varying comparison stimulus with durations ranging from 400 ms and 600 ms were subsequently presented. The presentation order of the stimuli pseudo-randomly vary across trials. After the second stimulus was presented, the participants had to decide which of the stimuli had a longer duration. The response time was measured after the second stimulus was presented until a response key was pressed. **B.** Graph showing two theoretical psychometric functions that map the response probability to

different durations of the comparison stimulus. The point where these curves cross the horizontal dashed grey line indicates the point of subjective equality. When the duration of the comparison is not equal the duration of the standard at this point then we observe a Type A error, which is the case with the blue solid curve. **C.** Graph showing response probabilities as a function of the duration of the comparison for the two presentation orders separately. The two psychometric functions differ in their slope and thus indicate that the discrimination performance depends on the presentation order (Type B error)

more complex models have been proposed with the aim of accounting for such context effects. However, while these models often accurately describe the discrimination performance they are agnostic about the decision process itself. In the present study, we present a new approach that aims to incorporate variants of the state of the art model of perceptual discrimination into a decision process model. Although studies have investigated perceptual discrimination and the influence of contextual effects in a wide variety of stimulus features, the focus of the present study lies on the discrimination of short stimulus durations (below 1 s). In what follows, we briefly explain typical context effects in discrimination tasks and how these are accounted for, and then describe our modeling approach.

As noted above, decisions based on subjective stimulus intensity estimates can be biased by various contextual factors (so-called *carryover effects*). These effects can be broadly divided into *perceptual* and *decisional* context effects. Whereas the former refers to biases that occur based on the perception of previously encountered stimuli,

the latter describes contextual effects as a result of prior decisions. Every context effect can either be *assimilative* (i.e., the perception or decision is pulled towards previous ones) or *contrastive* (i.e., the perception or decision diverges from previous trials; Wiener et al., 2014). Further, these effects can broadly be classified into *global* or *local* effects. Global context effects describe the impact of the total set of past stimuli or decisions on a given trial. Conversely, *local* effects refer to the influence of immediately preceding trials (de Jong, Akyürek, & van Rijn, 2021). A typical example of a perceptual context effect is the central-tendency effect, also known as Vierordt’s law (Vierordt, 1868; Lejeune & Wearden, 2009). According to this law, humans tend to overestimate relative short durations and underestimate relative long durations (e.g., see Karin M. Bausenhart, Dyjas, & Ulrich, 2014; Grondin, 2005; Gu & Meck, 2011; Taatgen & van Rijn, 2011). Humans, thus, show a regression to the mean where the single stimulus is biased towards a representation of an average of previous stimuli (this reflects an assimilative global effect).

The so-called *Type A effect*, also known as the *time-order error* (TOE), refers to the impact of stimulus order on the *point of subjective equality* (PSE; Fechner, 1860). This means that participants usually over- or underestimate one stimulus relative to the other depending on the presentation order (for a comprehensive review of this research see, Hellström, 1985). Figure 1B shows a graph with two psychometric functions. The blue dashed sigmoid curve shows equal choice probability of the two responses (comparison > standard; comparison < standard) when the comparison and standard stimulus had the same duration. This means that there was no systematic over- or underestimation of either stimuli. The solid blue line has the same slope but is horizontally shifted to the right. This shows us that the duration of the standard stimulus frequently overestimated, which means there is a Type A effect.

In the classical *difference model* the Type A effect is accounted for by assuming a response bias parameter as an additive constant (Yeshurun, Carrasco, & Maloney, 2008; Alcalá-Quintana & García-Pérez, 2011). However, subsequent studies have invalidated this assumption at least to some extent (Hellström, 1977; Jamieson & Petrusic, 1976). These studies suggest that the cause of the TOE lies in perceptual – that is, pre-decisional – processes, which again cannot be explained by the *difference model*. However, possible decisional biases cannot be ruled out entirely. To date, the interplay of perceptual decisional processes in the origin of the TOE are not fully understood.

Dyjas et al. (2012) showed that when participants have to discriminate the duration between a constant standard stimulus s and a varying comparison c their discrimination performance is better when the standard stimulus precedes, rather than follows, the comparison stimulus. This effect is often referred to as a negative Type B effect² (Ulrich & Vorberg, 2009) and has also been shown in other domains such as *weight* (Ross & Gregory, 1964) or *contrast* discrimination (Nachmias, 2006). The Type B effect reflects a decreased slope of the sigmoid function for trials with the stimulus order [cs] compared to trials with reversed order (see, Fig. 1C). The Type A effect, however, reflects merely a lateral shift of the sigmoid function mapping response probabilities to the difference in stimulus duration. The Type B effect – albeit being observed across different modalities (e.g., visual, auditory) and stimulus attributes (e.g., duration, frequency, intensity, and numerosity) – received much less attention in research (Ellinghaus, Gick, Ulrich, & Bausenhart, 2018). Although most studies found a negative Type B effect (i.e., better discrimination when the comparison stimulus is presented after the standard),

some studies found a reversed effect (Hellström, Patching, & Rammsayer, 2020), especially, when stimulus duration and the inter-stimulus-interval (ISI) are very short (≤ 300 ms).

Both Type A and Type B effects are assumed to be *global* context effects, in the sense that their cause lies in the history of many previously encountered stimuli. Raviv, Ahissar, and Loewenstein (2012) demonstrated with an absolute stimulus duration identification task that immediately preceding trials are positively correlated. To investigate whether this effect is due to the perception of the previous stimulus or due to the previous decision, Wiener, Thompson, and Coslett (2014) conducted a study in which they counterbalanced the order of the different stimuli. They observed that decisions biased perception in the following trial, such that the interval was judged similarly. Further, they also found a *contrast* effect of stimulus perception on the subsequent perception. Further, an order effect on PSE (Type A effect) has also been observed (Dyjas, Bausenhart, & Ulrich, 2012; de Jong et al., 2021). Altogether, several *global* and *local* effects influence perception and decisions in discrimination tasks which cannot be explained by merely considering differences in stimulus magnitudes.

The predominant explanation for all these context effects is that decisions concerning the magnitude of a stimulus feature (e.g., duration) are derived not only based on the currently presented stimulus but also on the distribution of previously encountered stimuli. Thus, it is apparent that information stored in the memory system influences perception of and decisions about later presented stimuli. More recent modeling approaches aiming to improve the theoretical accounts for context effects on duration discrimination performance all share the theoretical rationale that responses to interval timing are based on a triad of cognitive processes: (1) A perceptive clock system that systematically changes over time. (2) A temporal reference memory system that stores past encounters with the stimulus. And, (3) a decision process that determines how the current output of the perceptive system relates to the values stored in the memory system and how to take any action based on this comparison (for a recent review see, van Rijn, 2016). Many different models have been proposed, which try to explain how the brain keeps track of time and implements such a clock system for time perception (for a review see, Balci & Simen, 2016).

Internal reference model

The reference memory system in particular is assumed to play an important role in the occurrence of context effects. Lapid, Ulrich, and Rammsayer (2008), for example, assume that participants store an internal reference of a prototype stimulus in the memory and update this reference over

²Rammsayer and Wittkowski (1990) called this effect the *position effect* and defined it the opposite way.

time (e.g., Durlach & Braida, 1969). Dyjas et al. (2012) proposed a quantitative model, the *internal reference model* (IRM), that describes how such an internal reference (I) is established and updated over time. According to this model, the internal reference I_n on a given trial n is computed as a weighted sum of the internal reference of the previous trial I_{n-1} and the internal representation $X_{1,n}$ of the first stimulus of the actual trial. This means that the internal reference is updated on a trial-by-trial basis, such that the internal reference I_n on trial n follows a geometrically moving average (Roberts, 1959):

$$I_n = g \cdot I_{n-1} + (1 - g) \cdot X_{1,n}, \quad (1)$$

with a weight g , $0 \leq g \leq 1$. This parameter indicates how much weight is given to the internal reference. To make a decision, participants compare this internal reference I_n with the internal representation of the second stimulus $X_{2,n}$, $D_n = I_n - X_{2,n}$. When this difference D_n is greater (smaller) than 0 then they decide that the first stimulus was longer (shorter). This model simplifies to the standard difference model if the weight g is set to 0.

Different studies showed that the IRM succeeds in predicting various context effects such as Vierort's law (Bausenhardt, Dyjas, & Ulrich, 2014), (Bausenhardt et al., 2014), Type A and B effect, as well as n-1 effects (Dyjas et al., 2012; Dyjas & Ulrich, 2014; Bausenhardt, Dyjas, & Ulrich, 2015; Ellinghaus, Ulrich, & Bausenhardt, 2018), comparison stimulus precedes a constant standard stimulus, the internal reference is no longer stable across trial because the variable stimulus gets integrated. This variation of the internal reference representation then causes a decreased discrimination performance. Thus, the size of the Type B effect, for example, should increase with increasing g because the percept is then influenced more strongly by the varying internal reference (Dyjas & Ulrich, 2014). A recent study by Ellinghaus et al. (2018) showed that this weight decreases when the interval between two stimuli increases. The idea behind this finding is that the internal representation decays over time. Dyjas et al. (2012) showed in two experiments that this model successfully accounts for the behavioral patterns (e.g., Type B effect) in a duration discrimination task where the stimulus order of a constant standard and a variable comparison stimulus was manipulated.

Sensation weighting model

The *sensation weighting model* (SWM) proposed by Hellström (Hellström, 1979; 1985) is a more general account. This model does not incorporate a trial-by-trial updating of an internal reference and it is formalized as follows:

$$D = [w_1 X_1 + (1 - w_1) R_1] - [w_2 X_2 + (1 - w_2) R_2] + b, \quad (2)$$

where D is the subjective difference between sensation magnitudes of two stimuli X_1 and X_2 each weighted by w_1 and w_2 . The parameter b reflects a bias. R_1 and R_2 are reference levels (similar to the internal reference) that indicate the average subjective level of stimulation of the stimuli. The crucial difference of the SWM (compared to the IRM) is that not only the first but also the second stimulus has a corresponding internal reference. This model simplifies to the same discrimination process as the IRM with $s_2 = 1$ and $b = 0$. Within this model context effects are explained by different weights for the stimuli. As shown by Hellström (1979) a larger weight for the second stimulus results in a Type B effect. Hellström, Patching, and Rammsayer (2020) argue that the SWM but not the IRM accounts for the full range of observed Type B and Type A effects. As soon as the standard stimulus is not fixed anymore (*roving standard tasks*), the IRM has problems accounting for the effects. Also, the sometimes observed positive Type B effects are difficult to explain with the IRM model (Hellström et al., 2020; de Jong et al., 2021). However, as Dyjas and Ulrich (2014) stated, it could be a promising approach to combine the generality of the SWM and the trial-by-trial updating mechanism of the IRM.

Both models discussed here ground on the notion of stimulus comparison, described by a linear model with different weights for the two stimuli and/or an integration of past stimulus experiences. Altogether, these models pose important progress compared to the standard difference model in accounting for various context effects. However, they focus on the memory system of the cognitive triad and are rather agnostic about the subsequent decision-making processes. The present study aims to provide a more detailed description of the processes involved in duration discrimination by incorporating the concepts of the IRM and the SWM into a *diffusion decision model* (DDM). In such a framework, choice and response time data are jointly analyzed. This allows for a more fine-grained analysis of the ongoing cognitive processes. It is well known that accuracy trades off with speed (e.g., Heitz, 2014). Accounting for response times can prevent potential inferential biases. Furthermore, using an additional source of information (response times) can act as a useful constraint for parameter estimation and can increase parameter recoverability (Shahar et al., 2019; Ballard & McClure, 2019).

Diffusion decision model

The DDM, originally developed by Roger Ratcliff (Ratcliff, 1978; for recent reviews see, Ratcliff, Smith, Brown, & McKoon, 2016; Voss, Nagler, & Lerche, 2013), belongs to the broader class of *evidence accumulation models*, sometimes also referred to as *sequential sampling models*.

Its core assumption is that noisy evidence is accumulated over time until a decision boundary (one for each decision alternative) is reached. This terminates the evidence sampling process, and the decision corresponding to the crossed boundary is made.

The standard DDM consists of four parameters, which all correspond to a specific cognitive aspect of the decision process: The drift rate v refers to the average rate of evidence accumulation. The boundary separation α is the distance between the boundaries and indicates how much evidence one considers necessary to reach a decision. Hence, this parameter is interpreted as a measure for response caution. The starting point z determines where the evidence integration process starts relative to the distance to the decision boundaries. If the starting point is equidistant from both decision boundaries, both decision alternatives have the same probability before the evidence accumulation process starts. When the starting point is shifted toward one of the boundaries, participants show an *a priori* bias toward one of the alternatives. Finally, the non-decision time τ , which is associated with the process of stimulus encoding and the execution of some action after one of the boundaries has been reached. The DDM is one of the most influential cognitive process models and is used in a wide variety of research domains involving two-alternative forced choice tasks (for a review see Wagenmakers, 2009). One of its advantages lies in the joint analysis of choices and the full response time distributions for correct and error responses. This allows the model, for example, to account for the prominent speed-accuracy trade-off (Luce, 1986; Heitz, 2014). Combining variants of the psychophysical memory models described above and the DDM into one framework could be beneficial for the field of stimulus discrimination and for the field of decision-making. On the one hand, it may be a fruitful extension for the models of perceptual discrimination, which are agnostic about the decision process itself and – on the other hand – it advances decision process models that usually ignore sequential and contextual effects in experiments. To account for the context effects (e.g., Type B effects) some of the DDM parameters must vary systematically from trial to trial. The DDM has already been extended by trial-by-trial variability parameters for the drift rate, starting point and also the non-decision time, which is often referred to the “full” Ratcliff diffusion model (Ratcliff & Tuerlinckx, 2002). It has been argued that these variabilities improve the quality of data fit, especially for fast responses (Lerche & Voss, 2016; Boehm et al., 2018). However, these trial-by-trial variabilities are usually assumed to be random. For the present model, our goal is to inform trial-by-trial variability, especially of the drift rate and the starting point, based on the proposed mechanics of the discrimination models described above.

To our knowledge, the study from Patching, Englund, and Hellström (2012) is the only attempt in this direction. In their study, the authors modeled data from a paired visual stimuli size and brightness discrimination experiment with a DDM that regressed the drift rate on differently weighted magnitudes of the stimuli:

$$v_n = w_1 X_{1,n} - w_2 X_{2,n} + b, \quad (3)$$

where $X_{1,n}$ and $X_{2,n}$ are the stimulus magnitudes on a given trial n with their respective weights w_1 and w_2 , and b is a constant. This corresponds to the mechanisms proposed by the SWM. Besides, they included also a random trial-by-trial variance for the drift rates, starting points and non-decision times. Their model succeeded in predicting the Type A effect. In the present study, we build on this study and apply different models with a similar rationale to empirical data from a duration discrimination experiment. First, we use these models to predict not only possible Type A effects but also the Type B effects. Second, we integrate different variants of the IRM and the SWM into a DDM framework and compare their fits to the data.

In discrimination task experiments, such as described earlier, response times (RT) are usually measured from the offset of the second stimulus. This means that the evidence accumulation process of the DDM technically starts when the presentation of the second stimulus ends. Balci and Simen (2014) proposed a two-stage sequential diffusion model with a similar idea in mind. The first stage is a diffusion process that delivers a noisy estimate of a time interval, which arises from a balance between excitation and inhibition and is referred to as a *time-adaptive, opponent Poisson drift diffusion model* (TOPDDM). The starting point as well as the drift rate of the second diffusion process, which corresponds to the actual decision process required in the experiment, is then influenced by the first stage’s first passage time. Within the TOPDDM framework, different intervals are timed by adjusting the accumulation rate; a higher drift rate is used to time shorter intervals. It is worth mentioning that this sequential DDM was applied to experimental data from a bisection task. On each trial in this task, a single stimulus had to be classified in one of two categories. Therefore, the TOPDDM formalizes the perception and decision involving a single stimulus and cannot account for comparative decisions of two stimuli. Also, the model does not incorporate any memory system mechanism that could account for context effects. Still, the study showed that the perception of stimulus can influence the starting point of the subsequent evidence accumulation process.

We assume that information from the first stimulus additionally affects the starting point of the evidence

accumulation. To evaluate whether our model assumptions are plausible, we fitted different models to the data from Dyjas et al. (2012). The relative fit to the data for all models was assessed with an approximate leave-one-out cross-validation procedure (Vehtari, Gelman, and Gabry, 2017). In addition to the relative goodness-of-fit it is important to also assess the absolute fit to the data (i.e., the degree to which they can capture quantitative and qualitative patterns in the empirical data), because even the relative best-fitting model could be a bad model for describing the data generating process (Palminteri, Wyart, & Koechlin, 2017). Therefore, we performed posterior predictive checks with the relative best-fitting model for response and response time data.

Methods

This study is based on a re-analysis of the data from Dyjas et al. (2012). More details about the methods can be found in the original article.

Participants

26 volunteers with normal hearing and sight participated in 3 sessions on different days in the first experiment. Data from 5 participants were eliminated from all analyses due to non-cooperative participation (see the *contaminants handling* subsection for more information about our elimination procedure). This resulted in a final sample $N = 21$ participants (15 female; 6 male) with an average age of 24.85 years ($SD = 7.3$, range = 18–41). For the second experiment, a novel sample consisting of 24 female participants was recruited. We excluded 3 individuals due to non-cooperative participation which resulted in a final sample of 21 participants with an average age of 20.19 years ($SD = 2.6$, range = 18–28).

Experimental task

In Experiment 1, participants had to decide which of two subsequently presented auditory stimuli (white noise) had a longer duration. On each trial, a stimulus (s) had a standard duration of 500 ms, while for the comparison stimulus c durations ranging from 400 to 600 ms were used. The inter-stimulus interval was always 1000 ms. Participants had to decide whether the first or the second stimulus had a longer duration. Response times were recorded starting from the offset of the second stimulus until a response has been made. After an inter-trial interval of 1600 ms, the next trial began. The experiment consisted of three conditions that differed in the order of the two stimuli and were tested in separate sessions. In the [sc] *blocked* condition the standard stimulus s always preceded the comparison stimulus c .

In the [cs] *blocked* condition, this order was reversed. In the *random* condition, both stimulus orders were presented randomly intermixed.

In Experiment 2, the task was the same except that visual (*discs*) instead of auditory stimuli were used and the range of durations of the comparison stimulus was increased to 300–700 ms. In the original study (Dyjas et al., 2012), no substantial differences have been found between the fixed and random conditions. For brevity, we focus our data analysis on the *random* condition of both experiments.

Contaminants handling

Generally, it is important to have an appropriate strategy for handling data points that are not a product of the process in consideration but from another process that is not in the focus of the research question (*contaminants*; Zeigenfuse & Lee, 2010). *Fast guesses* are one type of such contaminants, which are very fast responses (e.g., ≤ 300 ms) with chance level performance. In the case of diffusion modeling, it is particularly important to appropriately deal with this type of contaminants because otherwise, it can lead to biased parameter estimation and incorrect standard errors (Ratcliff, 1993; Ratcliff & Tuerlinckx, 2002; Ulrich & Miller, 1994). Furthermore, an analysis of contaminants helps to detect non-cooperative participants who can then be excluded from further data analysis.

Therefore, we applied a method called *exponentially weighted moving average* (EWMA; Chandra, 2007; Vandekerckhove & Tuerlinckx, 2001) to identify fast guesses for each individual separately. If the proportion of fast guesses exceeded 10% of all responses then this participant was excluded from further analyses. Some participants showed only a few very fast responses. In this case, it is not appropriate to calculate average accuracy because performance could exceed chance level randomly. Therefore, we additionally removed all responses faster than 100 ms.

Cognitive process models

With this study, we want to evaluate whether the DDM is an appropriate model for explaining decision processes involved in duration discrimination. Our core assumption is that – when two stimuli are presented sequentially – then the first stimulus influences the starting point of the evidence accumulation process, while the drift rate depends on the magnitude of both stimuli. We further enrich the diffusion model with a memory model that determines the influence of previously seen stimuli, as proposed by different variant of the IRM, SWM or a combination of both. We estimated several different models to test whether these assumptions are justified given our data (see Table 1 for an overview of all tested models).

Table 1 The different specifications of DDMs with the number of free individual-level parameters, their goodness of relative fit to the data quantified by the expected log-predictive density (elpd), and the corresponding uncertainty of these values quantified by the standard error (SE)

Model	Specification	N pars	Exp 1		Exp 2	
			elpd	SE	elpd	SE
1	$v_n = v_0 + v_1(X_{1,n} - X_{2,n})$	5	-6925	148	-4767	121
2	$v_n = v_0 + v_1(I_{1,n} - X_{2,n})$	6	-6709	148	-4636	121
3	$v_n = v_0 + v_1 X_{1,n} + v_2 X_{2,n}$	6	-6706	148	-4630	121
4	$v_n = v_0 + v_1 I_{1,n} + v_2 X_{2,n}$	7	-6681	148	-4609	121
5	$v_n = v_0 + v_1(I_{1,n} - X_{2,n})$	7	-6677	148	-4597	120
	$z_n = z_0 + z_1 I_{1,n}$					
6	$v_n = v_0 + v_1 X_{1,n} + v_2 X_{2,n}$	7	-6637	147	-4591	120
	$z_n = z_0 + z_1 X_{1,n}$					
7	$v_n = v_0 + v_1 I_{1,n} + v_2 X_{2,n}$	8	-6611	147	-4609	121
	$z_n = z_0 + z_1 I_{1,n}$					

v_n and z_n refer to the drift rate and starting point, respectively, on a given trial n . $X_{1,n}$ and $X_{2,n}$ denote the internal representation of the first and second stimulus on trial n . $I_{1,n}$ is the internal references computed by the mechanism suggested by the IRM (Equation 1)

Model 1 is a baseline model, which incorporates the simple difference model. Here, the drift rate is the only parameter that is allowed to vary between trials. This is modeled as a linear function of the difference between stimulus durations on the present trial. This model comprises a total of 5 free individual-level parameters: The intercept and slope of the drift rate, the starting point, the boundary separation, and the non-decision time.

Models 2-4 differ from the Model 1 in the linear function describing the drift rate. Model 2 implements the concept of the IRM by replacing the internal representation of the first stimulus X_1 with the internal reference I_1 , which is calculated for every trial following Equation 1. This introduces one additional parameter g that weights past internal references and the currently present first stimulus. In Model 3, the drift rate is calculated according to the SWM. Here, the drift rate v on a given trial n is modeled as a linear function of an intercept v_0 and different weights (β_1 , β_2) for the first and the second stimulus. Model 4 implements a combination of the IRM and SWM concepts, as suggested by Dyjas and Ulrich (2014). This model uses not only an internal reference for the first stimulus but also allows for different weighting of both presented stimuli.

To examine whether the processing of the first stimulus influences the starting point of the evidence accumulation process we refitted the Models 2 to 4 with an additional linear function for the starting point z (Models 5-7). For simplicity reasons, we did not include the random trial-by-trial variability parameters like Patching et al. (2012) did. We fitted all 7 models to the data of the *random* condition of Experiment 1 and 2.

Model fitting and evaluation

All models were implemented in a Bayesian hierarchical framework (Vandekerckhove, Tuerlinckx, and Lee, 2011). For each parameter, a Gaussian hyper-distribution was estimated from which individual parameters for each participant were sampled. This procedure allows to investigate inter-individual differences in the stimulus comparison process and also serves to account for a source of variability in the average parameter estimates (Lee, 2011). See Appendix A for a description of all (hyper-) priors used in our models.

All models were implemented in Stan (Carpenter et al., 2017; Stan Development Team, 2020b) and estimated with the R interface package RStan (Stan Development Team, 2020a). Samples were drawn using a Hamiltonian Monte Carlo sampler (HMC; Betancourt, 2018) with 4 chains and 2000 iterations of which 50% were used as warm-up samples and later discarded. To ensure model convergence we inspected the \hat{R} statistic (Gelman & Rubin, 1992) and assured that $\hat{R} < 1.01$ for all parameter estimates. We compared the relative fit of all models using *approximate leave-one-out cross-validation* (Loo R package; Vehtari et al., 2020; Vehtari et al., 2017). The best-fitting model, indicated by the highest expected log-predictive density (elpd), was then rigorously evaluated in terms of absolute fit by the means of posterior predictive checks. We sampled 500 parameters set from the posterior distribution. We then simulated new datasets with these parameters and calculated summary statistics for the responses (*mean* and *95% highest density interval*; HDI) as well as response

times (*median* and 95% HDI) and compared those with summary statistics of the empirical data. All code and data are freely available on GitHub (https://github.com/LuSchumacher/timing_discrimination).

Individual context effects

In order to evaluate the degree of Type B and Type A, effect we fitted a Bayesian logistic regression to the data for each participant separately. This model predicted the proportion of a $c > s$ response with the predictors *duration of c*, *stimulus order*, and their interaction. We then calculated the difference limen for both stimulus orders separately by the difference between the 75% and 25% percentile of the resulting psychometric function divided by 2. The individual Type B effect was then determined by the difference between those two *DL*'s. The individual Type A effect was calculated from the difference between the *PSE* of both stimulus orders, which corresponded to the 50% percentile of the psychometric function.

Results

Relative model fit

The differences in relative goodness-of-fit for all models are depicted in Fig. 2 for both experiments separately. All

models' elpd values were compared relative to the best-fitting model. Hence, the model with the most accurate out-of-sample prediction has an elpd difference of 0. We clearly see that Model 1, which is an implementation of the simple difference model, performs worse compared to all other models in both datasets. The elpd values for the Models 2 and 3 are superior compared to Model 1 and fairly similar to each other (see Table 1). This suggests that the different implementations of the IRM and SWM predict data equally well. Model 4 which used different weighting of the stimuli as well as the internal reference mechanism showed a slightly better fit in both experiments compared Model 2 and 3.

Models 5 to 7 included an additional linear function for the starting point of the evidence accumulation process. These models tend to show a slightly better goodness-of-fit. However, we did not observe a clear increase in prediction accuracy when compared to model variants that did not include this additional predictor (Models 2–4). Again, the results differ between Experiments 1 and 2. In Experiment 1, Model 7, which is a combination of the IRM and SWM, showed the best goodness-of-fit. It appeared to predict the data more accurately compared to the model that was identical but without the linear function of the starting point (Model 4). This was not the case in Experiment 2. Neither was Model 7 the best-fitting model nor did it differ from Model 4 in a meaningful way. Our data suggest a large degree of model mimicry and it appears that the impact of

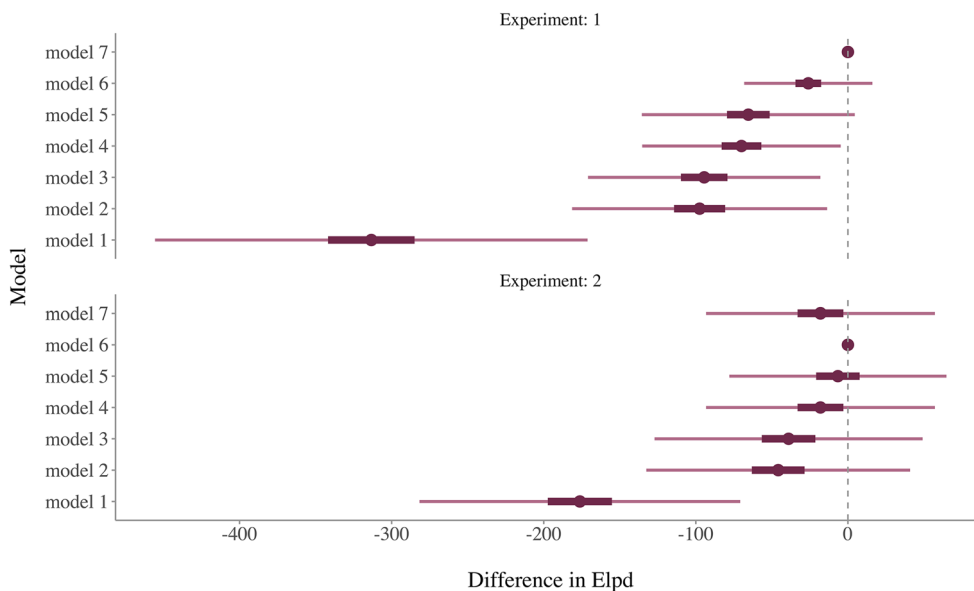


Fig. 2 Differences in the expected log-predictive density (elpd) for each model fitted to the data of Experiments 1 and 2. The difference for each model is computed relative to the best-fitting model, and thus, the elpd difference of the best-fitting model equals 0. The thick bars indicate ± 1 standard error and the thin bar ± 5 SE's of the elpd difference estimate

the first stimulus (or the internal reference) on the starting point is rather small.

Absolute model fit

Across both Experiments, Model 6 provided a good relative fit to the data. It predicted data best in Experiment 2 and was only slightly worse than the best-fitting model (Model 7) in Experiment 1 while using one less parameter. Thus, we evaluated this model in more detail in terms of absolute fit to data by performing posterior predictive checks. Figure 3AB

depict the probability of deciding that the comparison stimulus *c* was longer than the standard stimulus *s* for all durations of *c* and both stimulus orders separately. The solid lines and points indicate the empirical data averaged across all participants. The shaded areas indicate the 95% HDI of the mode from 500 simulated datasets and thus, describe the model’s prediction of the average performance and its uncertainty. As expected, the probability of choosing *c* > *s* increases with increasing duration of *c* in both experiments. The slope of the line is steeper when the comparison followed rather than preceded the standard stimulus. This

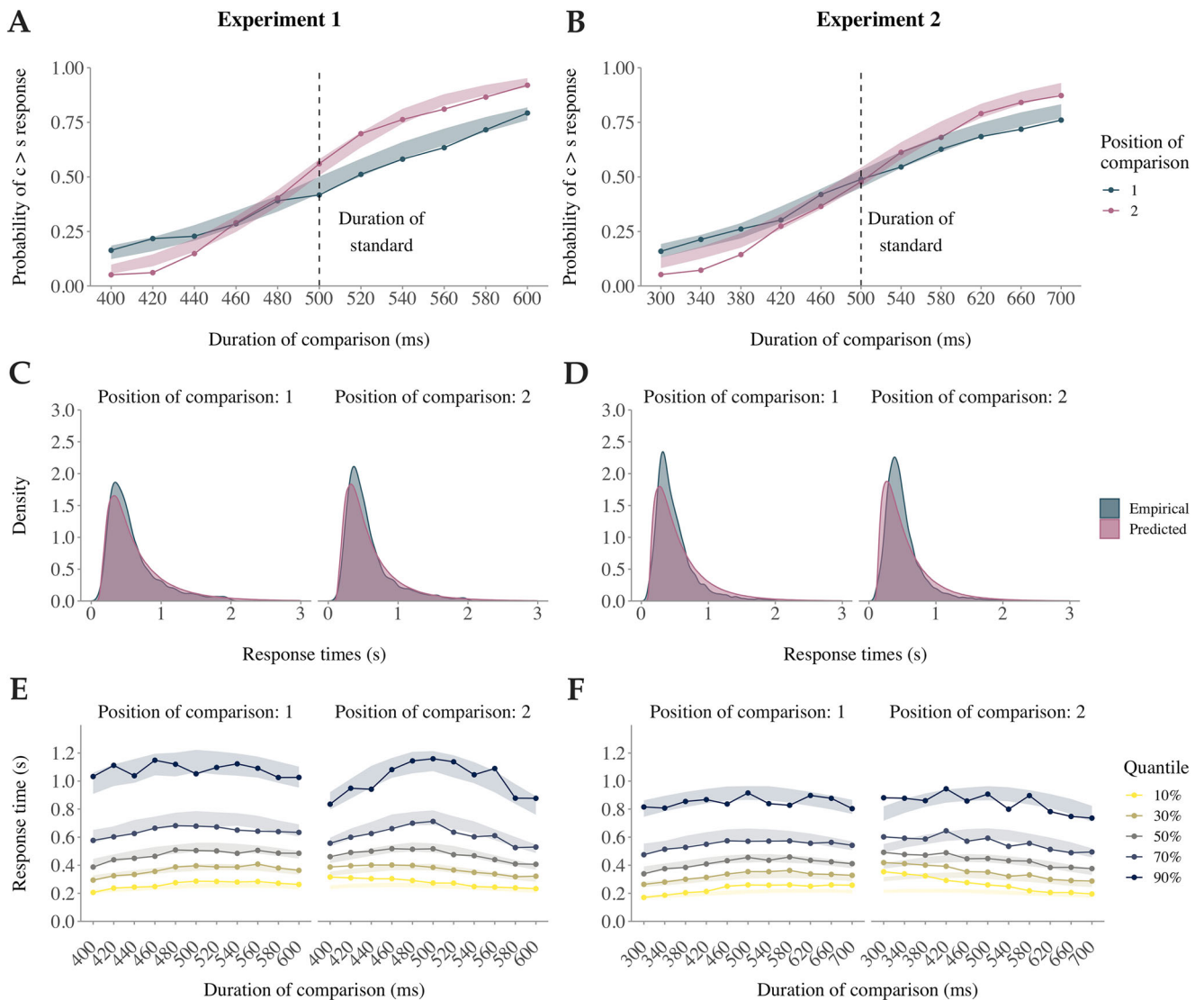


Fig. 3 Posterior predictions and empirical data of the average performance in Experiment 1 and 2. **A, B.** The average probability of a *c* > *s* response as a function of the duration of the comparison stimulus for both stimulus orders separately. The posterior prediction is shown with shaded areas (95% HDI) and the empirical data with solid lines and points. **C, D.** The densities of the predicted and the empirical raw

response time distribution of all participants for both stimulus orders separately. **E, F.** Different quantiles of the predicted (shaded area) and empiric (solid lines and points) average response time as a function of the duration of the comparison stimulus for both stimulus orders separately

indicates a negative Type B effect. Both patterns were successfully predicted by the model. However, the model slightly overestimates the average probability of $c > s$ responses for short durations of c in both experiments. This is probably due to shrinkage, which is a result of our hierarchical modeling approach. When the same model was fitted with complete pooling these divergences disappeared. Appendix B shows the posterior predictive checks based on the individual-level parameters for each participant separately. In these analyses, accurate predictions for all subjects were observed.

The empirical and predicted raw RT distributions are shown in Fig. 3CD for both stimulus orders and experiments separately. Comparing the empirical and the predicted densities reveal acceptable overlap in both conditions and experiments. However, the model was not able to capture the empirical data with very high precision. The mode of the predicted distribution was lower compared to the empirical distribution. Also, the model predicted heavier tails than have been observed in the data. All misfits were slightly more pronounced in Experiment 2 compared to Experiment 1 (see the Discussion section for possible explanations).

Figure 3EF shows a more fine grained picture of the RTs by depicting different quantiles (10%; 30%; 50%; 70%; 90%) of the observed and posterior predictive RT distributions for both stimulus orders and durations of stimulus c separately. In Experiment 1, RTs tend to increase with increasing task difficulty. This pattern seems to be more extreme in the tails of the RT distribution when stimulus c followed rather than preceded the standard stimulus. Both patterns were successfully predicted by our model. However, the uncertainty (95% HDI) increased in the tails of the RT distribution. This is due to the lower number of trials with such high RTs and also due to the greater variance.

The prediction of the RTs of Experiment 2 was not as good as for Experiment 1. Empirical RTs tend to be faster with short-duration c stimuli compared to longer durations if the comparison stimulus was presented first. The opposite pattern was found for the reversed order. A similar pattern was also observed in Experiment 1 although less pronounced. This is not a pattern we commonly would expect. Usually, RTs tend to increase with difficulty. Here, the most difficult decisions have to be made when the standard and comparison stimulus are the same or very similar. However, depending on the order of the stimuli either trials with short or long durations showed the slowest RTs. This is a pattern, which our model could not predict.

Parameter specific analyses

Table 2 shows the mode and 95% HDI of the group-level mean parameter posterior distributions for both experiments separately. The posteriors of the boundary separation are similar between both experiments and show plausible values. The estimates for the non-decision time are again similar and lower than typically observed in cognitive experiments. This could be a result of the experimental paradigm. We discuss this issue in more detail in the *discussion* section. Remember, the starting point of this model was modeled as a linear function of the duration of the first stimulus. z_1 corresponds to the beta-weight for the predictor *first stimulus* and its posterior is very small in both experiments. This means that in trials where the first stimulus showed the most extreme duration (-100 ms or 100 ms when centered on the standard stimulus) the influence of the first stimulus on the starting point would still be small (e.g., $0.0007 \cdot 100 = 0.07$ units of change in the starting point). Although this effect is very small, it is not zero.

As pointed out by Dyjas and Ulrich (2014) and Hellström et al. (2020), the Type A and Type B effect are explained

Table 2 The mode and lower/upper boundaries of the 95% HDI of all group-level mean posterior distributions for both experiments separately

Parameter	Experiment 1			Experiment 2		
	mode	lower	upper	mode	lower	upper
a	1.4736	1.3577	1.5986	1.3921	1.3320	1.4597
ndt	0.1456	0.0986	0.1760	0.1226	0.0889	0.1468
z_0	0.4411	0.4250	0.4559	0.4192	0.3967	0.4402
z_1	0.0007	0.0005	0.0009	0.0002	0.0001	0.0003
v_0	0.0437	-0.0769	0.1671	0.2327	0.0959	0.3480
v_1	0.0096	0.0073	0.0120	0.0060	0.0047	0.0075
v_2	-0.0198	-0.0225	-0.0170	-0.0087	-0.0095	-0.0078

within the SWM as a result of the different weighting of both stimuli. Figure 4 shows the correlation between the individual proportional difference between the weighting parameters ($\frac{v_1 - v_2}{v_1 + v_2}$) and the empirical Type A and Type B effect in Experiment 1 (left panel) and Experiment 2 (right panel). All participants except one showed either a tendency or a clear negative Type B effect which was already reported in the original study (Dyjas et al., 2012). We observe a large correlation of $r = 0.82$, which suggests that individual estimates of the weighting parameters are meaningful predictors for the size of individual Type B effects. Although a smaller association between the weights and the empirical Type A effect was found, a similar conclusion can be drawn.

Discussion

Human performance in duration discrimination is influenced by several context effects. For instance, the order of two successively presented stimuli not only affects the

point of subjective equality (Type A effect) but also the discrimination sensitivity (Type B effect). Current models that account for such effects propose that an internal representation of the stimulus history interferes with the perception of the current stimulus. Although they describe how a decision variable evolves, they are agnostic to the dynamics of the decision process itself.

In this work, we presented a novel modeling approach for perceptual decision-making in duration discrimination. We demonstrate that integration of current models of stimulus discrimination (IRM, SWM) into a Bayesian hierarchical diffusion decision model offers good prediction of the average as well as individual discrimination performance by taking not only responses but also the entire response time distribution into account. Moreover, we demonstrated that the estimates of the model parameters are meaningful predictors for two intensively studied context effects, the Type B effect and the Type A effect.

In the field of perceptual stimulus comparison two different models have been proposed: the internal reference model, and the sensation weighting model. To this date,

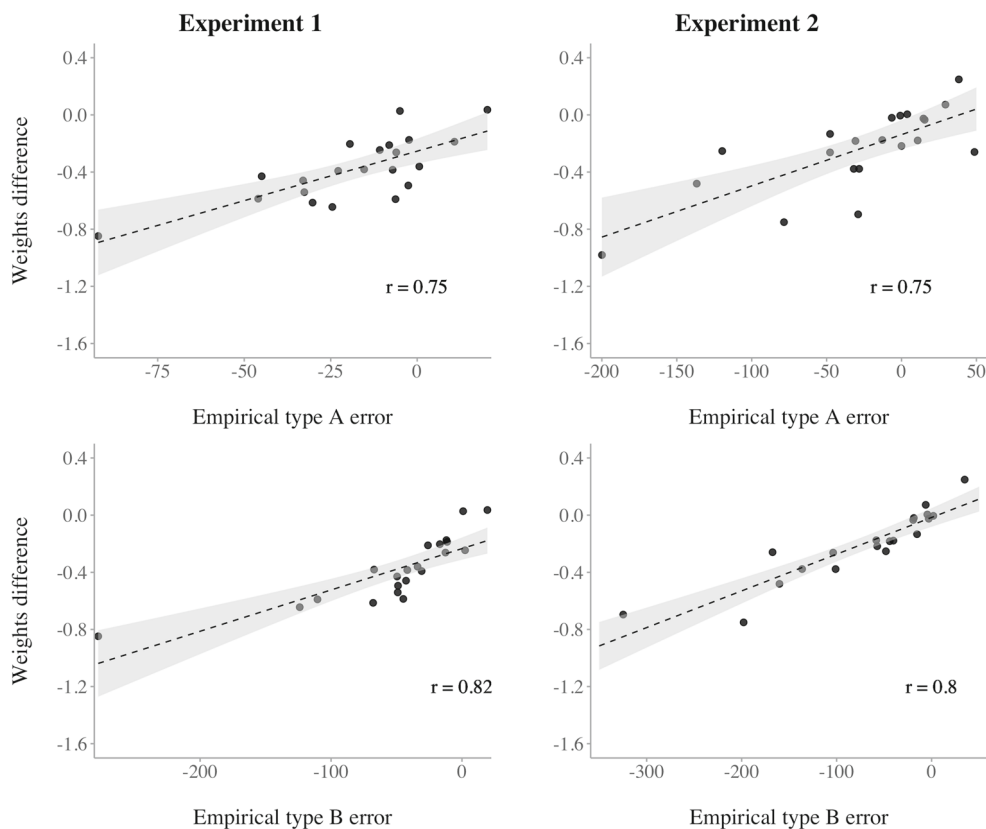


Fig. 4 Correlation between the proportional difference in weights and the empirically observed Type A (A) and Type B effect (B) of Experiment 2 (left panel) and Experiment 2 (right panel). Each point corresponds to a single participant's context effect and their individual-level proportional difference in weights

it remains unclear, which of those models best describes human discrimination processes. Little effort has been made, to rigorously compare the prediction of these models. The present study is a step towards this direction. However, we observed large model mimicry and were not able to discriminate between the different models based on the empirical data. This is not too surprising because the authors of the original study pointed out that it makes little difference whether the IRM or SWM is applied (Dyjas et al., 2012). We think, it would be a promising avenue for future investigation to compare the different models with our framework based on more complex task such as, for example, the roving standard task or task with very brief stimulus presentations. Furthermore, these models could also be tested against various local context effects (Wiener, Thompson, & Coslett, 2014; de Jong et al., 2021). Our approach could prove particularly useful since effects on response times are often found in such experiments (Wiener et al., 2014).

Our hierarchical modeling approach showed accurate prediction of the performance of most individual participants. Individual weight parameters w_1 and w_2 for the SWM were estimated. We showed that these parameters highly correlated with the individual extent of the Type A and Type B effect. It has already been pointed out that individuals can significantly differ in context effects (Dyjas, Bausenhart, & Ulrich, 2014). The ability to estimate individual parameters that correspond to these effects combined with individual latent variables provided by the DDM could be fruitful for future studies on individual differences in stimulus discrimination.

Future work also should address the whole range of sequential effects. Differentiation between global and local context effects and also the difference between perceptual and decisional carryover effects which could all be either assimilative or contrastive. Wiener et al. (2014) found that the perception of time is susceptible to similar adaptive and decisional effects as other categorical stimuli, where the responses for any given interval are simultaneously assimilated by the prior response and contrasted away from the prior interval. Urai, de Gee, Tsetsos, and Donner (2019), for example, showed that previous choices biased the average rate of evidence accumulation and not, as previously thought, the starting point of the accumulation process. It would be interesting to see whether this is also the case in duration discrimination tasks.

The models proposed in this study mathematically formalize two aspects of the cognitive triad involved in interval timing discrimination; an updating process of a memory system such as the internal reference in the IRM and a decision process formalized by the diffusion decision model as an evidence accumulation process. However, the model does not include a perception mechanism such as

a clock system implemented. Several promising models such as the TOPDDM (Balci & Simen, 2014) or the pace-maker accumulator model exist (Church, 1984; Gibbon, Church, & Meck, 1984; Hartcher-O'Brien, Brighthouse, & Levitan, 2016). Implementing such a mechanism into the modeling framework proposed here could lead to an even more complete formalization of all the processes involved in interval timing discrimination and could provide a promising framework for future studies. Moreover, recent advances in joint modeling enable us to incorporate neural data into diffusion models (Ghaderi-Kangavari, Rad, & Nunez, 2022; Turner, Forstmann, & Steyvers, 2019), which could be a fruitful approach to studying brain-behavior relationships during duration discrimination.

Further, Toso, Fassihi, Paz, Pulecchi, and Diamond (2021) showed that non-temporal stimulus features (e.g., loudness of a tone) can influence the perceived stimulus duration. In their duration comparison task, participants' responses were biased depending on the intensity of the stimulus which resulted in a horizontally shifted psychometric curve. They argue that this bias is a perceptual rather than a decisional phenomenon because it occurred whether the non-relevant feature was manipulated in the first or the second stimulus. In their opinion, the first stimulus is dissociated from any decisional process. The models proposed in the present study could precisely test this assumption as the starting point of the evidence accumulation process directly represents a decisional bias.

In perceptual decision-making tasks, RTs heavily depend on the difficulty of a given trial, expecting longer decision times for relatively difficult trials. In this study's experiments, the difficulty of a trial was relatively high when the duration of the comparison and the standard stimulus was very similar. The task difficulty decreases with larger duration differences because it gets more obvious which of the two stimuli' duration was longer. Thereby trials with relatively short durations of c and relatively long durations should be equally difficult. As the difficulty of a trial decreases, also the RTs should decrease and not differ between trials with short and long c durations.

Surprisingly, that is exactly what was observed in the empirical data. In both stimulus order conditions, RTs for relatively long durations of c differed from RTs when relative short durations were presented. When the c stimulus preceded the s stimulus participants responded slower when relative long durations were presented compared to relatively short durations. The opposite pattern was found when the c stimulus followed the s stimulus. In this case, participants tend to show slower RTs in trials when relative short durations of c were presented. Therefore, RTs were generally slower when the first stimulus was clearly longer than the second stimulus compared to trials when the first stimulus had a shorter duration than the second one.

This particular pattern in empirical RT data was most pronounced in Experiment 2 but also present in Experiment 1. Our model was not able to capture such RT differences. This is not surprising as the model assumes that the drift rate depends on the weighted difference between the duration of the stimuli presented in a trial. If this difference is low, the model produces more slow and erroneous responses. Conversely, when the weighted difference is relatively large, the model predicts more accurate and faster responses, independent of whether the duration of c was short or long.

It is an important open question if this pattern in the RTs can be explained as a phenomenon of the perceptual decision-making process in duration discrimination or whether it is a result of the experimental task used to measure the discrimination performance. In this study's experiments, the RTs were measured from the offset of the second stimulus until a response button was pressed. A possible explanation for this asymmetry in RTs could be that in trials where the duration of the first stimulus is shorter, participants can already know their choice around the time when the duration of the second stimulus exceeds the duration of the first stimulus. This would lead then to relatively fast responses as participants are already waiting for pressing the response button before they are allowed. Conversely, in trials where the first stimulus was longer than the second one, participants could be surprised by the abrupt ending of the second stimulus. This could then lead to delayed start-ups and thus explain the longer RTs in those trials.

Further, it is worth mentioning that some posteriors of the individual-level non-decision time parameters were odd. The mode of these distributions sometimes took on unrealistic small values of below 100 ms (range: 0.019 to 0.302). It is rather implausible that it took a participant only 19 ms to execute the motor action that was needed to give a response. This also contributes to the assumption that participants with such low non-decision time parameters at least sometimes already decided and prepared their response before the presentation of the second stimulus has finished.

Although we took good care of potential fast guesses by applying the EWMA method to each participants' responses, a substantial proportion of all responses were very fast. These fast responses, however, were clearly above chance performance. Although we observed such fast above chance performance responses, these did not occur exclusively in the [cs] order but also in the [sc] order. As we have not programmed nor experienced the stimuli in the experiment it is hard to come up with a satisfying explanation for these odd findings and simply disclose that the estimates of the non-decision time in our study have to be interpreted with caution.

We suggest that future studies take a deeper look into these surprising behavioral patterns. One way could be to start the RT measurement from the onset of the second stimulus. This possibly leads to even shorter RTs in trials where the first stimulus is clearly shorter than the second because participants would no longer have to wait to press the response button until the second stimulus presentation is finished.

In summary, the model proposed in this work provides a novel approach to predict human performance in duration discrimination. It not only incorporates perceptual mechanism like stimulus weighting or internal representation updating but also decisional processes such as processing speed or decision caution. We think the proposed model lay a good starting point to further investigate perceptual and decisional context effects.

Open Practices Statement

All data and code used in this study are freely available on GitHub (https://github.com/LuSchumacher/timing_discrimination). None of the experiments was preregistered.

Appendix A: Prior distributions

A list of the prior distributions used for all models. \mathcal{N} refers to a Gaussian normal distribution with a mean μ and standard deviation σ parameter. Γ denotes a gamma distribution with a shape κ and scale θ parameter. The beta distribution is written as \mathcal{Be} and uses two shape parameters α and β . All individual-level parameters were sampled from a group-level Gaussian normal distribution $\mathcal{N}(\mu, \sigma)$, with a mean μ (listed below) and a standard deviation $\sigma \sim \Gamma(1, 5)$. The group-level distributions for the boundary separation α and non-decision time τ parameter were truncated with a lower bound at 0. The group-level distribution for the g parameter, which governs the trial-by-trial updating mechanism of the internal reference was truncated with a lower bound at 0 and an upper bound at 1.

$$\begin{aligned}\mu_{v_0} &\sim \mathcal{N}(0, 5) \\ \mu_{v_1}, \mu_{v_2} &\sim \mathcal{N}(0, 1) \\ \mu_{z_0} &\sim \mathcal{Be}(5, 5) \\ \mu_{z_1} &\sim \mathcal{N}(0, 1) \\ \mu_g &\sim \mathcal{Be}(1, 1) \\ \mu_\tau &\sim \Gamma(3, 15) \\ \mu_\alpha &\sim \Gamma(2, 2)\end{aligned}$$

Appendix B: Individual-level posterior predictive checks

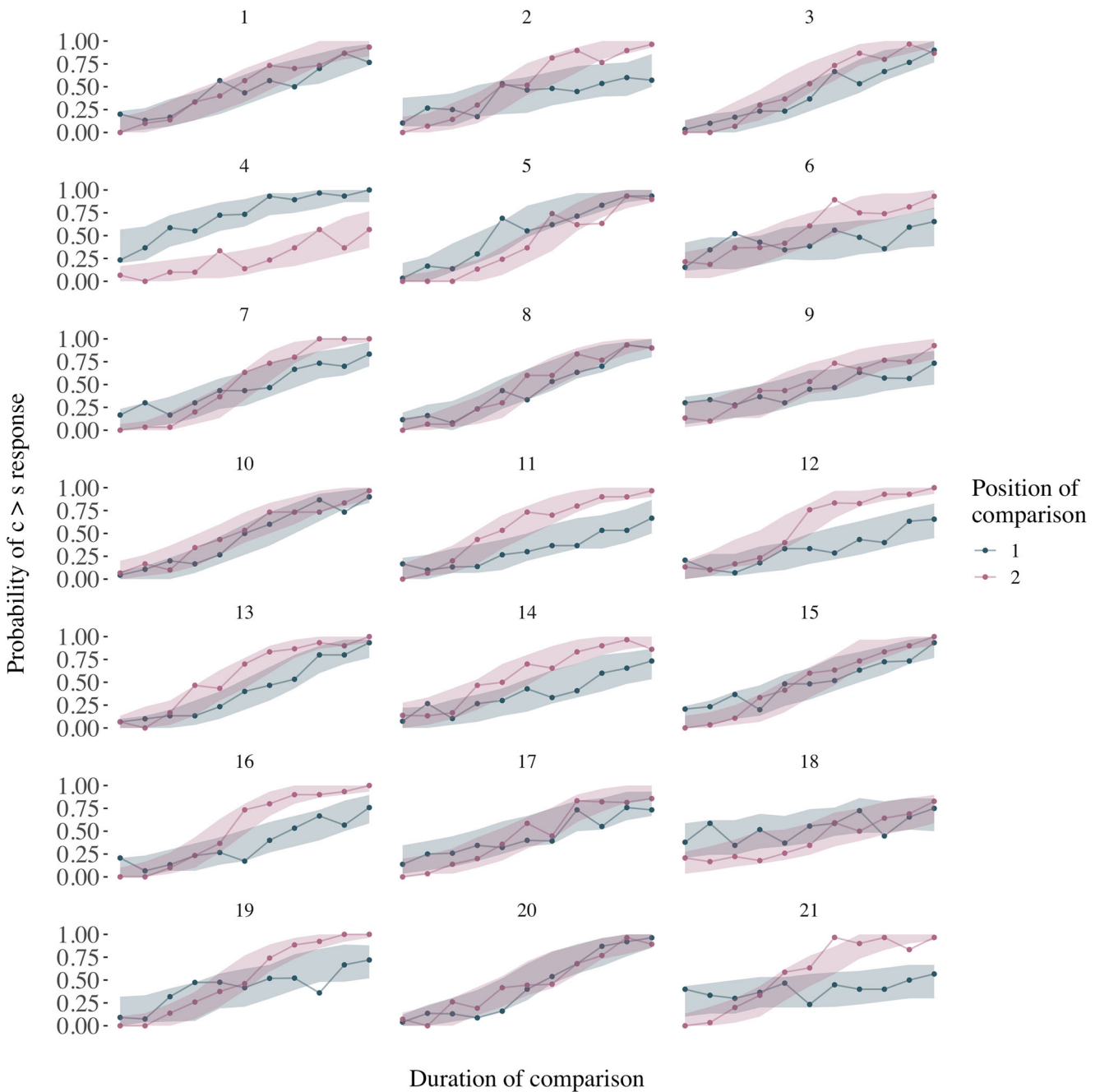


Fig. 5 Posterior predictions (shaded area) and empirical data of Experiment 1 (solid line and points) of the mean probability of a $c > s$ response for both positions of c and each participant separately. The

shaded area represents the 95% HDI over the mean of 500 simulated datasets based on individual-level parameter sets randomly sampled from the posterior distribution

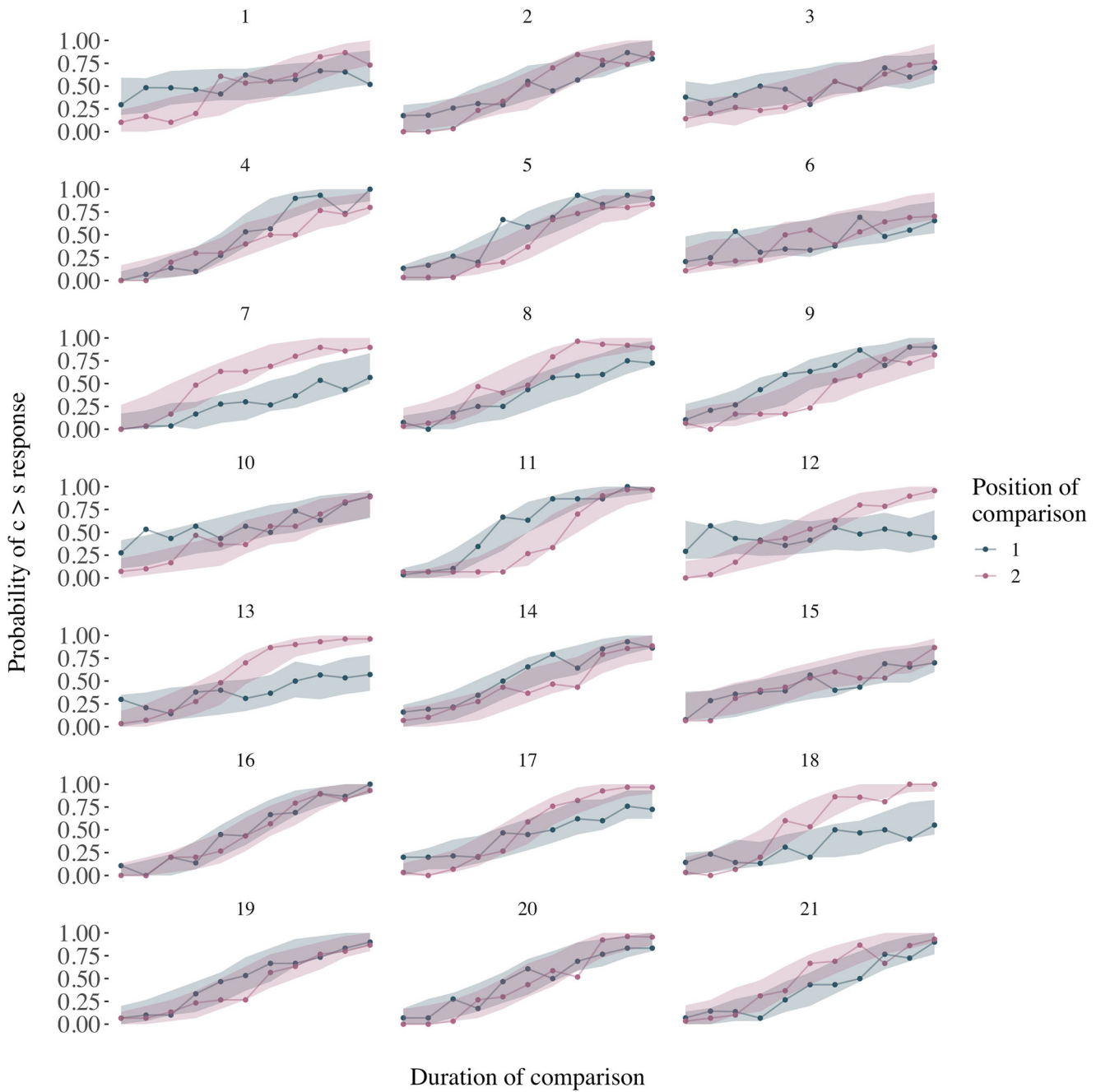


Fig. 6 Posterior predictions (shaded area) and empirical data of Experiment 2 (solid line and points) of the mean probability of a $c > s$ response for both positions of c and each participant separately. The

shaded area represents the 95% HDI over the mean of 500 simulated datasets based on individual-level parameter sets randomly sampled from the posterior distribution

Author Contributions LS and AV developed the concept of the study. LS performed the data analysis and prepared the draft manuscript. Both LS and AV approved the final version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; grant number GRK 2277 “Statistical Modeling in Psychology”).

Data Availability All data used in this work are freely available on GitHub (https://github.com/LuSchumacher/timing_discrimination).

Code Availability The code used for the analysis in this study is freely available on GitHub (https://github.com/LuSchumacher/timing_discrimination).

Declarations

Ethics approval and consent to participate This work uses datasets that were previously collected and published by Dyjas et al. (2012). The data collection was ethically approved by their local committee and was conform with the Declarations of Helsinki. The use of these data was permitted by the authors of the original article. Informed consent was obtained from all participants by the authors of the original article.

Consent for Publication Consent to publish the collected data was obtained from each participant by the authors of the original article.

Conflict of Interests The authors have no conflicts of interest to declare that are relevant to the content of this work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alcalá-Quintana, R., & Garcáa-Pérez, M. A. (2011). A model for the time-order error in contrast discrimination. *The Quarterly Journal of Experimental Psychology*, 64(6), 1221–1248. <https://doi.org/10.1080/17470218.2010.540018>
- Balci, F., & Simen, P. (2014). Decision processes in temporal discrimination. *Acta Psychologica*, 149, 157–168. <https://doi.org/10.1016/j.actpsy.2014.03.005>
- Balci, F., & Simen, P. (2016). A decision model of timing. *Current Opinion in Behavioral Sciences*, 8, 94–101. <https://doi.org/10.1016/j.cobeha.2016.02.002>
- Ballard, I. C., & McClure, S. M. (2019). Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. *Journal of Neuroscience Methods*, 317, 37–44. <https://doi.org/10.1016/j.jneumeth.2019.01.006>
- Bausenhardt, K. M., Dyjas, O., & Ulrich, R. (2014). Temporal reproductions are influenced by an internal reference: Explaining the vierordt effect. *Acta Psychologica, and Across Senses - Part-1*, 147, 60–67. <https://doi.org/10.1016/j.actpsy.2013.06.011>
- Bausenhardt, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics*, 77(4), 1033–1043. <https://doi.org/10.3758/s13414-015-0875-8>
- Bausenhardt, K. M., Bratzke, D., & Ulrich, R. (2016). Formation and representation of temporal reference information. *Current Opinion in Behavioral Sciences*, 8, 46–52. <https://doi.org/10.1016/j.cobeha.2016.01.007>
- Betancourt, M. (2018). A Conceptual introduction to Hamiltonian Monte Carlo. arXiv: 1701.02434 [stat].
- 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. <https://doi.org/10.1016/j.jmp.2018.09.004>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . , Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 1, 76.
- Chandra, M. J. (2001). *Statistical quality control*. Boca Raton: CRC Press.
- Church, R. M. (1984). Properties of the internal clock. *Annals of the New York Academy of Sciences*, 423(1), 566–582. <https://doi.org/10.1111/j.1749-6632.1984.tb23459.x>
- de Jong, J., Akyürek, E. G., & van Rijn, H. (2021). A common dynamic prior for time in duration discrimination. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-021-01887-z>
- Durlach, N. I., & Braida, L. D. (1969). Intensity perception: I. preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, 46, 372–383. <https://doi.org/10.1121/1.1911699>
- Dyjas, O., Bausenhardt, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception & Psychophysics*, 74(8), 1819–1841. <https://doi.org/10.3758/s13414-012-0362-4>
- Dyjas, O., Bausenhardt, K. M., & Ulrich, R. (2014). Effects of stimulus order on duration discrimination sensitivity are under attentional control. *Journal of Experimental Psychology: Human Perception and Performance*, 40(1), 292–307. <https://doi.org/10.1037/a0033611>
- Dyjas, O., & Ulrich, R. (2014). Effects of stimulus order on discrimination processes in comparative and equality judgements: Data and models. *Quarterly Journal of Experimental Psychology (2006)*, 67(6), 1121–1150. <https://doi.org/10.1080/17470218.2013.847968>
- Ellinghaus, R., Gick, M., Ulrich, R., & Bausenhardt, K. M. (2018). Decay of internal reference information in duration discrimination: Intertrial interval modulates the type B effect: Quarterly Journal of Experimental Psychology. <https://doi.org/10.1177/https://doi.org/10.1177/>
- Ellinghaus, R., Ulrich, R., & Bausenhardt, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance*, 44(1), 7–12. <https://doi.org/10.1037/xhp0000495>
- Fechner, G. T. (1860). *Elemente der psychophysik* Vol. 2. Leipzig: Breitkopf und Härtel.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Ghaderi-Kangavari, A., Rad, J. A., & Nunez, M. D. (2022). A general integrative neurocognitive modeling framework to

- jointly describe EEG and decision-making on single trials. <https://doi.org/10.31234/osf.io/pqv2c>
- Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences*, 423(1), 52–77. <https://doi.org/10.1111/j.1749-6632.1984.tb23417.x>
- Grondin, S. (2005). Overloading temporal memory. *Journal of Experimental Psychology: Human Perception and Performance*, 31(5), 869–879. <https://doi.org/10.1037/0096-1523.31.5.869>
- Gu, B.-M., & Meck, W. H. (2011). New perspectives on Vierordt's law: Memory-mixing in ordinal temporal comparison tasks. In Vatakis, A., Esposito, A., Giagkou, M., Cummins, F., & Papadelis, G. (Eds.) *Multidisciplinary Aspects of Time and Time Perception: COST TD0904 International Workshop, Athens, Greece, October 7–8, 2010, Revised Selected Papers, in Computer Science*, (pp. 67–78): Springer. https://doi.org/10.1007/978-3-642-21478-3_6
- Hartcher-O'Brien, J., Brighouse, C., & Levitan, C. A. (2016). A single mechanism account of duration and rate processing via the pacemaker-accumulator and beat frequency models. *Current Opinion in Behavioral Sciences*, 8, 268–275. <https://doi.org/10.1016/j.cobeha.2016.02.026>
- Hegelmaier, F. (1853). Ueber Das Gedächtniss Für Linear-Anschauungen. *Annalen der Physik*, 165(8), 610–620. <https://doi.org/10.1002/andp.18531650810>
- Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8. <https://doi.org/10.3389/fnins.2014.00150>
- Hellström, Å. (1977). Time errors are perceptual: An experimental investigation of duration and a quantitative successive-comparison model. *Psychological Research Psychologische Forschung*, 39(4), 345–388. <https://doi.org/10.1007/BF00308933>
- Hellström, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology. Human Perception and Performance*, 5(3), 460–477.
- Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, 97(1), 35–61. <https://doi.org/10.1037/0033-2909.97.1.35>
- Hellström, Å., Patching, G. R., & Rammsayer, T. H. (2020). Sensation weighting in duration discrimination: A univariate, multivariate, and varied-design study of presentation-order effects. *Attention, Perception, & Psychophysics*, 82(6), 3196–3220. <https://doi.org/10.3758/s13414-020-01999-z>
- Jamieson, D. G., & Petrusic, W. M. (1976). On a bias induced by the provision of feedback in psychophysical experiments. *Acta Psychologica*, 40(3), 199–206. [https://doi.org/10.1016/0001-6918\(76\)90011-1](https://doi.org/10.1016/0001-6918(76)90011-1)
- Lapid, E., Ulrich, R., & Rammsayer, T. (2008). On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task. *Perception & Psychophysics*, 70(2), 291–305. <https://doi.org/10.3758/PP.70.2.291>
- Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical bayesian models. *Journal of Mathematical Psychology*, 55(1), 1–7. <https://doi.org/10.1016/j.jmp.2010.08.013>
- Lejeune, H., & Wearden, J. H. (2009). Vierordt's the experimental study of the time sense (1868) and its legacy. *European Journal of Cognitive Psychology*, 21(6), 941–960. <https://doi.org/10.1080/09541440802453006>
- Lerche, V., & Voss, A. (2016). Model complexity in diffusion modeling: Benefits of making the model more parsimonious. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.01324>
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford: Oxford University Press.
- Nachmias, J. (2006). The role of virtual standards in visual discrimination. *Vision Research*, 46(15), 2456–2464. <https://doi.org/10.1016/j.vision.2006.05.016>
- Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6), 425–433. <https://doi.org/10.1016/j.tics.2017.03.011>
- Patching, G. R., Englund, M. P., & Hellström, Å. (2012). Time- and space-order effects in timed discrimination of brightness and size of paired visual stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, 38(4), 915–940. <https://doi.org/10.1037/a0027593>
- Rammsayer, T., & Wittkowski, K. M. (1990). Zeitfehler und positionseffekt des standardreizes bei der diskrimination kurzer zeitdauern. *Zeitfehler und Positionseffekt des Standardreizes bei der Diskrimination Kurzer Zeitdauern*, 142(2), 81–89.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59–108. <https://doi.org/10.1037/0033-295X.85.2.59>
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114(3), 510–532. <https://doi.org/10.1037/0033-2909.114.3.510>
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>
- Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9(3), 438–481. <https://doi.org/10.3758/BF03196302>
- Raviv, O., Ahissar, M., & Loewenstein, Y. (2012). How recent history affects perception: The normative approach and its heuristic approximation. *PLOS Computational Biology*, 8(10), e1002731. <https://doi.org/10.1371/journal.pcbi.1002731>
- Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3), 239–250. <https://doi.org/10.1080/00401706.1959.10489860>
- Ross, H. E., & Gregory, R. L. (1964). Is the weber fraction a function of physical or perceived input?. *The Quarterly Journal of Experimental Psychology*, 16(2), 116–122. <https://doi.org/10.1080/17470216408416356>
- Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., & Dolan, R. J. (2019). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLOS Computational Biology*, 15(2), e1006803. <https://doi.org/10.1371/journal.pcbi.1006803>
- Stan Development Team (2020). RStan: The R interface to Stan. R package version 2.21.2. Retrieved from <http://mc-stan.org/>
- Stan Development Team (2020). Stan modeling language users guide and reference manual. <https://mc-stan.org/>.
- Taatgen, N., & van Rijn, H. (2011). Traces of times past: Representations of temporal intervals in memory. *Memory & Cognition*, 39(8), 1546–1560. <https://doi.org/10.3758/s13421-011-0113-0>
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. <https://doi.org/10.1037/h0070288>
- Thurstone, L. L. (1927). Psychophysical analysis. *The American Journal of Psychology*, 38(3), 368–389. <https://doi.org/10.2307/1415006>
- Toso, A., Fassihi, A., Paz, L., Pulecchi, F., & Diamond, M. E. (2021). A sensory integration account for time perception. *PLOS Computational Biology*, 17(1), e1008668. <https://doi.org/10.1371/journal.pcbi.1008668>
- Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019). Joint models of neural and behavioral data. <https://doi.org/10.1007/978-3-030-03688-1>
- Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, 123(1), 34–80. <https://doi.org/10.1037/0096-3445.123.1.34>

- Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2afc tasks: Pitfalls and improved estimators. *Attention, Perception, & Psychophysics*, *71*(6), 1219–1227. <https://doi.org/10.3758/APP.71.6.1219>
- Urai, A. E., de Gee, J. W., Tsetsos, K., & Donner, T. H. (2019). Choice history biases subsequent evidence accumulation. *eLife*, *8*, e46331. <https://doi.org/10.7554/eLife.46331>
- van Rijn, H. (2016). Accounting for memory mechanisms in interval timing: A review. *Current Opinion in Behavioral Sciences*, *8*, 245–249. <https://doi.org/10.1016/j.cobeha.2016.02.016>
- Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026. <https://doi.org/10.3758/BF03193087>
- Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. <https://doi.org/10.1037/a0021765>
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Vierordt, K. (1868). Der Zeitsinn nach Versuchen. H. Laupp.
- Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology. *Experimental Psychology*, *60*(6), 385–402. <https://doi.org/10.1027/1618-3169/a000218>
- Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641–671. <https://doi.org/10.1080/09541440802205067>
- Wiener, M., Thompson, J. C., & Coslett, H. B. (2014). Continuous carryover of temporal context dissociates response bias from perceptual influence for duration. *PLoS One*, *9*(6), e100803. <https://doi.org/10.1371/journal.pone.0100803>
- Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*(17), 1837–1851. <https://doi.org/10.1016/j.visres.2008.05.008>
- Zeigenfuse, M. D., & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*(4), 352–362. <https://doi.org/10.1016/j.jmp.2010.04.001>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

APPENDIX B1 - MANUSCRIPT II

Manuscript II: Neural Superstatistics for Bayesian Estimation of Dynamic Cognitive Models



OPEN

Neural superstatistics for Bayesian estimation of dynamic cognitive models

Lukas Schumacher¹✉, Paul-Christian Bürkner², Andreas Voss¹, Ullrich Köthe³ & Stefan T. Radev⁴

Mathematical models of cognition are often memoryless and ignore potential fluctuations of their parameters. However, human cognition is inherently dynamic. Thus, we propose to augment mechanistic cognitive models with a temporal dimension and estimate the resulting dynamics from a superstatistics perspective. Such a model entails a hierarchy between a low-level observation model and a high-level transition model. The observation model describes the local behavior of a system, and the transition model specifies how the parameters of the observation model evolve over time. To overcome the estimation challenges resulting from the complexity of superstatistical models, we develop and validate a simulation-based deep learning method for Bayesian inference, which can recover both time-varying and time-invariant parameters. We first benchmark our method against two existing frameworks capable of estimating time-varying parameters. We then apply our method to fit a dynamic version of the diffusion decision model to long time series of human response times data. Our results show that the deep learning approach is very efficient in capturing the temporal dynamics of the model. Furthermore, we show that the erroneous assumption of static or homogeneous parameters will hide important temporal information.

Mathematical models are important tools for conceptualizing human cognition and predicting observable behavior. Such models aim to provide a mathematical formalization of cognitive processes by mapping latent cognitive constructs to model parameters and specifying how these generate manifest data¹. The surge of cognitive model applications has made it possible to test precise mechanistic hypotheses and to predict performance in various domains, such as decision-making^{2,3}, learning^{4,5}, or memory^{6,7}.

The majority of cognitive models treat human data as independent and identically distributed (IID) observations. The IID assumption implies that these models largely ignore the temporal changes of latent cognitive constructs. However, such constructs are inherently dynamic, regardless of a particular time scale^{8–11}. For instance, there is little dispute that constructs, such as working memory capacity¹² or mental speed¹³, change over the human life span. These constructs also vary on much shorter time scales, for example, within experimental sessions^{14,15}.

In psychological experiments, cognitive affordances are influenced not only by external task demands but also by internal mental processes and brain states that change over time. There are many possible explanations for the resulting systematic and unsystematic fluctuations, for instance, fatigue^{16,17}, practice^{18,19}, mind-wandering^{20,21}, or motivational factors^{22,23}. In this article, we argue that cognitive mechanisms should be treated as complex dynamic systems and that cognitive models should account for the dynamics of their components to fully understand and capture the rich structure of empirical human data.

Ignoring temporal fluctuations and changes in cognitive parameters can have drastic consequences for the descriptive, explanatory, and predictive merits of cognitive models. Consider a simple inverted U-shape hypothetical trajectory of a single parameter, as depicted in Fig. 1. Typical cognitive models assuming IID observations^{2,6} would estimate a flat trajectory (depicted in blue) whose uncertainty would match the width of the marginal parameter distribution (depicted in gray). Differently, dynamic models would account for temporal change and achieve a much greater information gain (depicted in red). Indeed, this is not just a hypothetical scenario, and we subsequently demonstrate its consequences in a real data application (cf. Fig. 8).

¹Institute of Psychology, Heidelberg University, Heidelberg, Germany. ²Department of Statistics, TU Dortmund University, Dortmund, Germany. ³Computer Vision and Learning Lab, Heidelberg University, Heidelberg, Germany. ⁴Cluster of Excellence STRUCTURES, Heidelberg University, Heidelberg, Germany. ✉email: lukas.schumacher@psychologie.uni-heidelberg.de

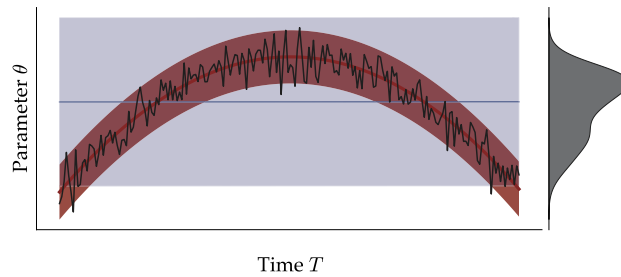


Figure 1. Conceptual illustration of a hypothetical parameter θ varying over time (solid black line). The solid blue line and shaded blue region depict the posterior mean and the 95% CI of a static model, respectively. The solid red line and shaded red region depict the posterior mean and 95% CI of a dynamic model, respectively. Treating the parameter as static (i.e., stationary) by marginalizing out the effects of time leads to inflated uncertainty estimates (matching the width of the marginal distribution, depicted in grey) and obscures the underlying change.

One way to mathematically formalize dynamic systems is by treating them as stochastic generative processes that produce data with temporal dependencies (i.e., time series data). As most complex systems are inherently non-linear, these time series often do not exhibit simple fluctuations around a stable mean with a fixed variance, but resemble a heterogeneous random walk²⁴. Beck and Cohen²⁵ coined the term *superstatistics*, which refers to a superposition of multiple stochastic processes on different temporal scales that can describe heterogeneous temporal dynamics²⁶. Thus, instead of assuming static model parameters, a superstatistics modeling approach introduces a hierarchy of at least two models: A low-level (i.e., observation or microscopic) model that formalizes the local behavior of a system and a high-level (i.e., transition or macroscopic) model that describes the parameter dynamics of the low-level model. Note that there is no absolute time scale for low- and high-level processes. The meaning of these terms is relative and always depends on the scale relevant to the research question.

A viable approach for modeling parameter transitions is offered by hidden Markov models (HMMs). For instance²⁷, accounted for different response states during a decision-making task by combining a HMM with an evidence accumulation model of decision-making. This model combination allows for discontinuous changes on longer time scales and continuous changes on shorter time scales. Similarly²⁸, extended a hierarchical version of the same decision-making model with a HMM and applied it to three existing long time series of response time and choice data. Both studies demonstrate that the HMM approach can reveal plausible fluctuations of decision model parameters in cognitive tasks.

However, the superstatistics framework is far more general and flexible in representing macroscopic fluctuations. First, it does not require modelers to pre-define a small set of possible modes (i.e., distinct system behaviors). Further, models within the superstatistics framework can be agnostic about the concrete dynamics of the model parameters—the most plausible dynamic can be directly estimated in a data-driven fashion. For example, using a superstatistics framework²⁹, demonstrated that the transition between different sleep stages is less abrupt than previously suggested.

The superstatistics framework has been utilized in physics^{30–32}, the life-sciences³³ and economics^{34,35}, but it has not yet been disseminated in the cognitive sciences. Under the assumption that cognitive processes are dynamic and complex, it seems natural to equip existing cognitive models with superstatistical aspects. However, to our knowledge, no previous study besides²⁹ has employed superstatistical methods for studying the dynamic aspects of cognitive parameters. Existing dynamic models of cognition fit stationary time series models (e.g., autoregressive models) to the observed behavior⁹ but do not incorporate a low-level mechanistic model that formalizes the underlying cognitive process(es). Thus, these time series models describe how behavior changes over time but do not explain how behavior occurs at a specific point in time. On the other hand, popular mechanistic models tailored to describe local behavior, such as diffusion decision models (DDM^{2,36,37}), either ignore the dynamic aspects of their parameters entirely or represent parameters as deterministic functions of time^{38–42}.

In this work, we argue that the superstatistics framework can reveal a more nuanced picture of cognitive dynamics and behavioral fluctuations. This is possible because we formalize the dynamic aspect of the low-level parameters as a higher-order stochastic process. Consequently, we estimate the low-level parameters at each time step directly from the data. Thus, their temporal evolution is only constrained by the modeler's choice of prior distributions and by the high-level transition model. Nevertheless, superstatistical models can be rigorously validated in the same way as their static counterparts, using standard model criticism methods, such as simulation-based calibration (SBC) to assess computational faithfulness, parameter recovery for inferential calibration, posterior re-simulation checks for assessing model adequacy, as well as cross-validation for assessing predictive performance^{43,44}. Superstatistical models allow us to address questions about how cognitive systems undergo distinct transitions in various settings²⁷. Further, one can examine which model parameters explain behavioral fluctuations without predefined equations that fix the hypothesized temporal evolution of specific parameters.

Superstatistical models can be quite challenging to estimate and compare for a number of reasons, especially in a Bayesian framework for principled uncertainty quantification²⁴. First, both the high-level and low-level models are stochastic, so there is considerable uncertainty about the values of all model parameters (i.e., static and dynamic) given a finite number of observations. Second, the low-level models might be complex and non-linear so that there is not always a closed-form analytic expression relating model parameters to data (i.e., the

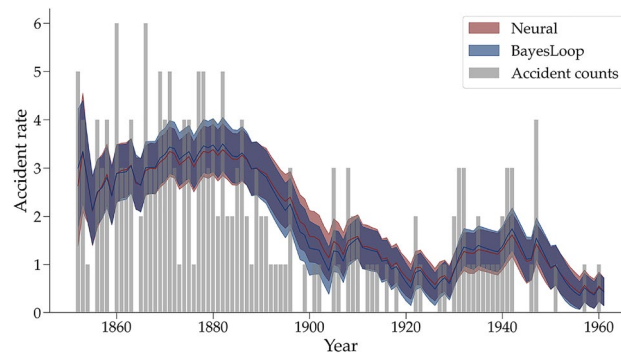


Figure 2. Coal mining disasters in the United Kingdom between 1852 and 1962. The annual reported accident counts are depicted using gray bars. The mean posterior of the rate parameter λ of a Poisson process with Gaussian fluctuation is shown with solid lines for both estimation methods separately. The shaded area represents ± 1 posterior standard deviation.

likelihood function is *intractable*), or the likelihood might be computationally very expensive to evaluate. Finally, even for stationary low-level models, the computational cost might become insurmountable when these models are applied to multiple data sets, since standard Bayesian methods are not amortized and thus need to be re-run sequentially (unless massively parallelized) and from scratch for each data set^{45,46}.

Indeed, estimation challenges may be the main reason for the underrepresentation of superstatistical models in psychology and the cognitive sciences. However, we argue that recent advances in (amortized) simulation-based inference (SBI^{45,47,48}) render estimation challenges secondary and allow researchers to create and test high-fidelity models of cognition without worrying about analytic tractability. SBI encompasses methods that use synthetic data to approximate intractable posterior distributions of unknown parameters. Moreover, amortized SBI with neural networks represents a particularly efficient way to perform posterior estimation on multiple data sets by investing the primary computational effort in a relatively expensive training phase^{47,48}. Once simulation-based training has converged, the trained networks can be applied to any number of observations or set of observations consistent with the model's structure.

The main purpose of this article is two-fold. First, we demonstrate and validate the use of superstatistics in cognitive modeling via an out-of-the-box extension of a popular mechanistic cognitive model, namely, the DDM. Second, we develop and validate a novel Bayesian estimation method grounded in the `BayesFlow` framework for amortized neural SBI⁴⁵. To this end, we first perform benchmark comparisons with existing frameworks on simulated data. We then specify a non-stationary DDM and fit it to long time series of response times obtained from human participants. Moreover, with this application, we empirically demonstrate how stationary models assuming IID observations can hide a number of interesting dynamic patterns and fluctuations present in behavioral data.

Results

Benchmark studies. To ensure the trustworthiness of our method, we first benchmark its performance against two existing Bayesian frameworks which use different estimation algorithms: `bayesloop`²⁴ and `Stan`⁴⁹. The former employs grid approximation for low-dimensional problems, whereas the latter relies on Hamiltonian Monte Carlo (HMC⁵⁰) sampling. Both frameworks operate in a non-amortized way and can only estimate superstatistical models with closed-form likelihoods.

Coal mining accidents. Currently, `bayesloop` cannot fit low-level models as complex as the DDM, nor high-level models such as the Gaussian process. Therefore, we compare the estimation performance of our method on a simpler example based on the coal mining accident data (freely available from²⁴). These data comprise counts of coal mining accidents in the United Kingdom between 1852 and 1962. The low-level model is a simple Poisson distribution with a parameter λ that corresponds to the accident rate. One can assume that the accident rate in coal mines was not constant during this more than a century-long period. Therefore, the accident rate λ is allowed to fluctuate over time according to the Gaussian random walk transition model (cf. Eq. 3). Both estimation methods use the same informative prior distribution for the low-level parameter $\lambda_0 \sim \text{Exp}(0.5)$ and high-level parameter $\sigma \sim \text{Beta}(1, 25)$.

Using the `bayesloop` software, we approximated a grid with 4000 equally spaced points ranging from 0 to 15 for λ and from 0 to 1 for σ , respectively. This calculation lasted approximately 38 minutes on a standard desktop computer. Training the neural network for 20 epochs took approx. 18 minutes, and obtaining 4000 posterior samples took less than a second. Thus, in this case, the training effort amortizes even after a *single* data set.

Figure 2 shows the annual count of coal mining accidents overlaid with the estimated dynamic accident rate λ (posterior mean and ± 1 standard deviation). Both methods estimate an almost identical latent trajectory for the low-level model parameter λ . Between the years 1880 and 1900, we observe a decrease in coal mining accidents followed by two temporary increases around the years 1905 and 1930. The estimated parameter dynamic

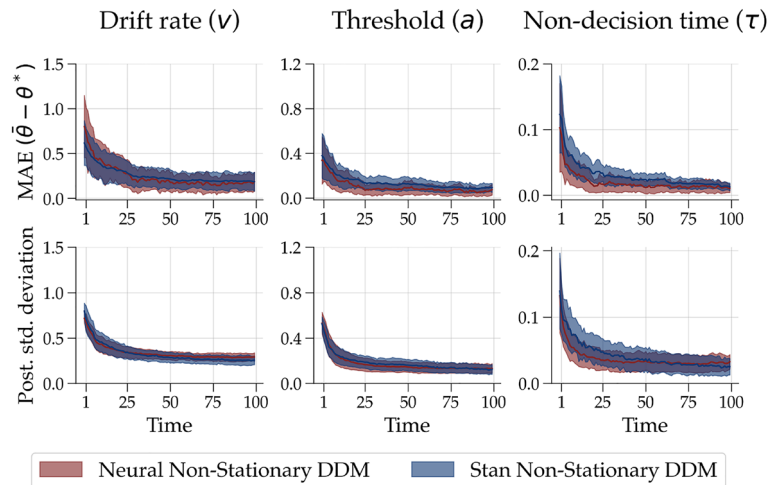


Figure 3. Comparison between the neural and Stan estimation method. First row: Median absolute error (MAE) between the ground truth data-generating parameters and the estimated posterior means across the 100 simulations over time. Second row: Posterior standard deviation aggregated across the 100 simulated data sets over time (solid lines). The shaded area depicts the median absolute deviation (MAD).

closely follows these data patterns. Thus, we conclude that our neural method can estimate a plausible parameter dynamic for a simple low-level model and performs equally well compared to `bayesloop`.

Static diffusion decision model. As a second benchmark, we compare our neural method to Stan in terms of Bayesian updating, assuming a “true” DDM with time-invariant parameters. This benchmark serves two goals. Firstly, it aims to compare the estimation performance of our method with that of Stan, which is regarded as the gold standard for sampling-based Bayesian inference. Secondly, it aims to assure that our method can correctly identify stationary parameters when fitting a dynamic model on data generated from a stationary process (i.e., it does not estimate “pseudo-dynamics”).

To this end, we simulated 100 data sets with 100 observations, each using a static DDM with 3 free parameters (see “A non-stationary diffusion decision model” section) without parameter fluctuation over time. Then, we fit a non-stationary DDM with a Gaussian random walk transition model (cf. Eq. 3) to all 100 data sets using both estimation methods. Again, we use the same prior distributions (see Appendix) to ensure comparability. We compared the two methods based on the following two performance metrics: (i) the median absolute error (MAE) between the estimated posterior means and the data generating stationary parameters averaged across all 100 simulations, and (ii) the average posterior standard deviation over time. These two metrics are common indicators for inferential model calibration, which aims to analyze the global behavior of the posterior distribution given possible observations from the prior predictive distribution⁵¹. The former metric informs us how well the posterior recovers the true model configurations (analogous to posterior z -scores). The latter metric indicates how much the posterior is informed by the data beyond the prior knowledge that was encoded in the prior distribution (analogous to posterior contraction)⁵¹.

The upper panel of Fig. 3 depicts the absolute difference between the true data generating parameters and the dynamically estimated posterior means over time, averaged over all 100 simulations for both methods separately. On average, the posterior means show a relatively large deviation from the true data generating parameters on early trials of the data. This difference then quickly decreases and flattens after approximately 25 trials. The performance of both methods concerning this metric is almost indistinguishable.

The lower panel of Fig. 3 displays the median posterior contraction measured as posterior standard deviation over time for all 3 parameters separately. We observe considerable posterior contraction within the first 25 time points. Again, the performance of both methods is nearly identical. However, there is a large difference in estimation time between the two methods. As we are interested in the filtering posterior distributions, the Stan model has to be refit with every additional observation of a time series. Hence, we fit the Stan model to each simulated $x_{1:t}$, $t = 1, \dots, T$, which amounted to $T = 100$ re-fits per simulated data set. Fitting the model to all 100 synthetic data sets resulted in 100×100 model fits. This procedure took over 1 week of non-stop computing on a standard desktop computer—whereas training the neural network lasted approximately 8 h with almost instantaneous fit to the 100 data sets thereafter. This is a non-negligible difference that will grow with longer time series, more data sets, or increased complexity until reaching a point where models can no longer be estimated with Stan due to limited processing resources or time constraints (see next section).

In summary, our method closely approximates the results obtained from `bayesloop` and Stan on the considered benchmark examples. Note, however, that our method is primarily designed for models where the above frameworks cannot be applied—higher dimensional models, possibly lacking a closed-form likelihood (i.e., available only as stochastic simulators), or many data sets consisting of long time series. The next application we present could be tackled with our neural approximators, but not with the above two frameworks.

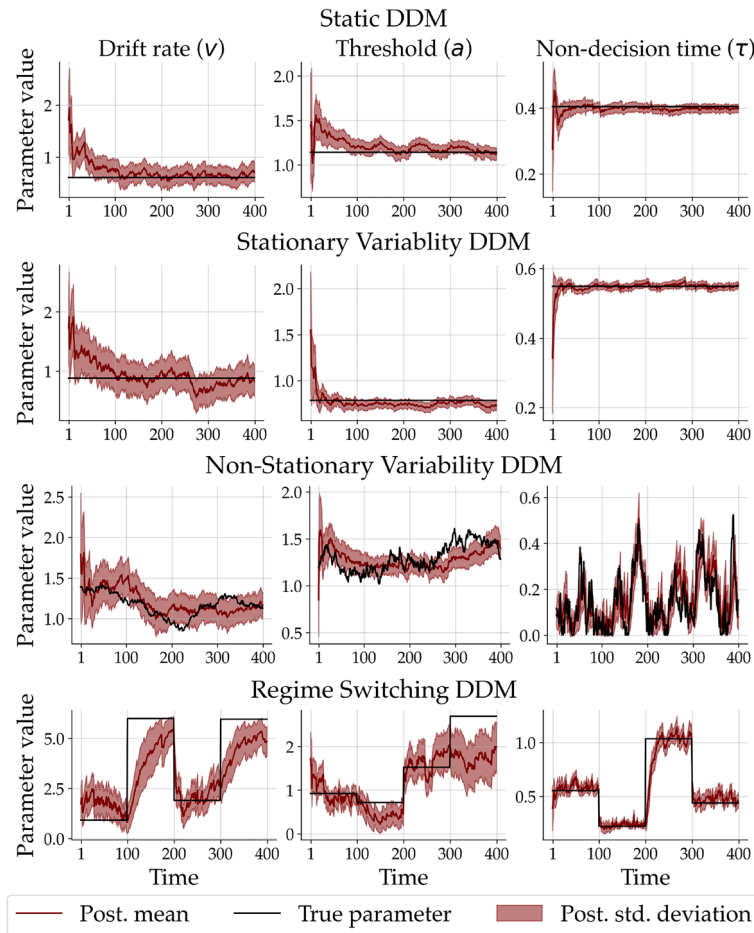


Figure 4. Example time-varying parameters estimated by our neural method in each scenario of the simulation study. Each row depicts the posterior estimates obtained from a single simulated person. The third row corresponds to the dynamic model used for training the network (i.e., well-specified case). The first, second, and fourth rows correspond to model variants not seen during training (i.e., misspecified cases).

Simulation study. Next, we probe the parameter recoverability of a non-stationary DDM under different induced misspecifications (i.e., models that differ from the one used for training the network). To this end, we performed an extensive study for which we simulated data sets consisting of $T = 400$ time points in four different scenarios: (i) A static DDM with constant parameters; (ii) a DDM with stationary variability (commonly referred to as “inter-trial variability”) where the 3 DDM parameters fluctuate randomly around a constant value; (iii) a non-stationary DDM with a Gaussian random walk transition model; (iv) and a DDM with constant parameters that jump abruptly and uniformly at three predefined time points (i.e., a regime switching model). Crucially, we trained the neural approximator only with simulations from the non-stationary model. However, during amortized inference, we applied the network to 200 data sets from each of the four scenarios. Thus, we could investigate the network’s response in the *open world* setting where the true data generator may differ from the reference model used during the training phase.

Figure 4 shows an exemplar fit of the non-stationary DDM with a random walk transition model to data sets from each of the four simulation scenarios. In the top row, we see that the estimated parameter trajectories converge to the constant ground-truth parameters. A similar pattern emerges when the ground-truth parameters randomly fluctuate around a constant value (second row), yet we observe less uncertainty reduction. The third row depicts the posterior estimates based on a data set simulated from the reference non-stationary DDM (i.e., the well-specified case). Besides some local deviations from the ground-truth parameter trajectory, the model is able to recover the overall trend of the dynamics. In the fourth row, we can inspect the posterior estimates from a data set simulated from the regime switching DDM which allows the parameters to “jump” uniformly at three time points to any value within the parameter bounds. Despite the severe misspecification, the random walk DDM is able to recover the discontinuous trajectories surprisingly well; still, the gradual adaptation and exhibits a notable lag after each switch.

Figure 5 depicts the true data generating and the estimated posterior means at time point $T = 99$ (right before the first jump of the regime switching transition model). We observe excellent recovery performance for all 3 parameters in all 4 simulation scenarios at the selected time point. The recovery performance at other time points as well as further details and analyses (i.e., MAE over time) can be found in the Appendix.

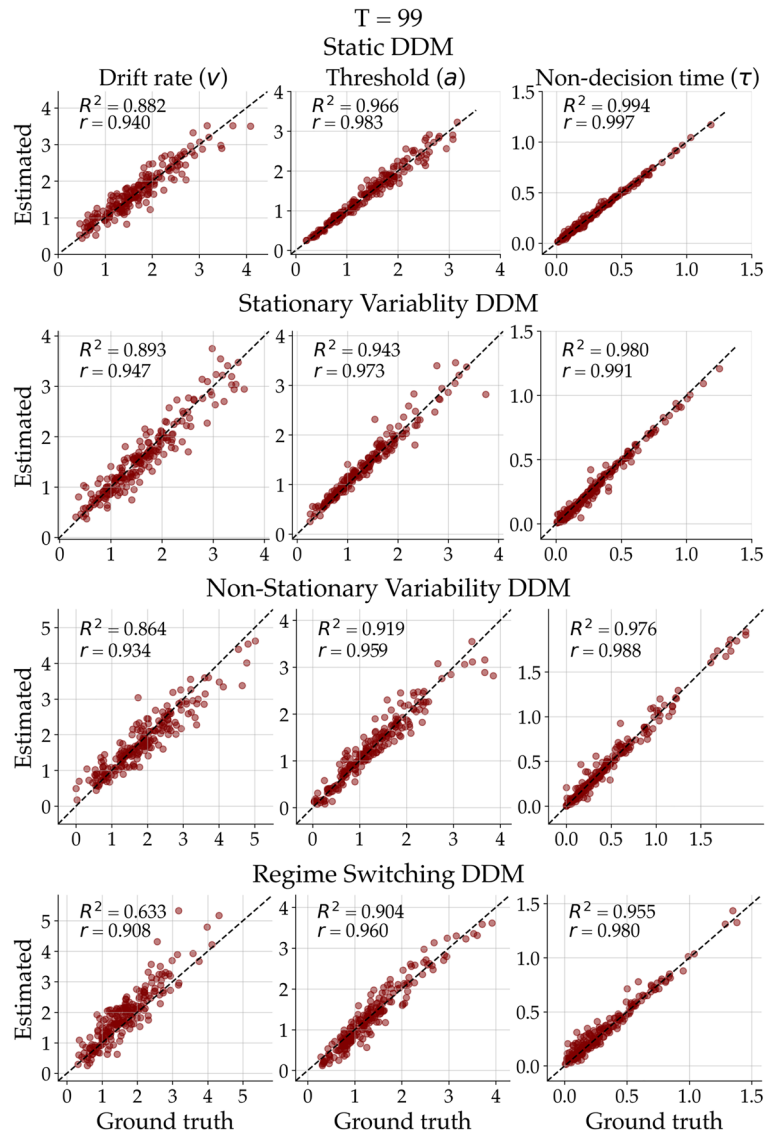


Figure 5. Ground truth–data generating parameters plotted against posterior means for all 3 parameters and simulation scenarios separately at time point $T = 99$ (just before the change of regime of the regime switching DDM).

Human data applications. Following our benchmarking and simulation studies, we applied non-stationary versions of the DDM to two separate data sets collected from response time (RT) experiments: (i) A standard random-dot motion task (a maximum of $T = 1320$ trials per participant), and (ii) very long time series (a maximum of $T = 3200$ trials per participant) from a lexical decision task. The first application serves as a starting point with data stemming from a popular task in experimental psychology. The second application showcases the utility of our method to estimate a complex non-stationary DDM with a Gaussian process (GP) transition model and multiple drift rate parameters for different difficulty conditions. Before fitting a model to empirical data, it is imperative to assess the faithfulness of the approximation method^{43,52}. To this end, we perform simulation-based calibration (SBC^{53,54}). These analyses suggest that our neural Bayesian method exhibits reasonable calibration, with slightly miscalibrated posteriors for the non-decision time parameter (see Appendix for more details on calibration).

Random-dot motion task. First, we fit a non-stationary DDM with a Gaussian random walk transition model to a data set retrieved from the experimental study of⁵⁵. We chose this data set because the purpose of the original study was to investigate the decline of the threshold parameter over time. The experiment had a 3 (*Low*, *Medium*, and *High* feedback) by 2 (*Time* and *Trial* condition) factorial between-subject design. Differently from our approach,⁵⁵ subdivided the time series into trial bins and fitted a stationary hierarchical Bayesian DDM to each bin separately. Therefore, we can compare the parameter trajectories recovered by our neural superstatistics method with the estimates obtained by the original authors using Markov chain Monte Carlo (MCMC).

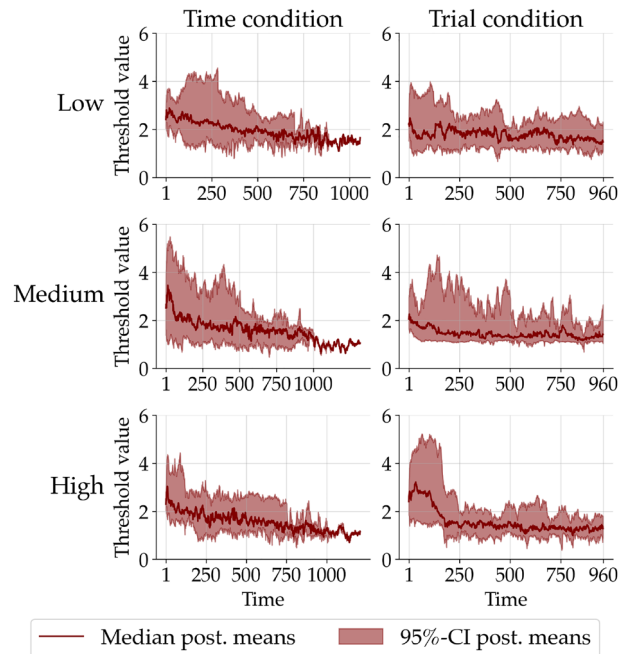


Figure 6. Estimated trajectories of the DDM threshold parameter aggregated across all individuals for each between-subject experimental condition. The first column corresponds to the *Time* and the second to the *Trial* condition. The rows correspond to the three feedback conditions, *Low*, *Medium*, and *High*, respectively. The red solid lines depict the median of the individual posterior means and the red shaded area the 95% credibility interval of these posterior means.

Figure 6 depicts the trajectory of the threshold parameter aggregated across all individuals in a separate panel for each experimental condition. Note, that in the *Time* condition participants had a fixed amount of time they could spend on the task resulting in different time intervals. When we compare our estimates to those obtained by⁵⁵, it becomes evident that both approaches yield similar qualitative and quantitative patterns. This result complements our promising results “in silico” and points to the convergent validity of our superstatistics approach in applications with real data.

Lexical decision task. We fit the non-stationary DDM with a GP transition model (cf. Eq. 5) to human behavioral data originating from a lexical decision-making task. The data consist of long RT and choice time series from four experimental conditions. For this application, we use four separate drift rates—one for each experimental condition. The length of these time series made it impossible to estimate the model with *Stan* (due to memory limitations and infeasible compute time). Thus, to increase the trustworthiness of the results obtained with our neural method, we resort to the established *fast-dm* software³⁶ as a benchmark, which is capable of estimating homogeneous (block) trial-by-trial fluctuations (i.e., inter-trial variabilities). We then compare the goodness of absolute fit in terms of re-simulation accuracy between both estimation methods and investigate the multi-horizon predictive performance of our method. Further, we analyze the main advantage of the non-stationary DDM, that is, the inferred trial-by-trial parameter dynamics, and compare those to the static *fast-dm* parameter estimates. Note that *fast-dm* is not a Bayesian method and is thus not included in our previous benchmark studies.

The left panel of Fig. 7 depicts the empirical RT time series data of an individual participant in black (Figures for the remaining participants are available in the Appendix). To evaluate whether the non-stationary DDM is capable of capturing the empirical data, we perform posterior re-simulations on the first 3 blocks of the experiment (trials 1–2500). To this end, we draw 100 samples from the posterior distributions over $\theta_{0:2499}$ to simulate 100 posterior re-simulated data sets. The resulting RT time series are then summarized with the median and the 95% credibility interval (CI) across simulations and depicted in red color. We smooth the trial-by-trial empirical data and model outputs via a simple moving average (SMA) with a period of 5 to ease visual inspection of potential trends. Note, that the re-simulation from the *fast-dm* model is only shown in the marginal RT distribution on the right panel to avoid visual clutter.

The overall time series show that the individual’s RTs decrease over time. Furthermore, the variability of the RTs, which is most pronounced in the first session, decreases considerably over time. The non-stationary DDM not only captures both of these overall trends, but also represents the shorter time oscillations within the empirical RT time series. The data also exhibits various sudden “jumps” in RTs, probably due to fluctuations in non-decisional processes, such as inattention. Unsurprisingly, these jumps are not fully accounted for by our non-stationary DDM since the high-level model (GP with squared exponential kernel) does not allow for sudden large changes in the low-level parameters.

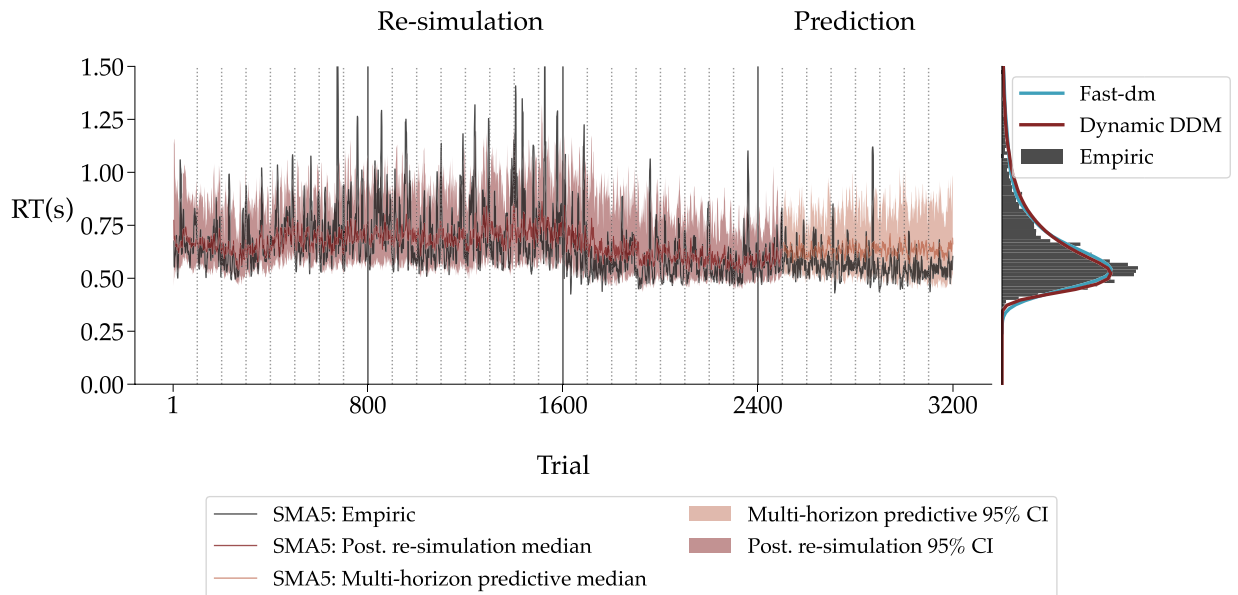


Figure 7. Model fit to human data. **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the non-stationary DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and multi-horizon prediction correspond to 95% credibility intervals. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the non-stationary DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

We purposefully leave out the remaining 700 trials from the posterior re-simulation analysis to also test the predictive capabilities of the non-stationary DDM against held-out empirical data^{56,57}. To this end, we generate 100 new parameter dynamics according to Eq. (5) with randomly drawn posterior samples of θ_{2499} as initial parameter values and posterior samples of the high-level Gaussian process parameters η . Then, we simulate 100 novel RT time series for the remaining 700 trials using the simulated parameter trajectories. The resulting RT time series are summarized in the same manner as before (median, 95% CI) and again smoothed with an SMA. The corresponding multi-horizon posterior predictions are depicted in Fig. 7 with an orange color. The dynamic model yields accurate predictions on the held-out data and thus does not overfit the training data. Moreover, the held-out time series remain in the 95% CI of the multi-horizon prediction, which is the case for all individual data sets (see Appendix).

The right panel of Fig. 7 depicts the empirical RT distributions (black) along with the data generated by the non-stationary DDM (red) and the static DDM (blue). Note that the three empirical RT distributions show a substantial overlap. Since the `fast-dm` re-simulations serve as a benchmark for the non-stationary DDM, it is essential to quantify if there are pronounced deviations between the re-simulated and the empirical RT distributions. To this end, we estimate the pairwise maximum mean discrepancy (MMD) between the three distributions for each individual separately and then average the resulting values across participants. MMD is a kernel-based statistical metric of equality between distributions⁵⁸.

Accordingly, our analysis reveals no pronounced differences between the three distributions. The average MMD between the empirical RT distributions and the ones predicted by the non-stationary DDM ($\overline{MMD} = 0.026$, $SD = 0.008$) is lower than between the empirical and the ones predicted by the `fast-dm` model ($\overline{MMD} = 0.042$, $SD = 0.027$). The SDs of the average MMD values indicate that data generated with the non-stationary DDM are not only slightly more accurate on average but also more consistent compared to data generated from the standard DDM. For the sake of completeness, we also compare both re-simulated RT distributions ($\overline{MMD} = 0.035$, $SD = 0.019$). This comparison reveals that the re-simulated RT distributions of the static DDM are more similar to the one obtained by the non-stationary DDM than to the empirical RT distribution. Altogether, both models can reproduce the empirical RT distributions with high fidelity, but the non-stationary DDM fits the data slightly better than the static DDM estimated with `fast-dm`.

In summary, our non-stationary DDM can closely re-simulate and predict the temporal trajectory of empirical RT time series as well as corresponding raw RT distributions from all individuals (see Appendix). Even though the standard DDM also accounts for the marginal RT distribution, it cannot generate the observed heterogeneous RT time series data (cf. Fig. 7).

However, the most decisive advantage of our non-stationary DDM over its stationary counterpart is that it can recover parameter dynamics directly from the empirical data. As the static parameters of `fast-dm` can only vary homogeneously around their mean, we cannot detect any systematic changes in the parameters over time. However, the dynamic parameters estimated with our neural method strongly suggest such systematic

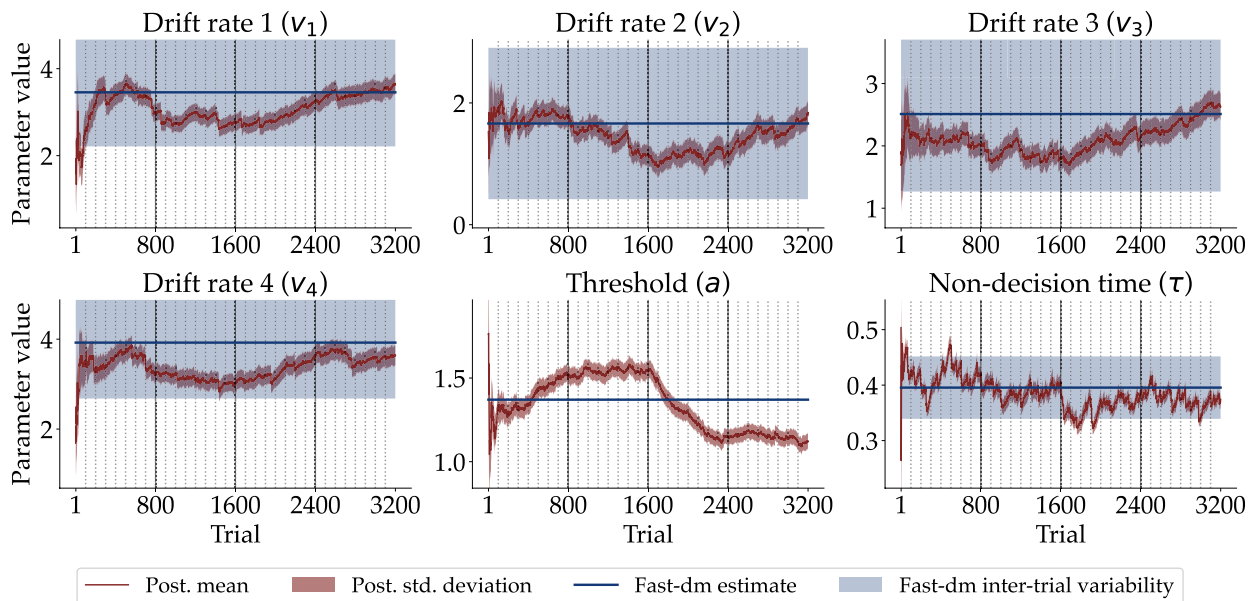


Figure 8. Estimated parameter dynamics. The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates v_1 – v_4 (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the stationary DDM parameters and the corresponding inter-trial variabilities (except for the threshold a) are shown in solid blue lines and blue shaded areas, respectively.

changes. Figure 8 depicts the dynamics of the estimated trial-by-trial posterior means and ± 1 standard deviation for all DDM parameters of the same participant as above in red (see Appendix for the parameter dynamics of the remaining participants as well as the average parameter dynamic). The corresponding point estimates (solid line) and inter-trial variabilities (shaded area) obtained with `fast-dm` are shown in blue.

All parameters of the non-stationary DDM seem to exhibit considerable fluctuations and notable oscillations throughout the experiment. Due to the assumption of homogeneous variation, the inter-trial variabilities inferred with `fast-dm` vastly overestimate the uncertainty in parameter estimates (cf. Fig. 8). The dynamic drift rates fluctuate roughly within the uncertainty corridors spanned by the homogeneous inter-trial variabilities, but exhibit much tighter error bars. As a consequence, local drift rates are much less uncertain than the homogeneous variability parameters indicate. On the other hand, the dynamic non-decision time τ fluctuates more than the corresponding flat inter-trial variability. Note that `fast-dm` does not support estimating inter-trial variability of the threshold a , so we only report the estimates of our neural method, suggesting a substantial decrease of the threshold parameter throughout the experiment. Notably, we observe a considerable mismatch between heterogeneous and homogeneous dynamics in almost all individuals (see Appendix).

Discussion

In this work, we explored the merits of superstatistics for representing non-stationary dynamics in cognitive processes, along with the utility of a neural Bayesian method for estimating superstatistical models. We verified the computational faithfulness and adequacy of our method using simulations and two benchmark studies. We then applied our method to a dynamic, non-stationary diffusion decision model and estimated the temporal trajectories of its key parameters, namely, drift rates, decision threshold, and non-decision time from the data of two experiments. We showed that such a non-stationary model (i) can indeed be fit to long time series of human data with high fidelity and (ii) that the inferred heterogeneous dynamics reveal patterns that would have remained hidden by traditional stationary models^{2,6}. To our knowledge, this is the first attempt to augment a stationary cognitive model by employing a superstatistics framework.

Previous research has suggested that response times often exhibit heterogeneous dynamics^{9,10}. It has also been shown that even the history of past choices can influence specific parameters of the DDM⁴⁰. Hence, it seems plausible that the cognitive processes represented by the DDM parameters vary over time even within an experimental session due to internal psychological factors. This is exactly what was implied by the individual parameter dynamics inferred from the lexical decision task data set. However, as the data originates from an experiment that was not designed explicitly to test dynamic modeling, we need to be wary of any *ad hoc* interpretations concerning the estimated parameter dynamics.

Nevertheless, some of the recovered patterns may suggest interpretable underlying changes. For instance, the threshold parameter seemed to decrease within an experimental session for many individuals. This indicates that participants generally responded less cautiously toward the end of an experimental session. A plausible explanation for this change in response caution might be that participants became increasingly bored during a session and started to decrease their ambitions. Note that current DDM modeling approaches rarely account for

such variation in the threshold parameter. Further, the drift rates generally tended to increase over time, suggesting that participants' increased their information processing speed over time. A change in the average rate of information uptake typically results in shorter RTs, which is precisely what we observed in most individual data sets (cf. Appendix). These increases in drift rates over time could imply the occurrence of learning effects. An important next step will be to tailor experiments with systematic manipulations from which we expect specific changes in some cognitive process and test whether the estimated parameter dynamics exhibit these changes.

Notwithstanding, our neural method has certain limitations. As can be seen in Fig. 3, the values for most parameters change strongly at the beginning of the time series. One could be tempted to (falsely) claim that the psychological constructs mapped to these parameters drastically change at the beginning of the first session of the experiment. However, these early parameter trajectories should be interpreted with great caution as they can be quite dependent on the initial prior. As a result, we cannot easily differentiate between initially large Bayesian updates to move away from the prior or actual changes in the underlying process. As is the case for any dynamic process, our modeling approach may also not be sensible for data sets with few observations. In the context of psychological experiments, a possible remedy could be to use burn-in trials at the beginning of an experiment that only serves the purpose of having some data points to inform the plausible parameter values. At the same time, these could serve as practice trials during which participants get accustomed to the task.

Furthermore, the simulation study has demonstrated that the non-stationary DDM exhibits a good performance in recovering parameters across various scenarios. However, it is essential to acknowledge that there still exists an error between the true and estimated parameters. Especially for the drift rate parameter errors around 0.25 have been observed frequently. Consequently, interpreting small local changes in parameter values requires caution. Despite this limitation, we firmly believe that the proposed method excels particularly in scenarios where moderately large changes in parameters are expected to occur over the course of a couple of time steps.

Another limitation concerns the implementation of the low-level mechanistic model, that is, the DDM itself. We assumed four different drift rates—one for each stimulus type—which is the standard procedure used in the application of stationary DDMs². This parameter is usually regarded as a proxy for average information uptake speed. However, in theory, there should only be one drift rate per participant³ that changes over time, for instance, due to experimental manipulation. Thus, a non-stationary DDM could also incorporate only one drift rate parameter. In our experiment, the manipulation (i.e., four conditions) was randomized throughout the experiment. This implies that besides fluctuation stemming from other sources, the drift rate would “jump” from trial to trial based on this change in task difficulty. To account for these jumps, we would need a different high-level transition model whose changes can be bigger than what a smooth Gaussian process or Gaussian random walk allows. In order to keep the content of this article manageable, we decided against proposing a novel transition model.

Finally, there are numerous degrees of freedom when implementing a computational model – not only with respect to the low-level observation model, but also regarding the high-level transition model. Exploring different model specifications and then deciding which is the most sensible for the type of task and data at hand requires Bayesian model comparison. Concerning dynamic cognitive models, it would be of particular interest to test which high-level transition model specification is most plausible for a given setting²⁴. Since Bayesian model comparison is a topic in its own right, future studies should investigate the utility of simulation-based methods^{59,60} for comparing competing superstatistical models.

We acknowledge that our study may not provide a definitive argument for when and why a non-stationary DDM is superior to a static DDM. The primary objective of this article is to showcase the implementation of non-stationary parameters within a superstatistics framework. However, we believe that the superstatistics framework, coupled with powerful neural approximators, gives rise to many new modeling opportunities and makes it possible to augment virtually any computational model with time-varying parameters. We think that there are many interesting research questions out there that could be investigated with the approach we propose in this work. Future studies can use our method to estimate even more challenging cognitive models than the DDM explored in this work and further extend its scope beyond cognitive science and psychology.

Methods

Experimental tasks. *Random-dot motion task.* The data set used in this study was adopted from⁵⁵. It includes data from 58 individuals, after excluding participants with a response accuracy below 70%. Each individual was randomly assigned to one of six groups, which were formed by two factors: the *time* versus *trial* condition and three levels of feedback details. During the experiment, participants solved a total of 24 blocks of the task. In the *trial* condition, each block comprised 40 trials, whereas in the *time* condition, each block lasted for 1 minute. In each trial, participants were presented with a random dot kinematogram and were required to determine if some of the dots coherently moved to the top-left or top-right direction. For more in-depth information about the experimental setup and methodology, refer to the comprehensive details provided in⁵⁵.

Lexical decision task. A total of 11 students from Heidelberg University participated in the experiment. Their average age was 23.81 ($SD = 3.30$) and 10 of the participants were female. All individuals gave written informed consent to the study, which was approved by the local ethics committee. The study was conducted according to the ethical declarations of Helsinki.

The participants performed a lexical decision-making task. On each trial, they had to assess if a presented letter string was a German word. As stimuli, we used high and low-frequency words, pseudo words that were generated by replacing vowels of existing words, and random letter strings. These four experimental conditions were pseudo-randomly presented throughout 3200 trials. All participants solved their task on 4 separate days (sessions) consisting of 800 trials each. The sessions were further split into 8 blocks of 100 trials with short breaks

between them. On each trial, participants' choice (German word; non-German word) and response time was recorded.

Model family. Following²⁴, we consider dynamic models that entail a low-level model with time-dependent parameters θ_t , which vary according to a high-level model with static parameters η . The low-level model is defined by a likelihood function \mathcal{L} , and the high-level model consists of a transition function \mathcal{T} .

In this work, we aim to tackle general superstatistical models for which the low-level model likelihood \mathcal{L} may not be available in closed-form. Such models are implemented as randomized stateful simulators that generate observable trajectories $\{x_t\}_{t=1}^T$ via the following (very general) recurrent system:

$$\theta_t = \mathcal{T}(\theta_{0:t-1}, \eta, \xi_t) \quad \text{with} \quad \xi_t \sim p(\xi|\eta) \quad (1)$$

$$x_t = \mathcal{G}(x_{1:t-1}, \theta_t, z_t) \quad \text{with} \quad z_t \sim p(z|\theta_t). \quad (2)$$

In the above equation, \mathcal{T} is an arbitrary high-level transition function parameterized by η , \mathcal{G} stands for an arbitrary (non-linear) transformation which encodes the functional assumptions of the low-level model. $\xi_t \sim p(\xi)$ and $z_t \sim p(z)$ are sources of random noise. The initial parameter configuration θ_0 follows a prior distribution $\theta_0 \sim p(\theta)$ which encodes available information about plausible parameter values.

One example of a transition model \mathcal{T} is a convolution with a Gaussian distribution, which implies a gradual change in the low-level model's parameters resembling a random walk:

$$\mathcal{T}(\theta_{t-1}, \eta, \xi_t) = \theta_{t-1} + \eta \xi_t \quad \text{with} \quad \xi_t \sim \mathcal{N}(0, 1). \quad (3)$$

Another similar example would be a convolution with a fat-tailed distribution, allowing for abrupt changes in the parameter space. Furthermore, since our simulation-based setting is not limited to transition models with a Markov property, we can also test more complex transitions, such as a vector autoregression (VAR⁶¹):

$$\mathcal{T}(\theta_{t-p:t-1}, \eta, \xi_t) = c + A_1\theta_{t-1} + \dots + A_p\theta_{t-p} + \xi_t, \quad (4)$$

where p is the order of the VAR model (i.e., its look-back period), $\xi_t \sim \mathcal{N}(0, \sigma)$, and $\eta = \{c, A_1, \dots, A_p, \sigma\}$ are the high-level parameters of the model.

We can even test transition models which depend on the entire history of the process, such as a Gaussian process (GP⁶²)

$$\theta_{1:T} \sim \mathcal{GP}(\mu_\theta, K_\theta) \quad (5)$$

with mean function μ_θ and covariance function K_θ defined through the vector of time indices. The high-level parameters η in this case would be the free kernel parameters, such as the amplitude σ or the length-scale l of a Gaussian kernel

$$k(\theta_t, \theta_{t'}) = \sigma^2 \exp\left(-\frac{\|\theta_t - \theta_{t'}\|^2}{2l^2}\right). \quad (6)$$

A typical task in Bayesian analysis of dynamic systems is to recover both the entire trajectory of dynamic parameters $\{\theta_t\}_{t=1}^T$ as well as the vector of static parameters η . Since for many discrete dynamic systems, the current data point x_t depends on the current parameter configuration θ_t as well as on the observable history of the system $x_{1:t-1}$, we can write the (implicit) point-wise likelihood as

$$\mathcal{L}_t = p(x_t|x_{1:t-1}, \theta_t). \quad (7)$$

The point-wise likelihood describes the probability of each data point, given the parameter values of the same time step and all past data points²⁴. Notably, we do not require this likelihood to be available in closed-form; we only need the ability to generate random draws through the forward-time generative process specified by Eq. (1).

Assuming the above factorization of the likelihood is possible, we aim to estimate the joint *filtering* posterior distribution of θ_t and η up to each discrete time-step t

$$p(\theta_t, \eta|x_{1:t}) \propto \mathcal{L}_t p(\theta_t|x_{1:t-1}, \eta) p(\eta|x_{1:t-1}). \quad (8)$$

This posterior encodes the reduction in uncertainty regarding the dynamic states evolving over time and the static parameter values being increasingly constrained by the data. From this joint distribution, we can derive the corresponding marginal posteriors as follows:

$$p(\theta_t|x_{1:t}) = \int p(\theta_t, \eta|x_{1:t}) d\eta, \quad (9)$$

$$p(\eta|x_{1:t}) = \int p(\theta_t, \eta|x_{1:t}) d\theta_t. \quad (10)$$

These distributions describe the average parameter dynamics over all possible high-level parameters and the best estimate for the high-level parameters up to discrete time-step t , respectively. Thus, learning both distributions amounts to standard Bayesian updating with an additional uncertainty factor due to the high-level transition model \mathcal{T} . Thus, posterior contraction over time will strongly depend on the form of the transition model and

may even increase in some cases, such as models allowing for sudden “jumps” in their parameters (i.e., regime switching behavior).

Neural Bayesian estimation. Various methods for estimating dynamic models have been proposed in the literature. Markov chain Monte Carlo (MCMC) methods offer a viable but computationally demanding approach based on random draws from the posterior⁶³. Variational inference (VI) methods approximate the true target posterior with simple, tractable densities and thus are a faster alternative to MCMC at the cost of a potential loss of posterior accuracy⁶³. A recent promising approach for low-dimensional problems is the grid-based method of²⁴, which represents parameter distribution on discrete lattices and enables efficient approximation of model evidence.

However, the above methods all depend on the ability to evaluate the likelihood function \mathcal{L}_t at each time point explicitly. This restriction makes it impossible to efficiently test the growing number of simulator-based or non-analytic models of cognition to observed data^{45,64}. Furthermore, MCMC and standard variational methods cannot leverage experience and require the same repeated computational effort for every new data set. For instance, when multiple participants complete a cognitive task, the same estimation procedures need to be repeated for each participant from scratch. Differently, hierarchical Bayesian models can be employed to jointly estimate group- and participant-level parameters, but they come with high computational costs and also rely on a closed-form likelihood function.

In contrast, *amortized inference* refers to methods with a “pre-paid” computational cost - after an expensive optimization or training phase, the same procedure can be instantly applied to any data set whose structure is compatible with the model^{45,46}. As a useful “side effect”, amortization allows us to easily perform extensive checks of computational faithfulness and parameter recoverability “in silico”, since we can obtain posterior samples from hundreds or even thousands of simulated data sets by applying the same pre-trained network. Amortized Bayesian inference is typically realized by specialized neural networks, which are trained to become estimation experts from repeated model simulations^{45,65}. The architecture of these networks can easily encode the probabilistic symmetry of the data, for instance, recurrent networks for temporal data⁶⁶ or permutation-invariant networks for IID data⁶⁷.

Crucially, dynamic models with time-varying parameters present a challenge to existing neural architectures since they induce a new joint posterior at each time-step $p(\theta_t, \eta|x_{1:t})$. However, most previous architectures can only estimate a single set of parameters with no temporal information^{45,47,65}. Thus, we propose to use a recurrent probabilistic neural architecture that estimates the joint posterior over all static and dynamic parameters for all discrete time points in a single forward pass.

Recurrent estimation method. Our proposed architecture consists of several neural components. First, a recurrent neural network (RNN) with learnable parameters $\psi^{(r)}$ embodying long short-term memory (LSTM) consumes the observed data sequentially:

$$h_t = \text{LSTM}(x_t, h_{t-1}; \psi^{(r)}), \quad (11)$$

where the hidden state h_t at each time point t represents the internal memory of the network over arbitrary temporal intervals. Thus, we can treat h_t as a compact representation of the observable history up to time point t . We employ a standard LSTM network, which consists of three gates: an input gate, an output gate, and a forget gate. These gates are responsible for weighing and integrating old and new information. Importantly, LSTM networks can naturally deal with sequences of varying length, which enables them to process streams of “online” data⁶⁶.

In order to recover the time-varying parameters θ_t of the low-level model as well as the static high-level parameters η , we use the hidden state h_t as a conditioning vector for a generative neural network with trainable weights $\psi^{(g)}$. This network can be implemented as a conditional variant of any popular generative architecture for inference, such as coupling networks⁶⁸, autoregressive flows⁶⁹, or standard neural networks with probabilistic outputs⁷⁰. The generative network is responsible for approximating the current joint posterior up to time step t given the outputs of the recurrent summary network: $q(\theta_t, \eta|x_{1:t}, \psi) \equiv q(\theta_t, \eta|h_t, \psi)$. To reduce notational clutter, we set $\psi = (\psi^{(r)}, \psi^{(g)})$ and assume that h_t is expressive enough to encode all information contained in the data for correctly updating the prior (i.e., h_t is a maximally informative *summary statistic* of $x_{1:t}$).

Alternatively, we can also directly target one of the two equivalent factorizations of the joint posterior, namely:

$$p(\theta_t, \eta|x_{1:t}) = p(\theta_t|x_{1:t}, \eta) p(\eta|x_{1:t}) \quad (12)$$

$$= p(\eta|x_{1:t}, \theta_t) p(\theta_t|x_{1:t}). \quad (13)$$

While being mathematically equal, these factorizations imply different neural architectures and corresponding ancestral sampling schemes. The former factorization (Eq. 12) requires a generative network for first sampling the high-level parameters from $p(\eta|x_{1:t})$ and then sampling the low-level parameters from $p(\theta_t|x_{1:t}, \eta)$, conditional on the sampled high-level parameters. On the other hand, the latter factorization (Eq. 13) requires a generative network for first sampling the low-level parameters from $p(\theta_t|x_{1:t})$ and then sampling the high-level parameters from $p(\eta|x_{1:t}, \theta_t)$, conditional on the sampled low-level parameters.

In the current work, we consistently target the factorization in Eq. (12), but we were able to obtain comparable filtering results with either ancestral sampling strategy. In practice, we can either assume a multivariate Gaussian posterior for $q(\theta_t|x_{1:t}, \eta, \psi)$ and $q(\eta|x_{1:t}, \psi)$ as a dynamic extension of the basic method in⁷¹ or estimate

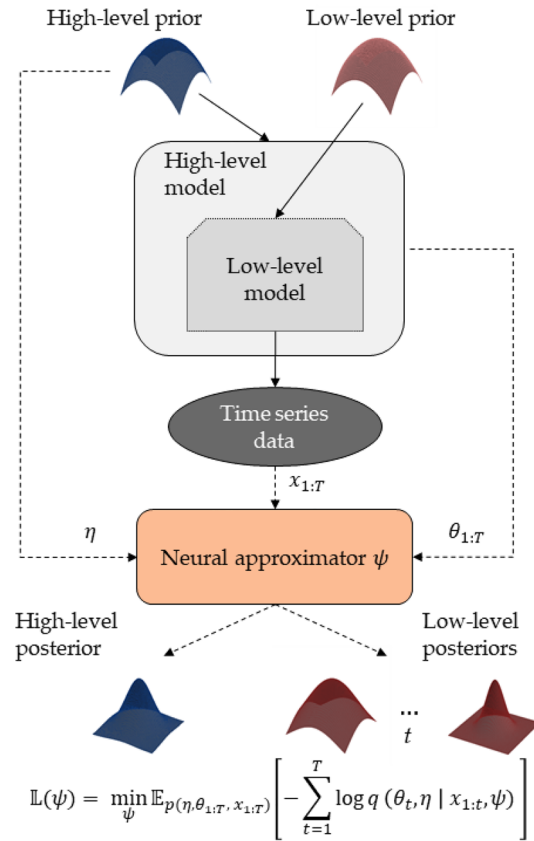


Figure 9. A graphical illustration of our neural inference method. A recurrent neural approximator updates the posterior of the low-level model parameters θ_t each time step t and yields the posterior over the high-level model parameters η considering all available data. The low-level prior constrains the initial dynamic parameter values θ_0 , which then get passed to the high-level transition model. Together, the two priors and the two models comprise a stochastic simulator that trains the neural approximator to perform amortized Bayesian updating.

free-form posteriors as a dynamic extension of the BayesFlow method⁴⁵. We use the former approach for the toy Coal Mining benchmark and the latter approach for all other experiments in this work.

Simulation-based training. Figure 9 graphically illustrates the rationale of our simulation-based inference approach. To train the networks, we treat the forward-time generative model as a simulator and employ Eq. (1) to generate multiple sets of simulated parameters and trajectories $(\eta, \theta_{1:T}, x_{1:T})$. We then minimize the Monte Carlo estimate of the following criterion

$$\mathbb{L}(\psi) = \min_{\psi} \mathbb{E}_{p(\eta, \theta_{1:T}, x_{1:T})} \left[- \sum_{t=1}^T \log q(\theta_t, \eta | x_{1:t}, \psi) \right], \tag{14}$$

where $\mathbb{E}[\cdot]$ denotes an expectation over the dynamic generative model and $\psi = (\psi^{(r)}, \psi^{(g)})$ denotes the collection of all trainable neural network parameters. This criterion ensures that the approximate posteriors match the analytic posteriors induced by the dynamic model and can be minimized either via online (i.e., generating dynamic simulations on the fly) or via offline training (i.e., using a set of pre-computed dynamic simulations).

A non-stationary diffusion decision model. To illustrate the potential of our approach, we will reformulate in superstatistical terms a popular cognitive model for analyzing human response times (RTs) in binary decision tasks, namely the DDM. The standard DDM describes the microscopic dynamics of perceptual evidence accumulations via a simple stochastic ordinary differential equation (SDE). Accordingly, the accumulated evidence x_j in experimental task j follows a random walk with drift and Gaussian noise:

$$dx_j = \nu dt_s + z \sqrt{dt_s} \quad \text{with} \quad z \sim \mathcal{N}(0, 1), \tag{15}$$

where t_s represents time on a continuous microscopic scale (i.e., during forced-choice decision making). A core assumption of the DDM is that task-relevant information (i.e., perceptual evidence) accumulates at a constant rate (ν). This process runs in a corridor with two absorbing boundaries, which represent two decision alternatives. As

soon as the accumulated evidence x_j reaches either a pre-defined threshold (a) or 0, the model makes a categorical decision D_j for the alternative favored by the collected evidence:

$$D_j = \begin{cases} 1, & \text{if } x_j \geq a \\ 0, & \text{if } x_j \leq 0 \end{cases} \quad (16)$$

Further, the model assumes a constant additive factor (τ) accounting for non-decision processes, such as encoding or motor responses. Thus, the standard (static) DDM has three key parameters $\theta = (v, a, \tau)$. The starting point of the decision process is either estimated as an additional parameter or fixed at $a/2$.

The typical assumption of the standard DDM is that the parameters θ remain stationary for the duration of a given cognitive task. In order to relax this restrictive assumption, the standard DDM has been extended to incorporate so-called inter-trial-variability for the drift rate and non-decision time parameters^{72,73}. In this way, the extended DDM concedes that these cognitive parameters are not static but vary over time. However, the assumed variation is homogeneous and memoryless, and the generative model still yields IID data, that is, the transition model coincides with independent sampling and reduces to $\theta_t = \mathcal{T}(\eta, \xi_t)$.

In contrast, our superstatistical model assumes a stateful Gaussian process (GP) high-level model, which describes the trial-by-trial dynamics of the DDM parameters according to Eqs. (5) and (6) (see the Appendix for more details).

Thereby, we want to demonstrate that our estimation method can tackle very flexible transition models \mathcal{T} , as long as we can simulate data from the low-level model. However, we also fit a DDM with a simpler Gaussian random walk transition model to the data described in the “Human data application” section. This simpler model corroborates our findings by suggesting qualitatively similar parameter dynamics, but yields less sharp predictions on unseen data than its GP counterpart (see Appendix for more details).

Data and code availability

All models, data, and scripts for reproducing the results of this work are publicly available in the project’s repository <https://github.com/bayesflow-org/Neural-Superstatistics>. The neural superstatistics method is implemented in the `BayesFlow` Python library for amortized Bayesian workflows⁷⁴.

Received: 15 March 2023; Accepted: 8 August 2023

Published online: 23 August 2023

References

- Farrell, S. & Lewandowsky, S. *Computational Modeling of Cognition and Behavior*. (Cambridge University Press). <https://doi.org/10.1017/CBO9781316272503> (2018).
- Voss, A., Nagler, M. & Lerche, V. Diffusion models in experimental psychology. *Exp. Psychol.* **60**(6), 385–402. <https://doi.org/10.1027/1618-3169/a000218> (2013).
- Ratcliff, R., Smith, P. L., Brown, S. D. & McKoon, G. Diffusion decision model: Current issues and history. *Trends Cogn. Sci.* **20**(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007> (2016).
- Eckstein, M. K. & Collins, A. G. E. Computational evidence for hierarchically structured reinforcement learning in humans. *Proc. Natl. Acad. Sci.* **117**(47), 29381–29389. <https://doi.org/10.1073/pnas.1912330117> (2020).
- Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Ann. Rev. Psychol.* **68**, 101–128. <https://doi.org/10.1146/annurev-psych-122414-033625> (2017).
- Oberauer, K. *et al.* Benchmarks for models of short-term and working memory. *Psychol. Bull.* **144**(9), 885–958. <https://doi.org/10.1037/bul0000153> (2018).
- Yoo, A. H. & Collins, A. G. E. How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. *J. Cogn. Neurosci.* **34**(4), 551–568. https://doi.org/10.1162/jocn_a_01808 (2022).
- Van Orden, G. C., Holden, J. G. & Turvey, M. T. Self-organization of cognitive performance. *J. Exp. Psychol. Gen.* **132**(3), 331–350. <https://doi.org/10.1037/0096-3445.132.3.331> (2003).
- Wagenmakers, E.-J., Farrell, S. & Ratcliff, R. Estimation and interpretation of $1/F\alpha$ noise in human cognition. *Psychon. Rev.* **11**(4), 579–615. <https://doi.org/10.3758/BF03196615> (2004).
- Gilden, D. L. Cognitive emissions of $1/f$ noise. *Psychol. Rev.* **108**(1), 33–56. <https://doi.org/10.1037/0033-295x.108.1.33> (2001).
- Collins, A. G. E. & Frank, M. J. Withinand across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proc. Natl. Acad. Sci.* **115**(10), 2502–2507. <https://doi.org/10.1073/pnas.1720963115> (2018).
- Brockmole, J. R. & Logie, R. H. Age-related change in visual working memory: A study of 55,753 participants aged 8–75. *Front. Psychol.* **4**, 12. <https://doi.org/10.3389/fpsyg.2013.00012> (2013).
- von Krause, M., Radev, S. T. & Voss, A. Mental speed is high until age 60 as revealed by analysis of over a million participants. *Nat. Hum. Behav.* **6**, 700–708. <https://doi.org/10.1038/s41562-021-01282-7> (2022).
- Riley, M. A. & Holden, J. G. Dynamics of cognition. *WIREs Cogn. Sci.* **3**(6), 593–606. <https://doi.org/10.1002/wcs.1200> (2012).
- Favela, L. H. Cognitive science as complexity science. *WIREs Cogn. Sci.* **11**(4), e1525. <https://doi.org/10.1002/wcs.1525> (2020).
- Ratcliff, R. & Van Dongen, H. P. Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proc. Natl. Acad. Sci.* **108**(27), 11285–11290. <https://doi.org/10.1073/pnas.1100483108> (2011).
- Walsh, M. M., Gunzelmann, G. & Van Dongen, H. Computational cognitive modeling of the temporal dynamics of fatigue from sleep loss. *Psychon. Bull. Rev.* **24**(6), 1785–1807. <https://doi.org/10.3758/s13423-017-1243-6> (2017).
- Kahana, M. J., Aggarwal, E. V. & Phan, T. D. The variability puzzle in human memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**(12), 1857. <https://doi.org/10.1037/xlm0000553> (2018).
- Evans, N. J., Brown, S. D., Mewhort, D. J. & Heathcote, A. Refining the law of practice. *Psychol. Rev.* **125**(4), 592. <https://doi.org/10.1037/rev0000105> (2018).
- Mittner, M., Hawkins, G. E., Boebel, W. & Forstmann, B. U. A neural model of mind wandering. *Trends Cogn. Sci.* **20**(8), 570–578. <https://doi.org/10.1016/j.tics.2016.06.004> (2016).
- Christoff, K., Irving, Z. C., Fox, K. C., Spreng, R. N. & Andrews-Hanna, J. R. Mind-wandering as spontaneous thought: A dynamic framework. *Nat. Rev. Neurosci.* **17**(11), 718–731. <https://doi.org/10.1038/nrn.2016.113> (2016).
- Brosowsky, N. P., DeGutis, J., Esterman, M., Smilek, D. & Seli, P. Mind wandering, motivation, & task performance over time: Evidence that motivation insulates people from the negative effects of MindWandering. *Psychol. Conscious. Theory Res. Pract.* <https://doi.org/10.1037/cns0000263> (2020).

23. Kiuru, N. *et al.* The dynamics of motivation, emotion, and task performance in simulated achievement situations. *Learn. Individ. Differ.* **80**, 101873. <https://doi.org/10.1016/j.lindif.2020.101873> (2020).
24. Mark, C. *et al.* Bayesian model selection for complex dynamic systems. *Nat. Commun.* **9**(1), 1803. <https://doi.org/10.1038/s41467-018-04241-5> (2018).
25. Beck, C. & Cohen, E. G. D. Superstatistics. *Phys. A Stat. Mech. Appl.* **322**, 267–275. [https://doi.org/10.1016/S0378-4371\(03\)00019-0](https://doi.org/10.1016/S0378-4371(03)00019-0) (2003).
26. Hanel, R., Thurner, S. & Gell-Mann, M. Generalized entropies and the transformation group of superstatistics. *Proc. Natl. Acad. Sci.* **108**(16), 6390–6394. <https://doi.org/10.1073/pnas.1103539108> (2011).
27. Kucharský, Š, Tran, N.-H., Veldkamp, K., Raijmakers, M. & Visser, I. Hidden Markov models of evidence accumulation in speeded decision tasks. *Comput. Brain Behav.* **4**(4), 416–441. <https://doi.org/10.1007/s42113-021-00115-0> (2021).
28. Gunawan, D., Hawkins, G. E., Kohn, R., Tran, M.-N. & Brown, S. D. Time-evolving psychological processes over repeated decisions. *Psychol. Rev.* **129**(3), 438. <https://doi.org/10.1037/rev0000351> (2022).
29. Metzner, C., Schilling, A., Traxdorf, M., Schulze, H. & Krauss, P. Sleep as a random walk: A super-statistical analysis of EEG data across sleep stages. *Commun. Biol.* **4**(1), 1–11. <https://doi.org/10.1038/s42003-021-02912-6> (2021).
30. Yalcin, G. C., Rabassa, P. & Beck, C. Extreme event statistics of daily rainfall: Dynamical systems approach. *J. Phys. A Math. Theor.* **49**(15), 154001. <https://doi.org/10.1088/1751-8113/49/15/154001> (2016).
31. Rabassa, P. & Beck, C. Superstatistical analysis of sea-level fluctuations. *Phys. A Stat. Mech. Appl.* **417**, 18–28. <https://doi.org/10.1016/j.physa.2014.08.068> (2015).
32. Williams, G., Schäfer, B. & Beck, C. Superstatistical approach to air pollution statistics. *Phys. Rev. Res.* **2**(1), 013019. <https://doi.org/10.1103/PhysRevResearch.2.013019> (2020).
33. Bogachev, M. I., Markelov, O. A., Kayumov, A. R. & Bunde, A. Superstatistical model of bacterial DNA architecture. *Sci. Rep.* **7**(1), 43034. <https://doi.org/10.1038/srep43034> (2017).
34. Van der Straeten, E. & Beck, C. Superstatistical fluctuations in time series: Applications to share-price dynamics and turbulence. *Phys. Rev. E* **80**(3), 036108. <https://doi.org/10.1103/PhysRevE.80.036108> (2009).
35. Denys, M., Gubiec, T., Kutner, R., Jagielski, M. & Stanley, H. E. Universality of market superstatistics. *Phys. Rev. E* **94**(4), 042305. <https://doi.org/10.1103/PhysRevE.94.042305> (2016).
36. Voss, A. & Voss, J. Fast-Dm: A free program for efficient diffusion model analysis. *Behav. Res. Methods* **39**(4), 767–775. <https://doi.org/10.3758/BF03192967> (2007).
37. Ratcliff, R. A theory of memory retrieval. *Psychol. Rev.* **85**(2), 59–108. <https://doi.org/10.1037/0033-295X.85.2.59> (1978).
38. von Krause, M. *et al.* Stability and change in diffusion model parameters over two years. *J. Intell.* **9**(2), 26. <https://doi.org/10.3390/jintelligence9020026> (2021).
39. Diederich, A. & Busemeyer, J. R. Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Percept. Psychophys.* **68**(2), 194–207. <https://doi.org/10.3758/BF03193669> (2006).
40. Urai, A. E., de Gee, J. W., Tsetsos, K. & Donner, T. H. Choice history biases subsequent evidence accumulation. *eLife* **8**, e46331. <https://doi.org/10.7554/eLife.46331> (2019).
41. van Rooij, M. M. J. W., Favela, L. H., Malone, M. & Richardson, M. J. Modeling the dynamics of risky choice. *Ecol. Psychol.* **25**(3), 293–303. <https://doi.org/10.1080/10407413.2013.810502> (2013).
42. Gasimova, F. *et al.* Dynamical systems analysis applied to working memory data. *Front. Psychol.* **5**, 687. <https://doi.org/10.3389/fpsyg.2014.00687> (2014).
43. Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. Bayesian workflow. <https://doi.org/10.48550/ARXIV.2011.01808> (2020).
44. van de Schoot, R. *et al.* Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* **1**(1), 1–26. <https://doi.org/10.1038/s43586-020-00001-2> (2021).
45. Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L. & Köthe, U. BayesFlow: Learning complex stochastic models with invertible neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **33**(4), 1452–1466. <https://doi.org/10.1109/TNNLS.2020.3042395> (2020).
46. Mestdagh, M., Verdonck, S., Meers, K., Loossens, T. & Tuerlinckx, F. Prepaid parameter estimation without likelihoods. *PLoS Comput. Biol.* **15**(9), e1007181. <https://doi.org/10.1371/journal.pcbi.1007181> (2019).
47. Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *Proc. Natl. Acad. Sci.* **117**(48), 30055–30062. <https://doi.org/10.1073/pnas.1912789117> (2020).
48. Bürkner, P.-C., Scholz, M., & Radev, S. Some models are useful, but how do we know which ones? Towards a unified Bayesian model taxonomy. <https://doi.org/10.48550/arXiv.2209.02439> (2022).
49. Carpenter, B. *et al.* Stan: A probabilistic programming language. *J. Stat. Softw.* **76**(1), 1–32. <https://doi.org/10.18637/jss.v076.i01> (2017).
50. Neal, R. M. *et al.* MCMC using Hamiltonian dynamics. *Handb. Markov Chain Monte Carlo* **2**(11), 2. <https://doi.org/10.1201/b10905-7> (2011).
51. Betancourt, M. Calibrating model-based inferences and decisions. <https://doi.org/10.48550/ARXIV.1803.08393> (2018).
52. Schad, D. J., Betancourt, M. & Vasisht, S. Toward a principled Bayesian workflow in cognitive science. *Psychol. Methods* **26**, 103–126. <https://doi.org/10.1037/met0000275> (2021).
53. Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. Validating Bayesian inference algorithms with simulation-based calibration. <https://doi.org/10.48550/ARXIV.1804.06788> (2018).
54. Säilynoja, T., Bürkner, P.-C., & Vehtari, A. Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. <https://doi.org/10.48550/ARXIV.2103.10522> (2021).
55. Evans, N. J. & Brown, S. D. People adopt optimal policies in simple decision-making, after practice and guidance. *Psychon. Bull. Rev.* **24**, 597–606. <https://doi.org/10.3758/s13423-016-1135-1> (2017).
56. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **12**(6), 1100–1122. <https://doi.org/10.1177/1745691617693393> (2017).
57. Bürkner, P.-C., Gabry, J. & Vehtari, A. Approximate leave-future-out cross-validation for Bayesian time series models. *J. Stat. Comput. Simul.* **90**(14), 2499–2523. <https://doi.org/10.1080/00949655.2020.1783262> (2020).
58. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **13**(25), 723–773 (2012).
59. Radev, S. T., D'Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., & Bürkner, P.-C. Amortized Bayesian model comparison with evidential deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **34**(8), 4903–4917. <https://doi.org/10.1109/TNNLS.2021.3124052> (2021).
60. Schmitt, M., Radev, S. T., & Bürkner, P.-C. Meta-uncertainty in Bayesian model comparison. <https://doi.org/10.48550/ARXIV.2210.07278> (2022).
61. Toda, H. Y. & Phillips, P. C. Vector autoregression and causality: A theoretical overview and simulation study. *Econ. Rev.* **13**(2), 259–285. <https://doi.org/10.1080/07474939408800286> (1994).
62. Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning*. Vol. **3176** (eds Bousquet, O.) 63–71 (Springer, Berlin, Heidelberg). https://doi.org/10.1007/978-3-540-28650-9_4 (2003).

63. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *J. A. Stat. Assoc.* **112**(518), 859–877. <https://doi.org/10.1080/01621459.2017.1285773> (2017).
64. van Rooij, I., Blokpoel, M., Kwisthout, J. & Wareham, T. Applications. *Intractability Guide Class. Parameterized Complex. Anal.* <https://doi.org/10.1017/9781107358331> (2019).
65. Greenberg, D., Nonnenmacher, M., & Macke, J. Automatic posterior transformation for likelihood-free inference, in *International Conference on Machine Learning*, 2404–2414. <https://doi.org/10.48550/arXiv.1905.07488> (2019)
66. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Comput.* **12**(10), 2451–2471. <https://doi.org/10.1049/cp:19991218> (2000).
67. Bloem-Reddy, B. & Teh, Y. W. Probabilistic symmetries and invariant neural networks. *J. Mach. Learn. Res.* **21**, 90–1 (2020).
68. Ardizzone, L., Lüth, C., Kruse, J., Rother, C., & Köthe, U. Guided image generation with conditional invertible neural networks. <https://doi.org/10.48550/arXiv.1907.02392> (2019).
69. Papamakarios, G., Pavlakou, T., & Murray, I. Masked autoregressive flow for density estimation. *Adv. Neural Inf. Process. Syst.* **30** (2017).
70. Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **30** (2017).
71. Radev, S. T., Mertens, U. K., Voss, A. & Köthe, U. Towards end-to-end likelihood-free inference with convolutional neural networks. *Br. J. Math. Stat. Psychol.* **73**(1), 23–43. <https://doi.org/10.1111/bmsp.12159> (2020).
72. Ratcliff, R. & Rouder, J. N. Modeling response times for two-choice decisions. *Psychol. Sci.* **9**(5), 347–356. <https://doi.org/10.1111/1467-9280.00067> (1998).
73. Ratcliff, R. & Tuerlinckx, F. Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychon. Bull. Rev.* **9**(3), 438–481. <https://doi.org/10.3758/BF03196302> (2002).
74. Radev, S. T., Schmitt, M., Schumacher, L., Else Müller, L., Pratz, V., Schälte, Y., Köthe, U., & Bürkner, P.-C. BayesFlow: Amortized Bayesian workflows with neural networks. <https://doi.org/10.48550/arXiv.2306.16015> (2023).

Acknowledgements

We thank Daniel Durstewitz and Lasse Else Müller for their helpful feedback on this project. We also thank Marie Wieschen for her efforts in data collection. L.S. and A.V. were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; Grant Number GRK 2277 "Statistical Modeling in Psychology"). P.-C.B. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2075 - 390740016 (the Stuttgart Cluster of Excellence SimTech). S.T.R. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2181 - 390900948 (the Heidelberg Cluster of Excellence STRUCTURES). For the publication fee we acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding programme "Open Access Publikationskosten" as well as by Heidelberg University.

Author contributions

L.S., S.T.R., designed the research, created and applied the models, and wrote the initial draft of the manuscript. P.C.B. significantly contributed with his statistical and scientific expertise to the methods and the written content. A.V. and U.K. supervised the project and contributed to all stages.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-40278-3>.

Correspondence and requests for materials should be addressed to L.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Appendix

Implementation Details

All experiments, neural networks, and simulation models are implemented using the BayesFlow library <https://github.com/stefanradev93/BayesFlow> built on top of TensorFlow [75]. Code and further instructions for reproducing the results from all experiments and applications in the current manuscript is available at <https://github.com/bayesflow-org/Neural-Superstatistics>.

Stan Benchmark Study

Data Simulation

To simulate the 100 data sets each consisting of $T = 100$ trials, we used the standard DDM implementation (cf. equation (15) in the main text). The diffusion constant was fixed to 1 and the starting point parameter to 0.5 (i.e., symmetric starting point between the two decision boundaries). For data simulation, we randomly sampled parameter sets from the following prior distributions:

$$\begin{aligned} v &\sim \Gamma(5.0, \frac{1}{1.3}) \\ a &\sim \Gamma(4.0, \frac{1}{3}) \\ \tau &\sim \Gamma(1.5, \frac{1}{5}) \end{aligned}$$

where $\Gamma(a, b)$ denotes a Gamma distributions with shape a as the first and scale b as the second argument.

Non-Stationary DDM fitting

We fitted a separate non-stationary DDM with a Gaussian random walk transition model to all 100 simulated data sets. The same implementation and likelihood was used for Stan and our neural estimation method. However, all 3 parameters were allowed to vary according to a Gaussian random walk (cf. equation (3) in the main text). The starting values were sampled from the same prior distributions as in simulation. The hyperparameters of the random walk transition model were sampled from the following distribution:

$$s_v, s_a, s_\tau \sim \mathcal{B}(1, 25)$$

where $\mathcal{B}(\alpha, \beta)$ denotes a Beta distribution with α and β parameters. In order to avoid implausible parameter values, the time-varying parameters v_t, a_t, τ_t were clipped to lower bounds $[0, 0, 0]$ and upper bounds $[6, 4, 2]$, respectively.

We trained the neural approximator via online learning (i.e., simulations on the fly) for 50 epochs with 1000 iterations each and a batch size of 8. We use an Adam optimizer with an initial learning rate of 5×10^{-4} and a cosine learning rate decay schedule. After training the network, we draw 4000 posterior samples (the same as with Stan) for each of the 100 data sets.

Simulation Study

In what follows, we describe the settings for the four different simulation scenarios, namely, the static DDM, the DDM with stationary variability, the DDM with non-stationary variability, and the static DDM with random uniform jumps at pre-defined time steps (i.e., regime switching DDM). For each scenario, we simulated 200 data sets, each consisting of $T = 400$ time steps.

We trained the neural approximator via online learning (i.e., simulations on the fly) for 75 epochs with 1000 iterations each and a batch size of 8. We use an Adam optimizer with an initial learning rate of 5×10^{-4} and a cosine learning rate decay schedule. After training the network, we draw 4000 posterior samples (the same as with Stan) for each of the 100 data sets.

Static DDM

To simulate the 200 data sets for the static DDM scenario, we used the same prior and likelihood as in the **Stan Benchmark Study**.

Stationary Variability DDM

For the stationary variability DDM, we used the same DDM implementation as in the static DDM scenario except that we used the following variability statements:

$$\begin{aligned} v_t &\sim \mathcal{N}(v, v_s) \\ a_t &\sim \mathcal{N}(a, a_s) \\ \tau_s &\sim \mathcal{U}\left(\tau - \frac{\tau_s}{2}, \tau + \frac{\tau_s}{2}\right) \end{aligned}$$

where $\mathcal{N}(\mu, \sigma)$ denotes a Normal distribution with location μ and standard deviation σ and $\mathcal{U}(\text{lower}, \text{upper})$ denotes an Uniform distribution with a lower and an upper bound.

The newly introduced variability parameters (v_s, a_s, τ_s) were sampled from the following prior distributions:

$$v_s, a_s, \tau_s \sim \mathcal{TN}_{[0, \text{inf}]}(0, 0.1)$$

where $\mathcal{TN}_{[a, b]}(\mu, \sigma)$ denotes the truncated normal distribution with location μ and standard deviation σ truncated within the interval $[a, b]$.

To avoid implausible values the per trial parameters v_t, a_t, τ_t were bounded with lower bounds $[0, 0, 0]$ and upper bounds $[6, 4, 2]$ respectively.

Non-Stationary DDM

We used the same non-stationary DDM implementation as described in **Stan Benchmark Study**.

Regime Switching DDM

The regime switching DDM is basically the same implementation as the static DDM, but the parameter jumped uniformly at 3 specific time steps ($T = 100; T = 200; T = 300$) and stayed again constant after the jump:

$$\theta_t = \begin{cases} \theta_{t-1}, & \text{if } t \notin \{100, 200, 300\} \\ \mathcal{U}(\text{lower}, \text{upper}), & \text{if } t \in \{100, 200, 300\} \end{cases} \quad (17)$$

where the lower and upper bounds of the Uniform distributions are $[0, 0, 0]$ and $[6, 4, 2]$, respectively. The starting values of the parameters were once again sampled from the same prior distributions as in the static DDM.

Amortized inference

We fitted the same non-stationary DDM with a Gaussian transition model as described above to all four scenarios. To train the networks we used 75 epochs with 1000 iterations each and a batch size of 16. After training the network we fitted the model to each simulation of each scenario separately and obtained 2000 posterior samples.

True vs. Estimated Parameters

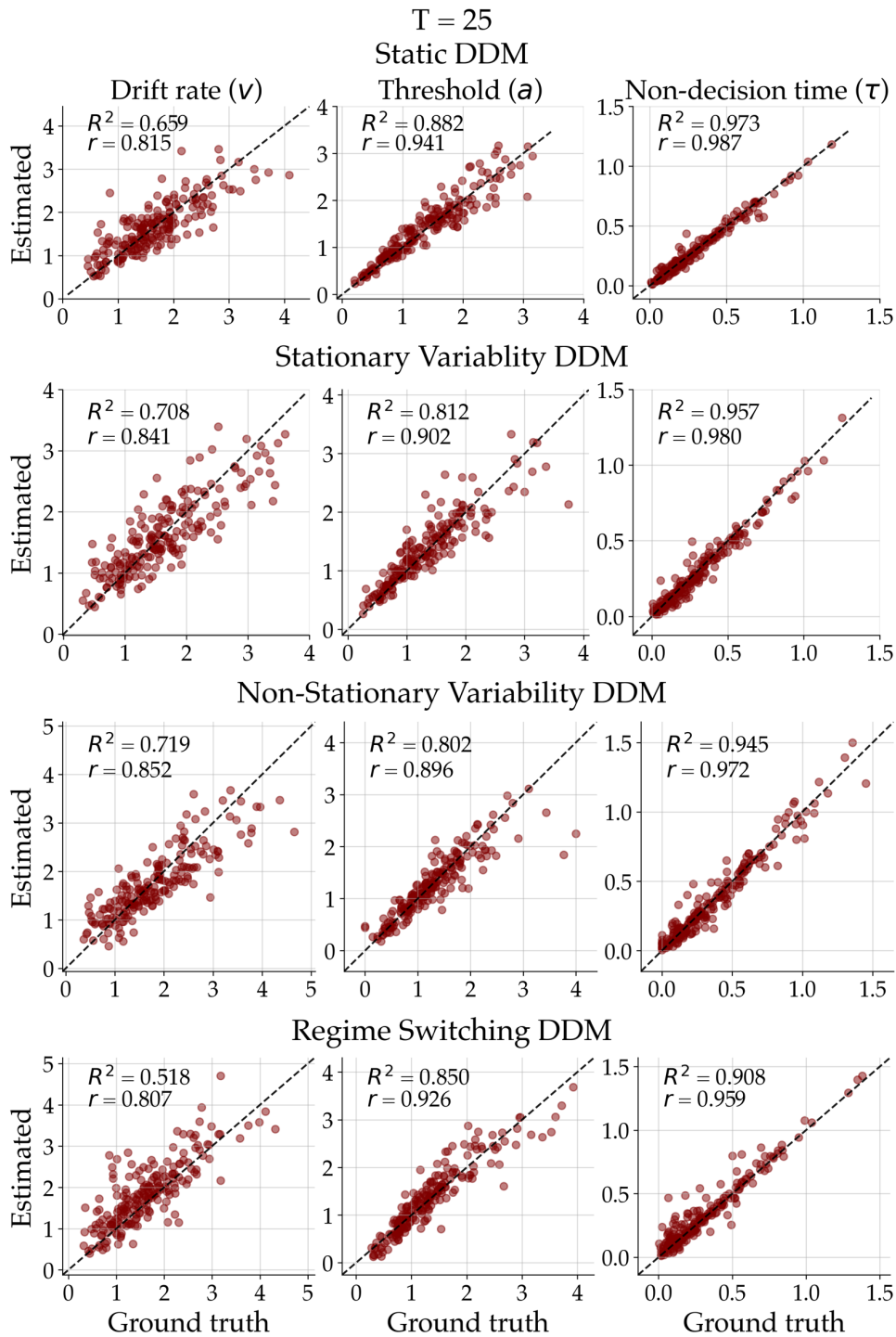


Figure A.10: True data generating parameters plotted against posterior means for all 3 parameters and simulation scenarios separately at time point $T = 25$.

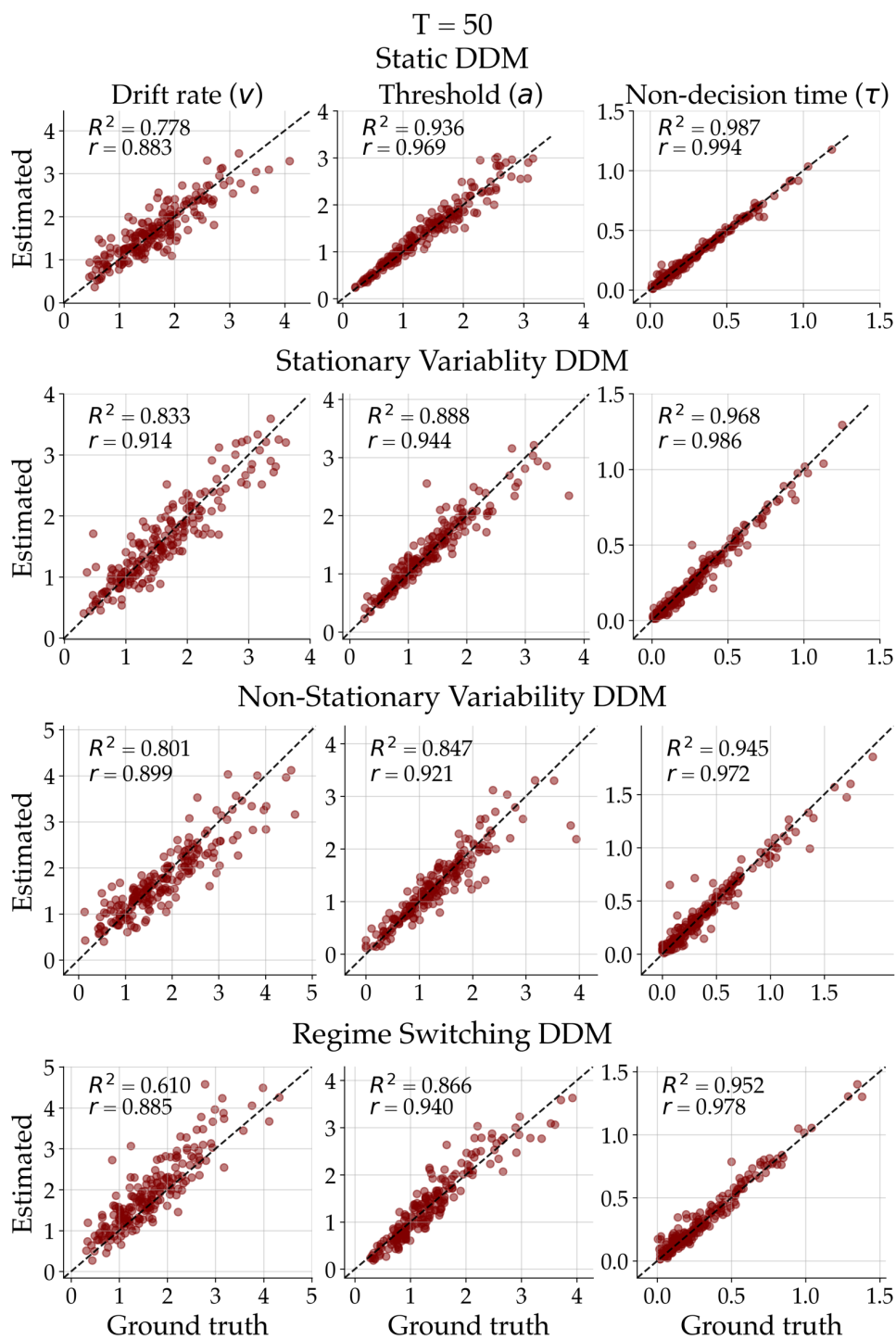


Figure A.11: True data generating parameters plotted against posterior means for all 3 parameters and simulation scenarios separately at time point $T = 50$.

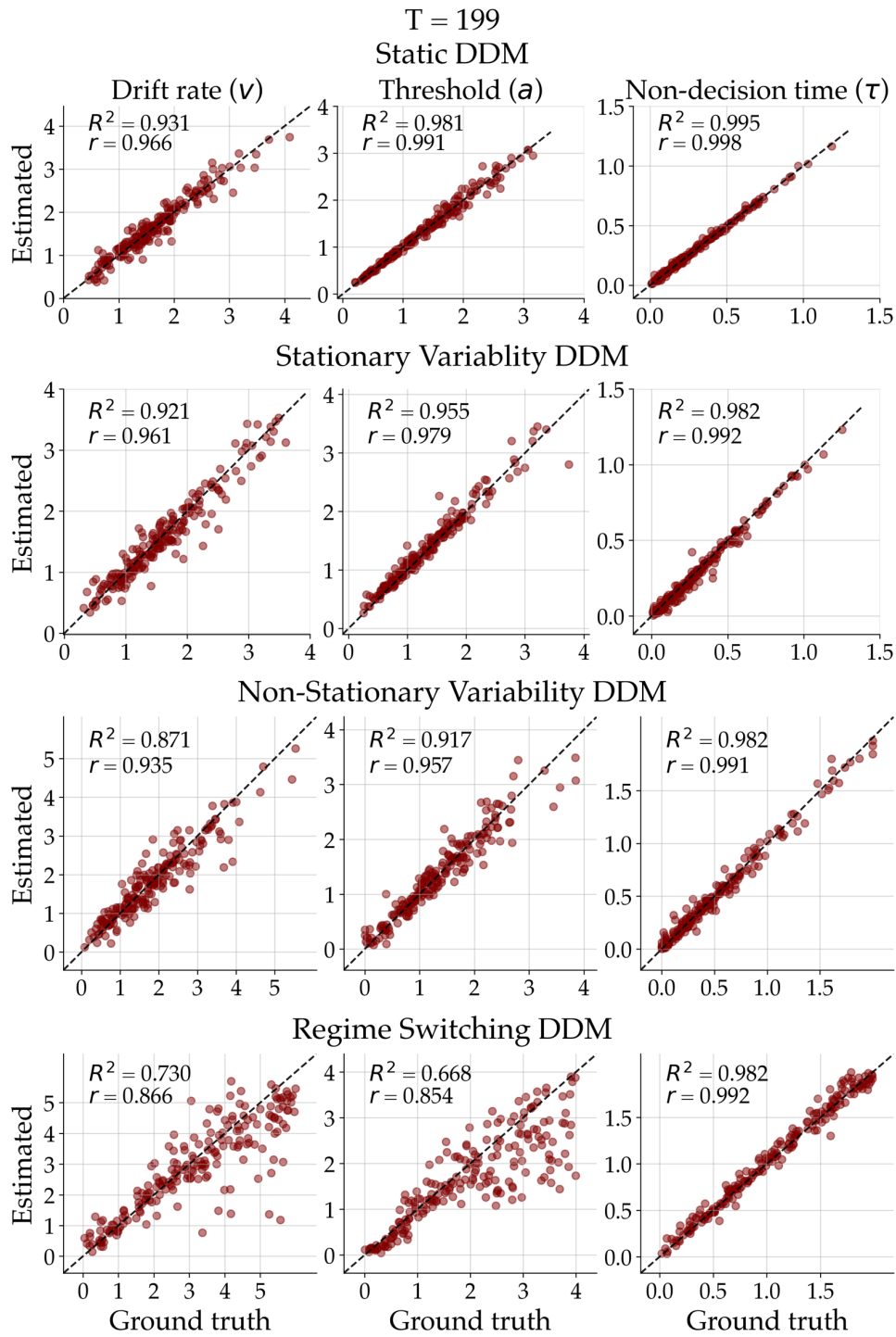


Figure A.12: True data generating parameters plotted against posterior means for all 3 parameters and simulation scenarios separately at time point $T = 199$.

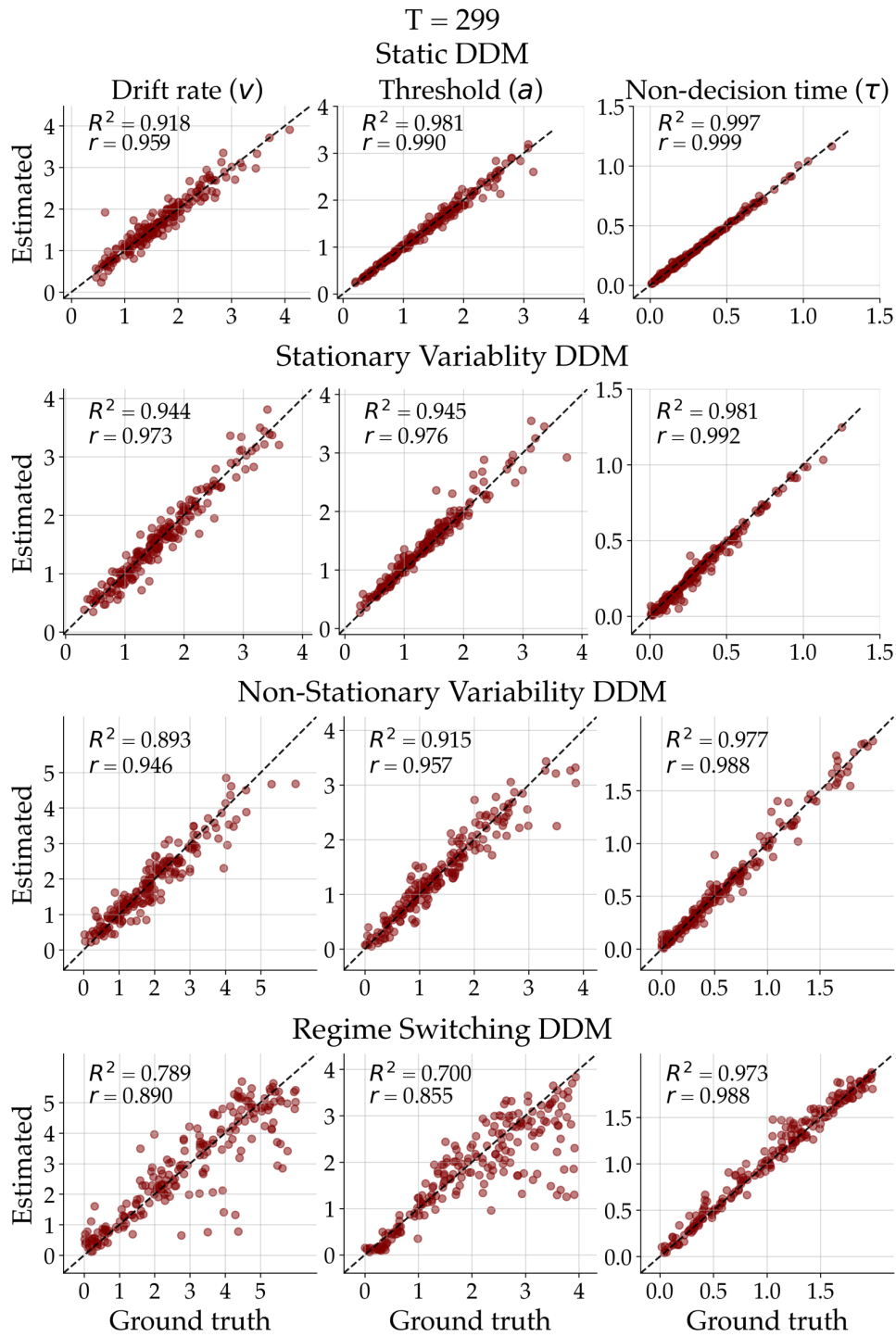


Figure A.13: True data generating parameters plotted against posterior means for all 3 parameters and simulation scenarios separately at time point $T = 299$.

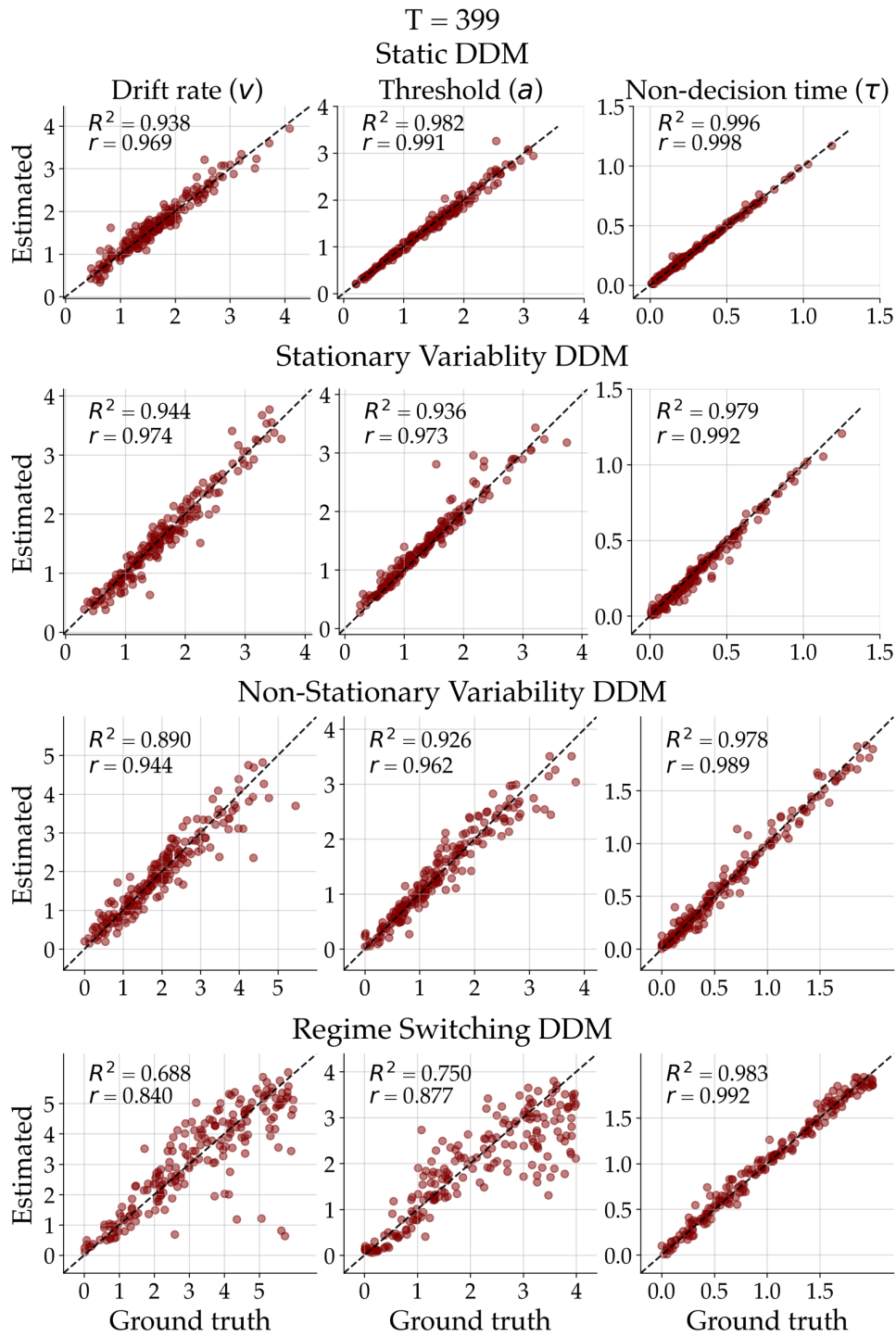


Figure A.14: True data generating parameters plotted against posterior means for all 3 parameters and simulation scenarios separately at time point $T = 399$.

Mean Absolute Error

As an additional analysis of the overall parameter recovery performance of the non-stationary DDM, we computed the median absolute error (MAE) between the true data generating parameter and the posterior mean for all DDM parameters and simulation scenarios separately. In the top row of [Figure A.15](#) we can see that the posterior estimates of the non-stationary DDM quickly approach the true data-generating parameter when the true parameter was constant over time. That said, there remains some error between the true and estimated parameter even after 400 time steps. This error is the largest in the drift rate parameter (≈ 0.15). We see similar recovery performance in the scenario, where the parameters were allowed to randomly fluctuate around a constant value (second row in [Figure A.15](#)). However, we observe a larger variability in the MAE. The third row depicts the MAE when data was simulated with the same model as we fitted to the data (i.e., the well-specified case). Once again, the MAE quickly decreases in the beginning and then flattens out. However, in this scenario, the MAE remains on a larger level than in the previous two scenarios. Also, the variability of the MAE between data fits is larger. This is not surprising because the estimation of non-stationary model parameters is more difficult than static or stationary variable parameters. The last row in [Figure A.15](#) shows how the parameter estimates of the non-stationary DDM react to sudden jumps in otherwise constant parameters. We observe that the MAE significantly increases when a jump occurred and then decreases again.

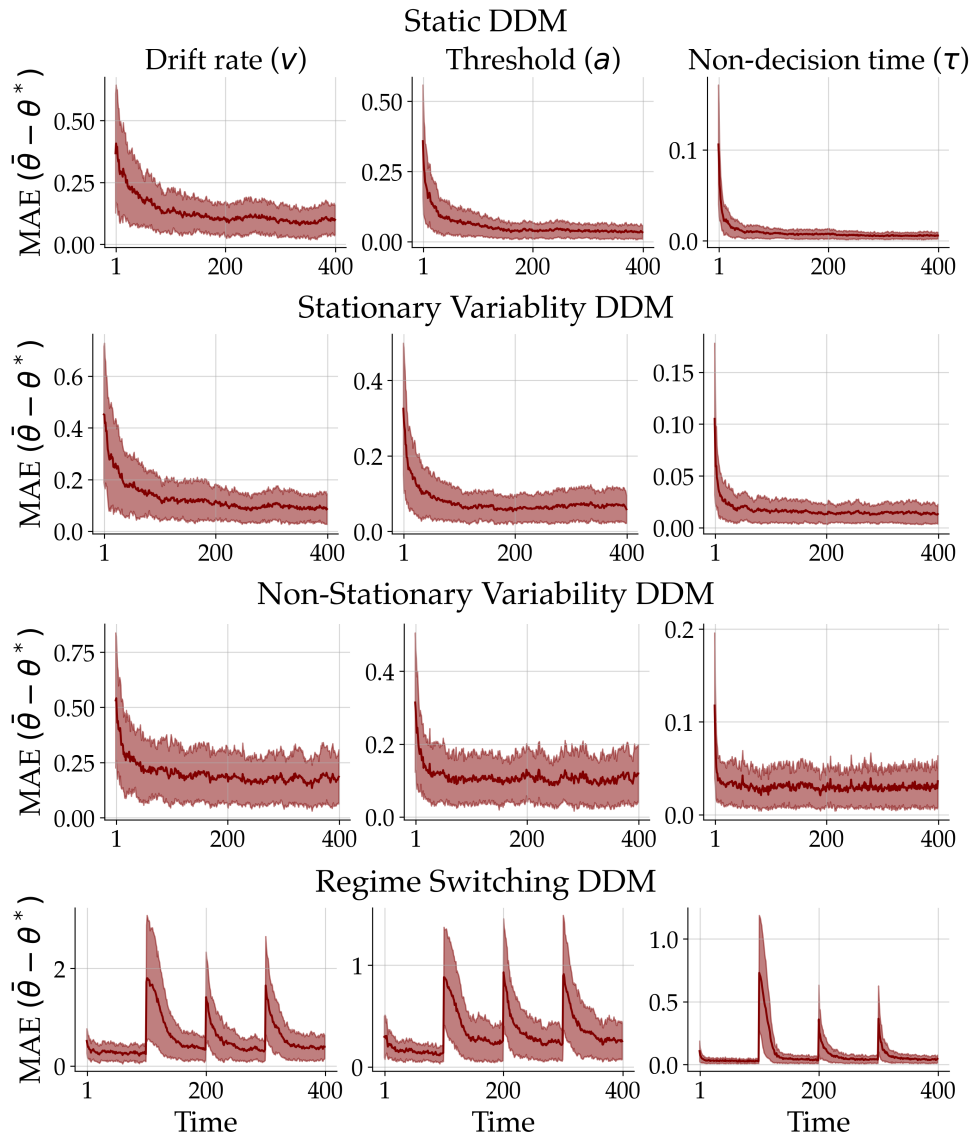


Figure A.15: Median absolute error (MAE) between the data generating parameters and the estimated posterior means aggregated across the 200 simulations over time for each DDM parameter (columns) and simulation scenario (rows) separately. The red shaded areas depict the median absolute deviation of the absolute errors.

Human Data Application: Random-Dot Motion

We fitted a non-stationary DDM with a Gaussian random walk transition model to each individual in the data set separately. The model implementation was the same as in the **Stan Benchmark** and the **Simulation Study**. For the training of our neural estimation method we used 75 epoch each consisting of 1000 iteration with a batch size of 16. After training we obtained 2000 posterior samples.

Human Data Application: Lexical Decision

Gaussian Process DDM

For the Gaussian Process transition model, we first create a $T \times T$ squared distance matrix with $T = 3200$. Based on this distance matrix we calculate the radial basis function kernel (cf. equation (6) in the main text) given the two parameters, amplitude σ and length-scale l , resulting in the covariance k for the multivariate normal distribution of the Gaussian Process:

$$\theta_{1:T} \sim \mathcal{MVN}(\mu_\theta, k)$$

where μ_θ is the mean parameter value. For these means we used the same priors we otherwise used for the starting values of the DDM parameters $(v_{0,i}, a_0, \tau_0)$. In the following we present a list of the priors used by the Gaussian Process DDM simulator to generate data for the simulation study and for training the neural networks. $\Gamma(a, b)$ refers to a Gamma distribution parameterized with shape a and scale b . The same prior distribution was used for all $i = 4$ drift rates $v_{0,i}$. $\mathcal{U}(a, b)$ stands for a continuous uniform distribution with a lower limit a and an upper limit b . l_j denotes the length-scale parameters of the GP transition model. The same prior distribution was used for all $j = 1, \dots, 6$ length-scale parameters governing the transitions of the DDM parameters. The amplitude parameter σ of the Gaussian kernel is usually highly correlates with the length-scale l . Thus, we fixed σ to sensible values for all low-level parameter transitions.

$$\begin{aligned} v_{0,i} &\sim \Gamma(2.5, \frac{1}{1.5}) \\ a_0 &\sim \Gamma(4.0, \frac{1}{3}) \\ \tau_0 &\sim \Gamma(1.5, \frac{1}{5}) \\ l_j &\sim \mathcal{U}(0.1, 10) \\ \sigma_{v_{1:4}} &= 0.15 \\ \sigma_a &= 0.1 \\ \sigma_\tau &= 0.05 \end{aligned}$$

Simulation-Based Calibration

We validate the computational faithfulness of our Bayesian inference algorithm using simulation-based calibration, a robust method for ensuring unbiased posterior distributions. The underlying principle is that an ensemble of posterior distributions should be indistinguishable from the prior distribution. To accomplish this, we carry out 2000 simulations with the dynamic DDM, each generating a separate data set. For each simulated data set, we fit the model and obtain 250 posterior samples. These posterior distributions collectively form an ensemble.

When we calculate rank statistics for the ensemble relative to the prior distribution then these should be uniformly distributed. To assess the uniformity at predefined time points, we utilize the empirical cumulative distribution function (ECDF) for each marginal rank distribution. Comparing it with a uniform ECDF allows us to gauge how the data is distributed. We further draw ECDF simultaneous bands using simulations from the uniform, providing an intuitive graphical test for uniformity. For clarity, [Figure A.16](#) presents the ECDF difference, providing a more dynamic range for the visualization. The red line (ECDF difference) should consistently fall within the gray shaded area (confidence band) across the entire range of fractional rank statistic values. In the majority of cases, this criterion is met for most parameters at all selected time points. Some slight deviations are observed for the threshold and non-decision parameters; however, these are typically small and not a major cause for concern.

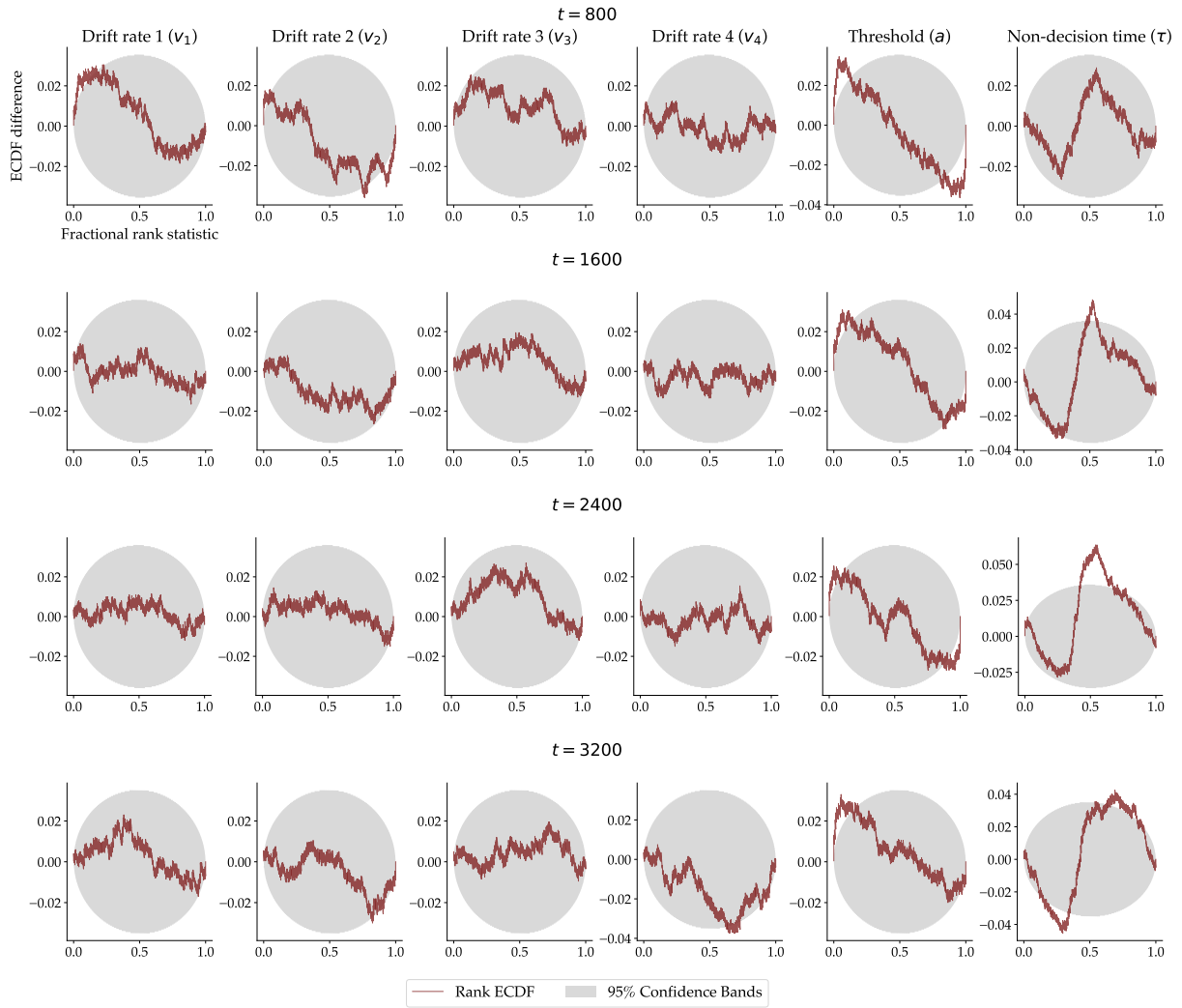


Figure A.16: **ECDF difference plot** 95% simultaneous confidence bands (gray) for the empirical cumulative distribution function (ECDF; red) for all 6 parameters at four selected time points (800, 1600, 2500, 3200) separately.

Parameter Recovery Study

A simulation study was performed to probe the dynamic DDM's capability of recovering data-generating parameter dynamics. To this end, we simulated 1000 data sets with the dynamic DDM and fit it to these data. The following figures show posterior predictions of 3 randomly selected simulated data sets and the comparison between the inferred and the true data-generating low-level parameter dynamics. The parameter recovery performance across all 1000 data sets over all 3200 time points for all 6 model parameters can be inspected as a GIF in our GitHub repository (<https://github.com/bayesflow-org/Neural-Superstatistics>)

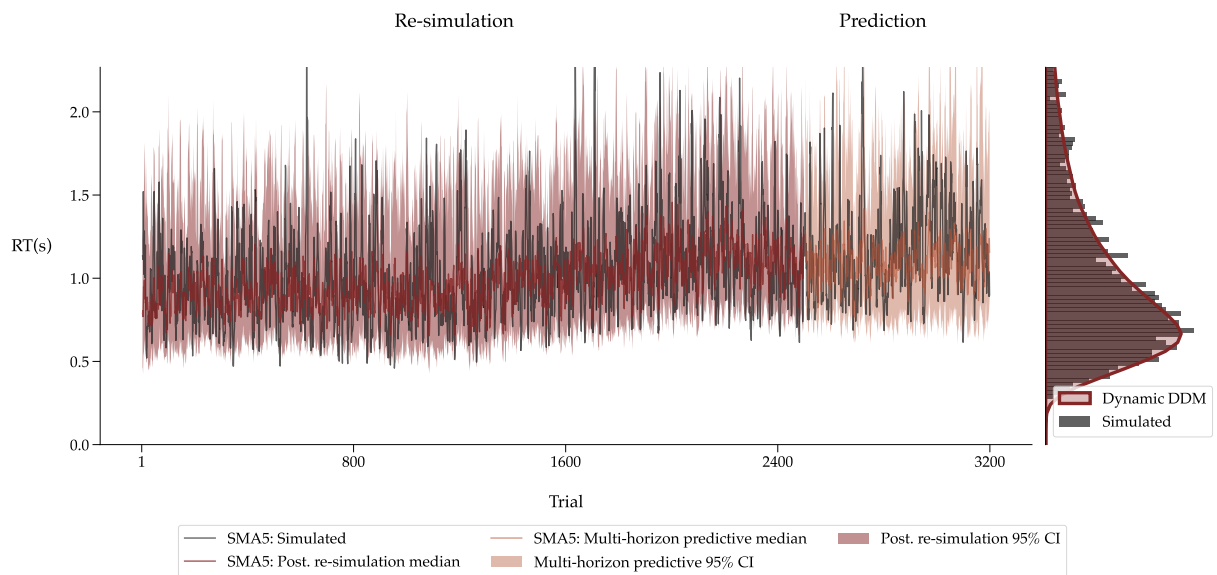


Figure A.17: **Left panel** The simulated RT time series is shown in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. **Right panel** The raw simulated RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM are shown as kernel density estimates (KDEs) in red.

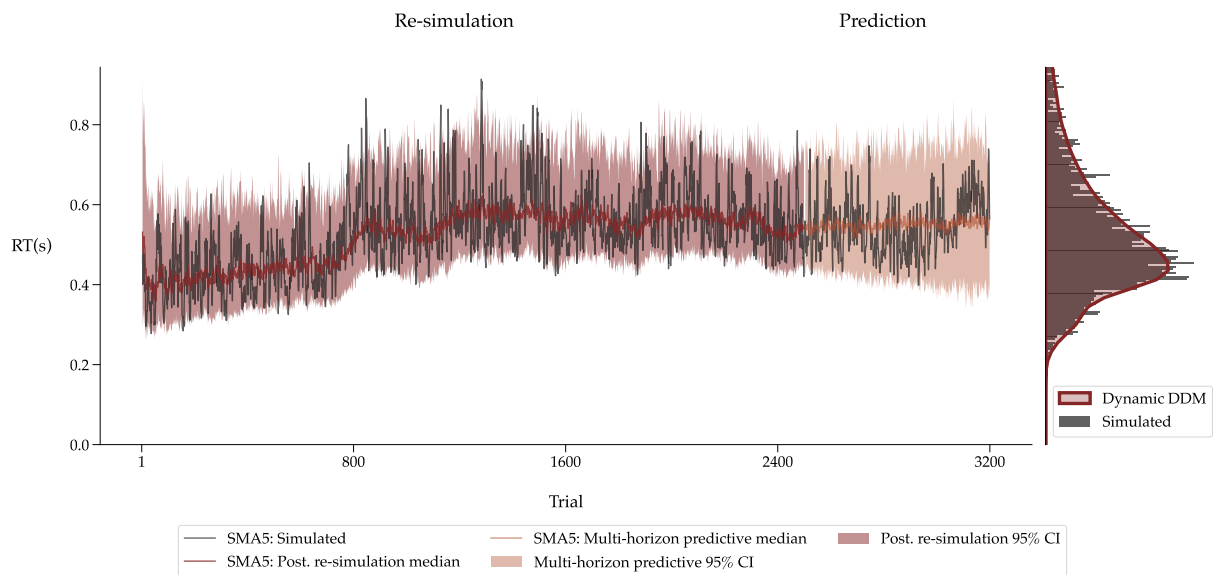


Figure A.18: **Left panel** The simulated RT time series is shown in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. **Right panel** The raw simulated RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM are shown as kernel density estimates (KDEs) in red.

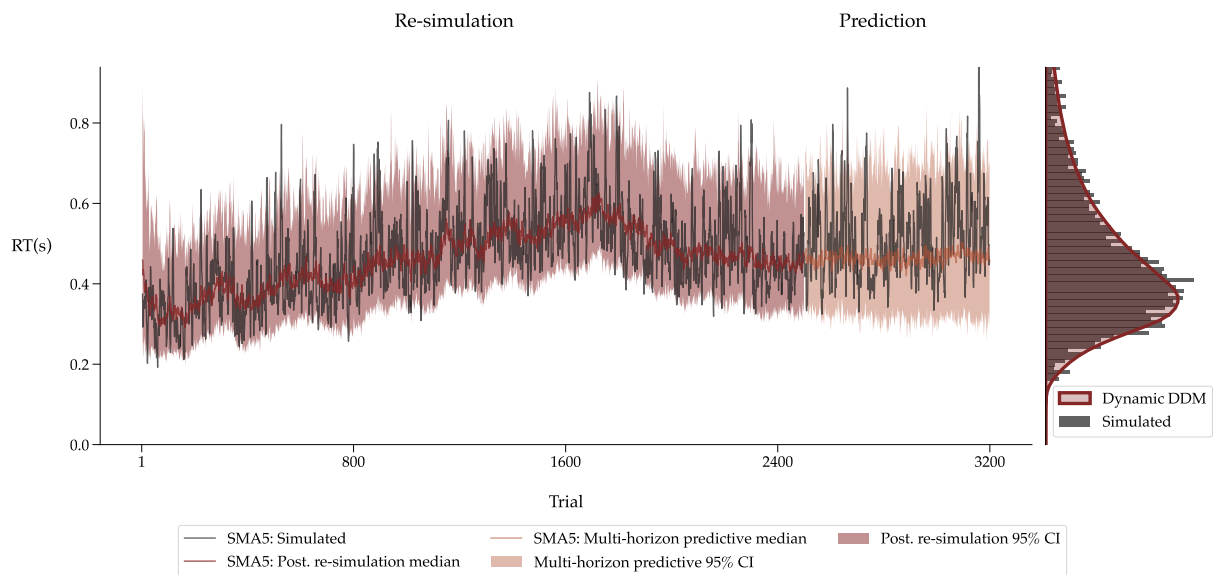


Figure A.19: **Left panel** The simulated RT time series is shown in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. **Right panel** The raw simulated RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM are shown as kernel density estimates (KDEs) in red.

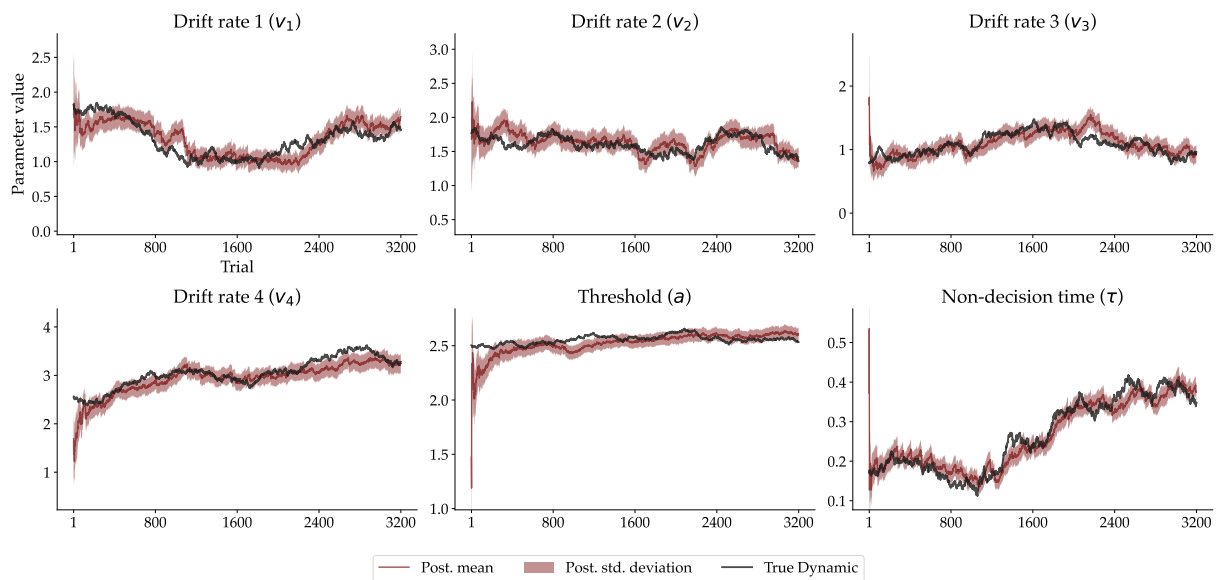


Figure A.20: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ in red. The true data generating parameter dynamic in black.

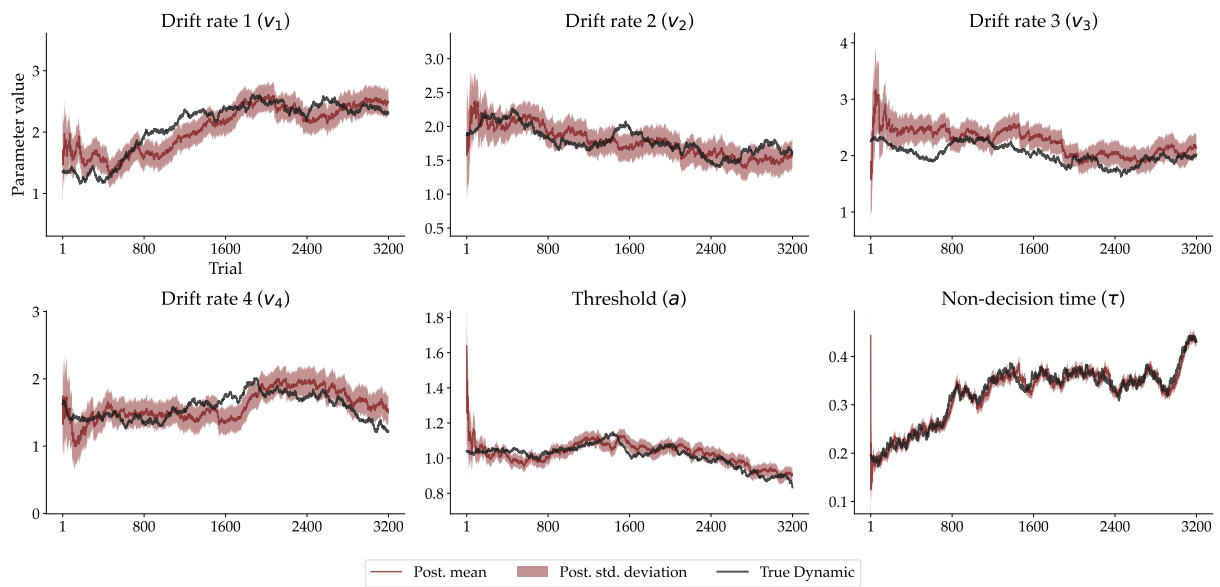


Figure A.21: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ in red. The true data generating parameter dynamic in black.

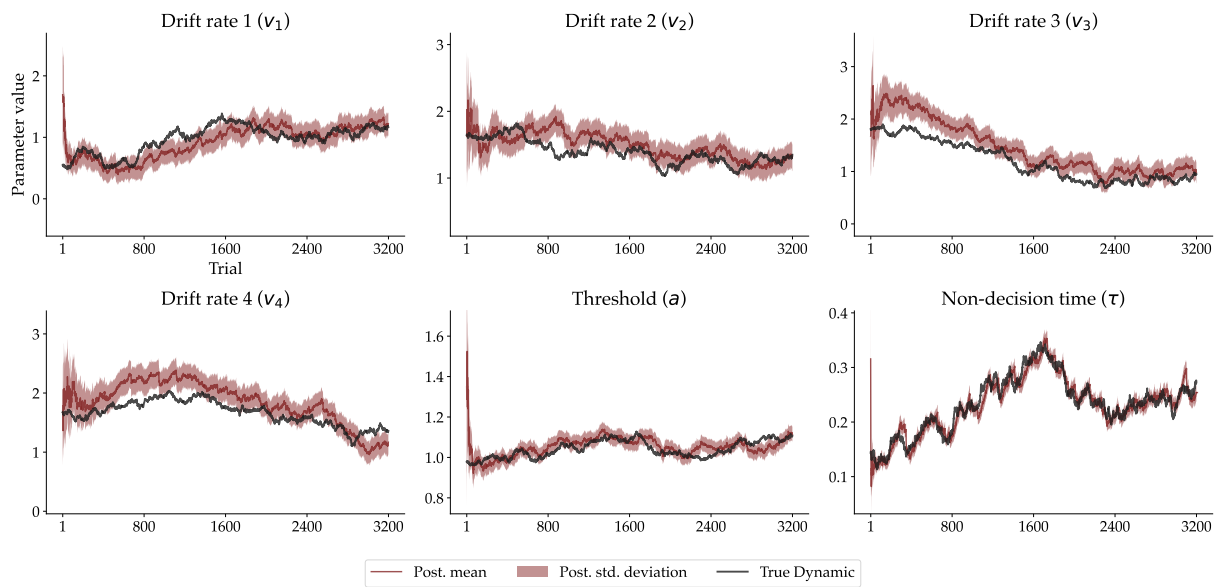


Figure A.22: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ in red. The true data generating parameter dynamic in black.

Individual Model Fits and Predictions

In the following, we show the fit and multi-horizon predictions of the dynamic DDM on the individual data of the remaining 10 participants not shown in the main text.

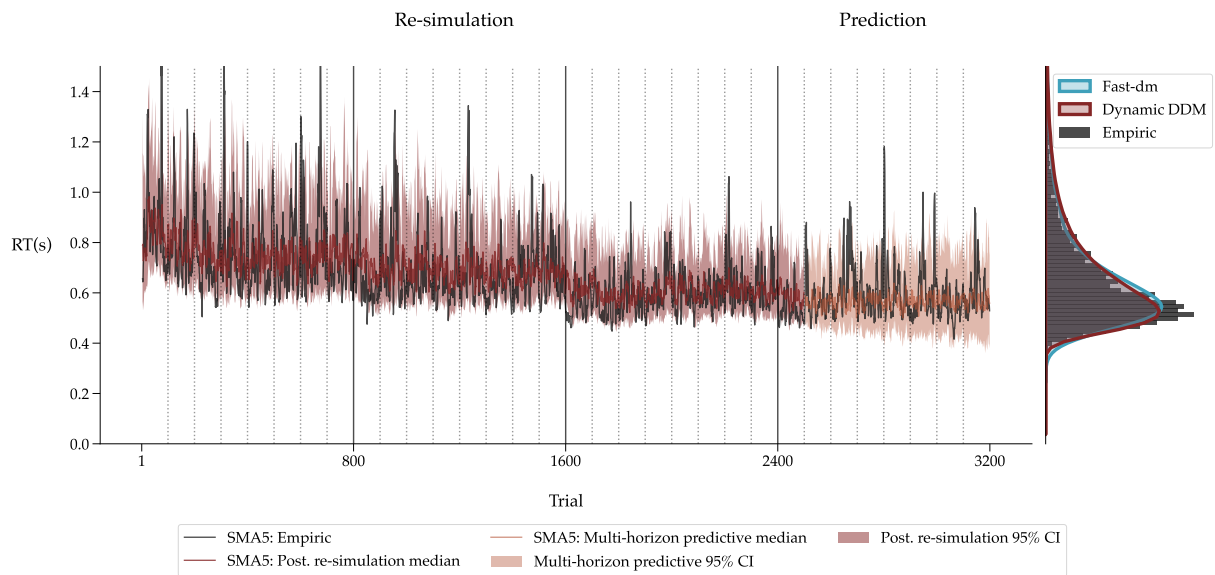


Figure A.23: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

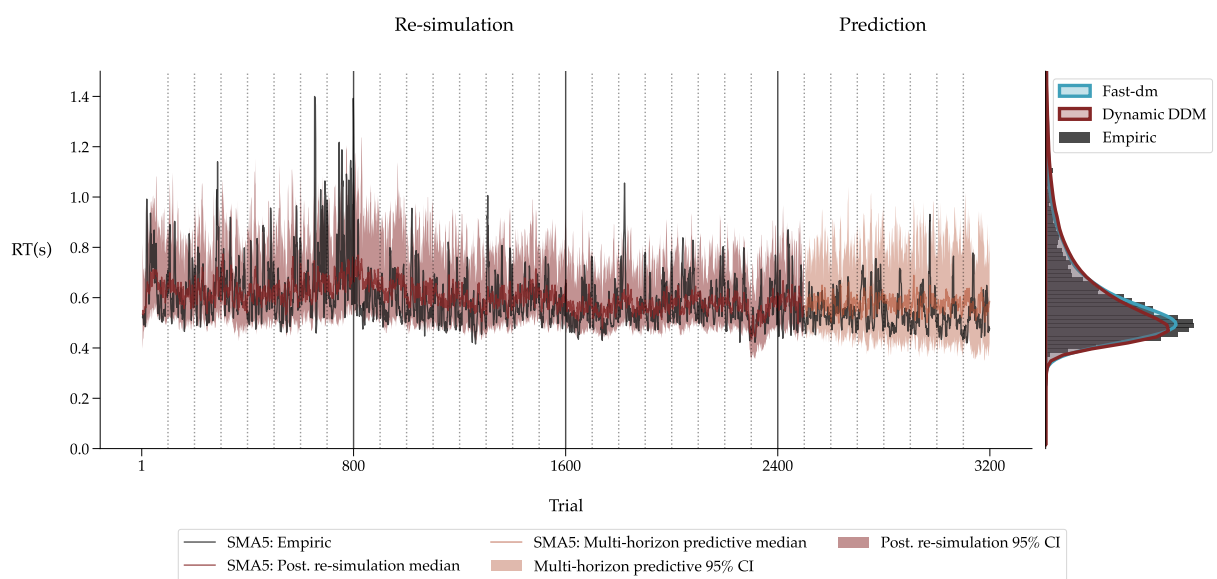


Figure A.24: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

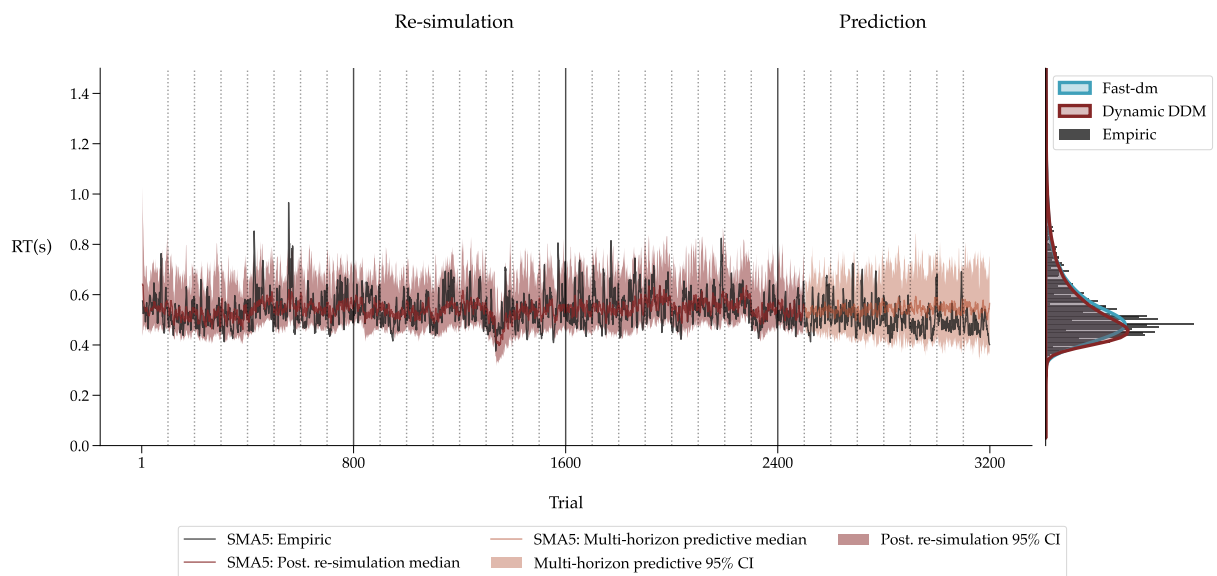


Figure A.25: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

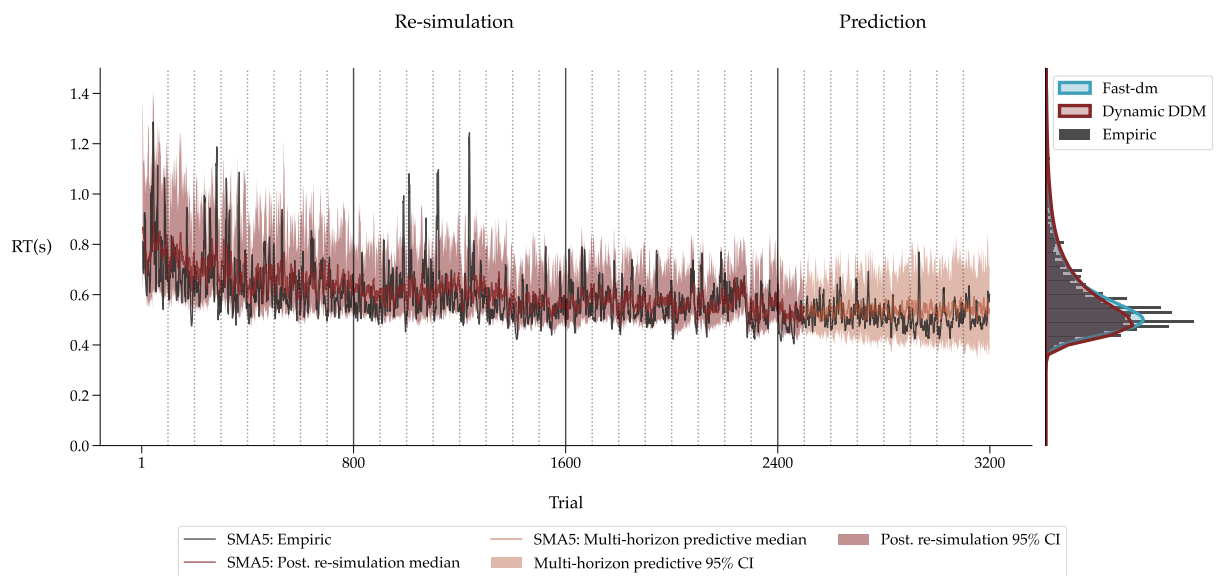


Figure A.26: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models’ multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

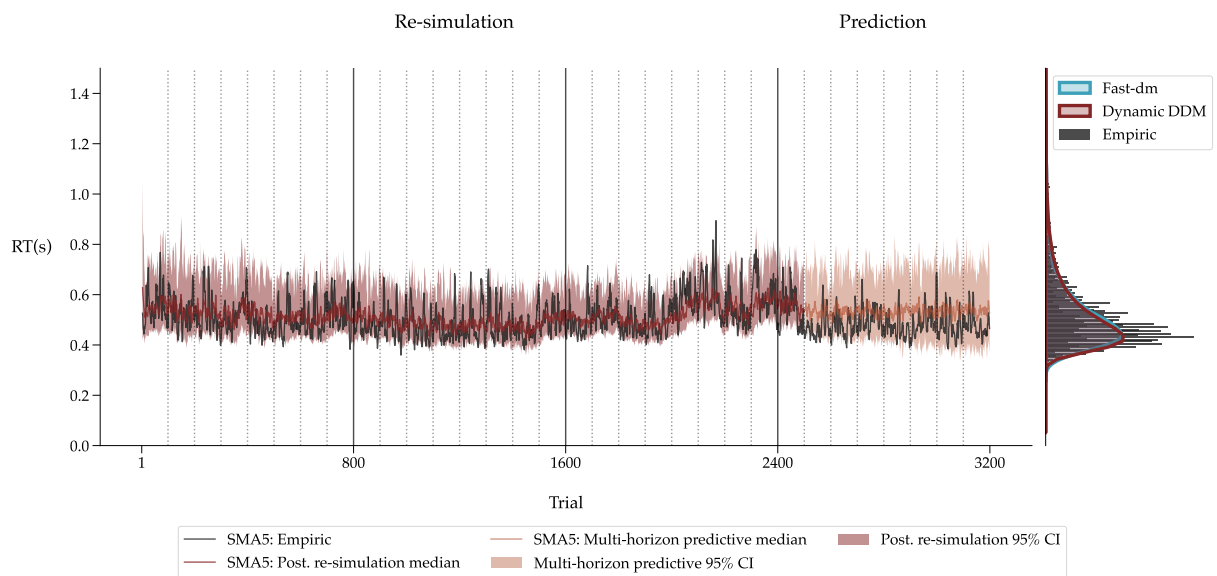


Figure A.27: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

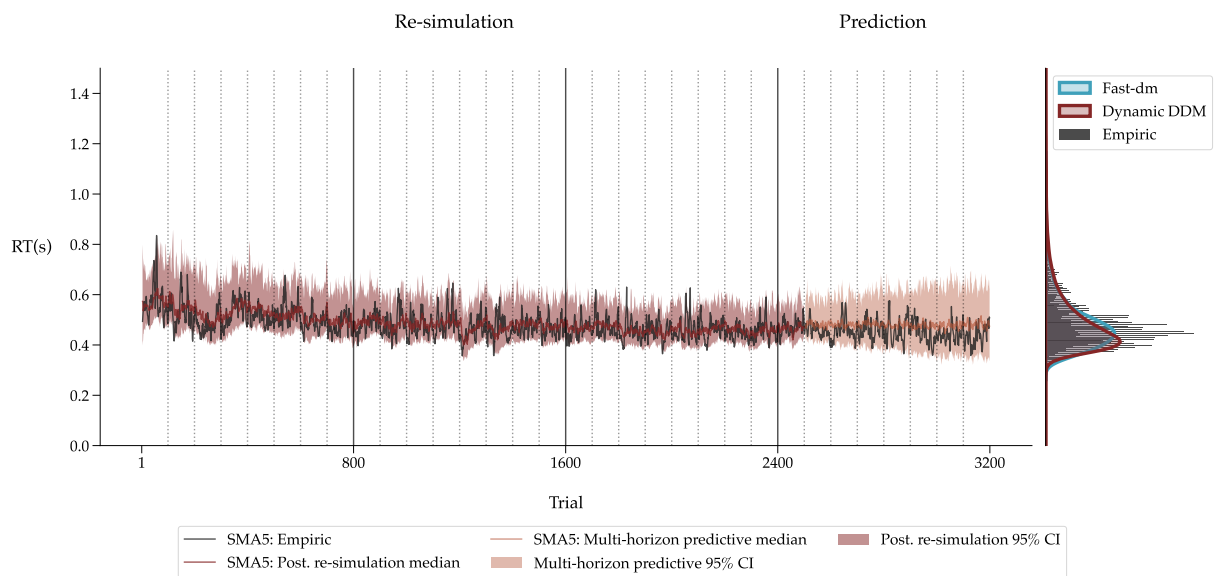


Figure A.28: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

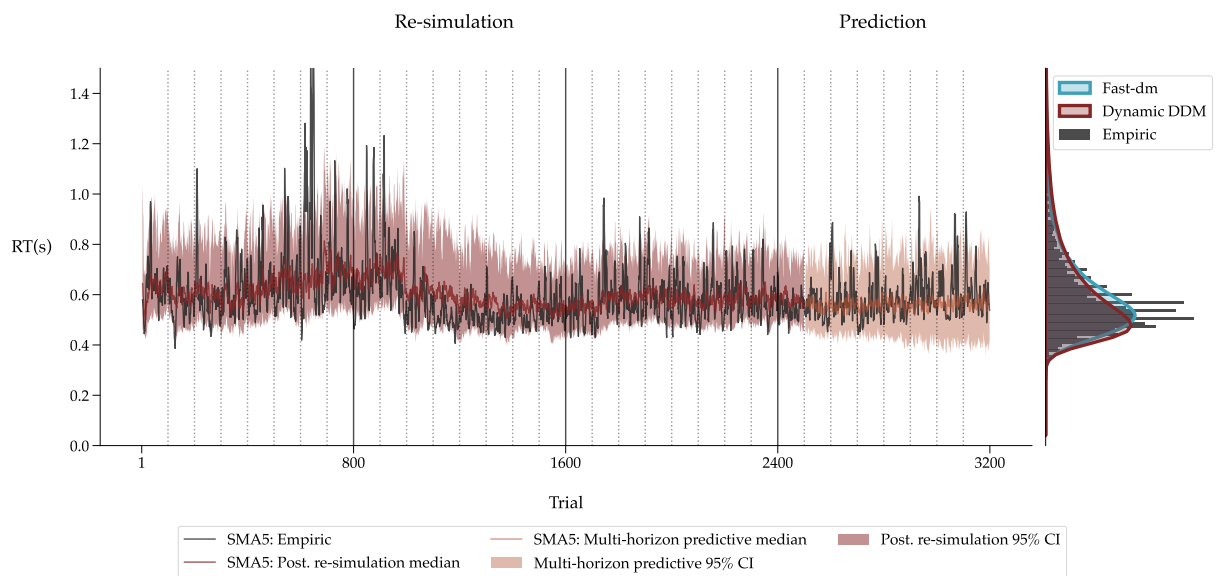


Figure A.29: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

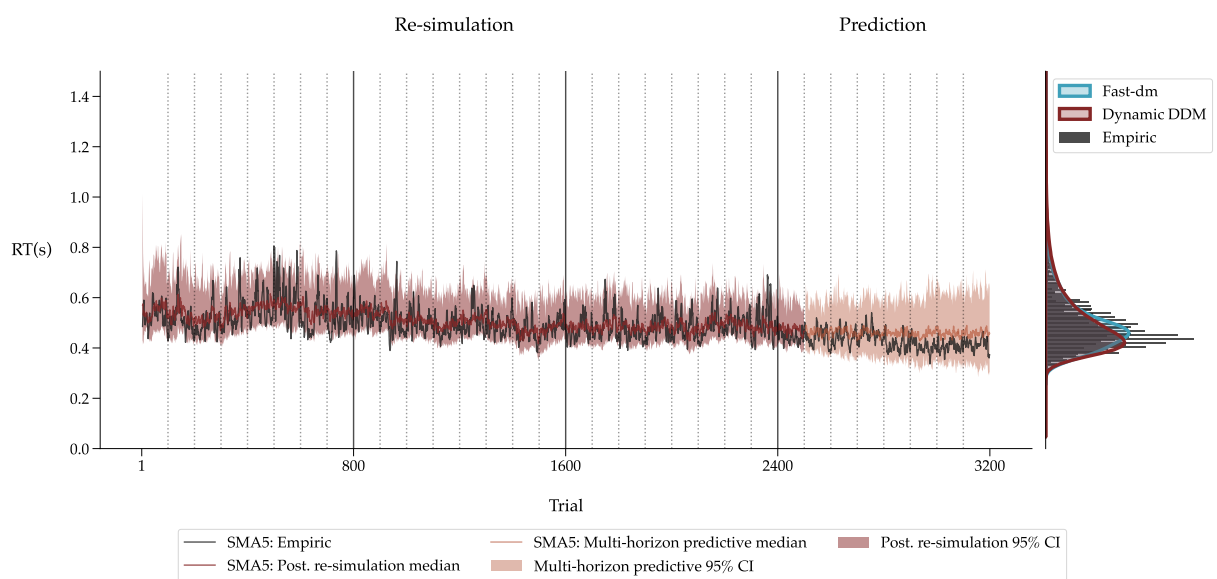


Figure A.30: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

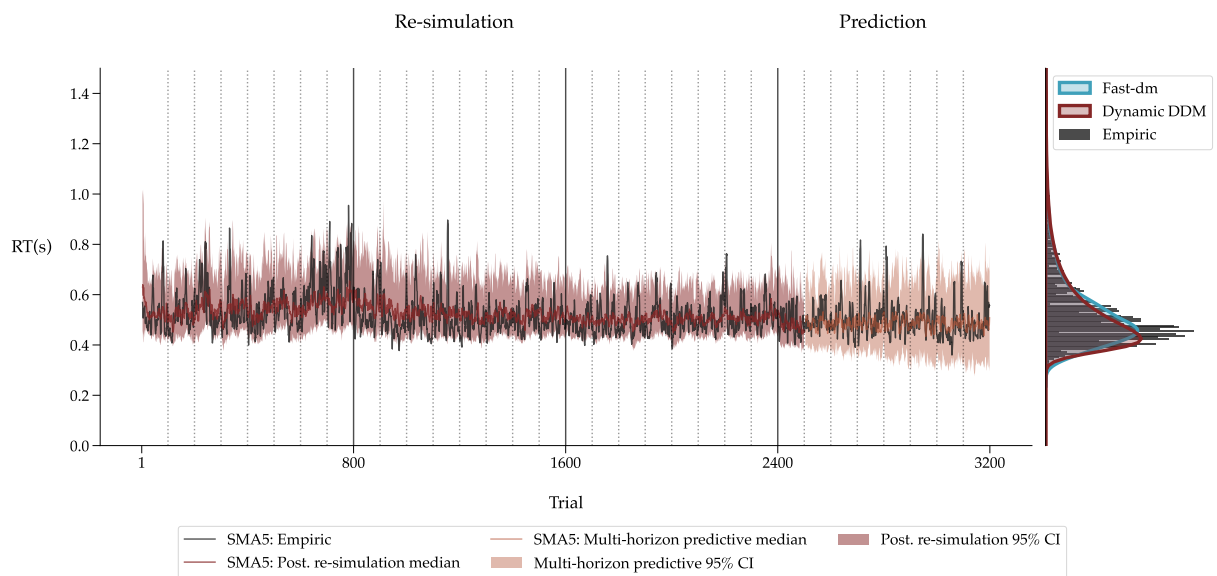


Figure A.31: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

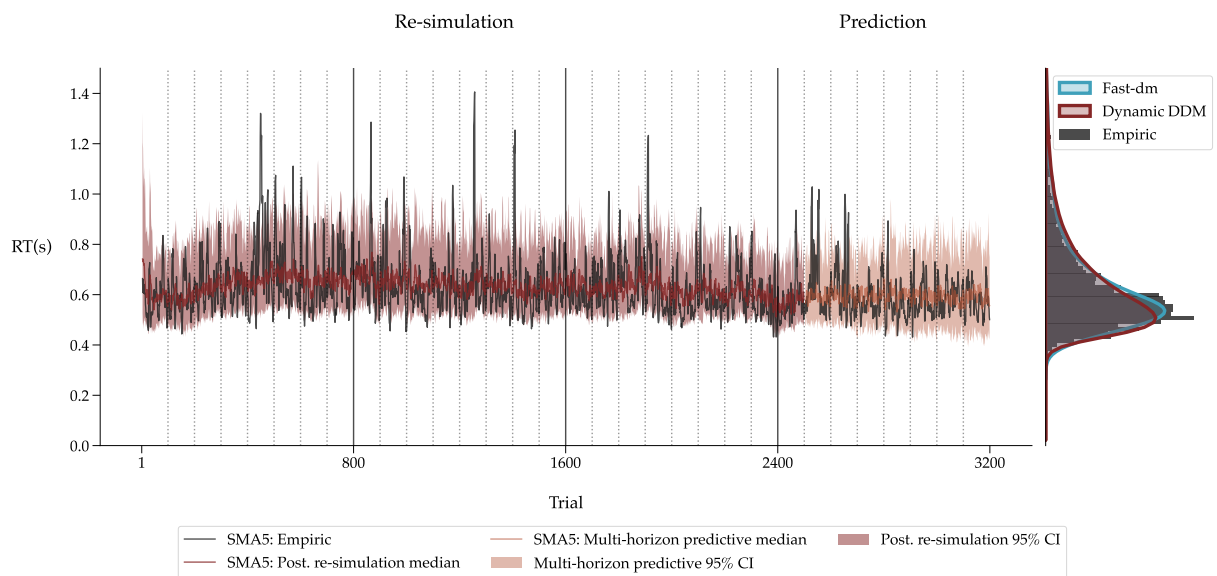


Figure A.32: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

Individual Parameter Dynamics

In the following, we show the inferred parameter dynamics of the remaining 10 participants not shown in the main text.

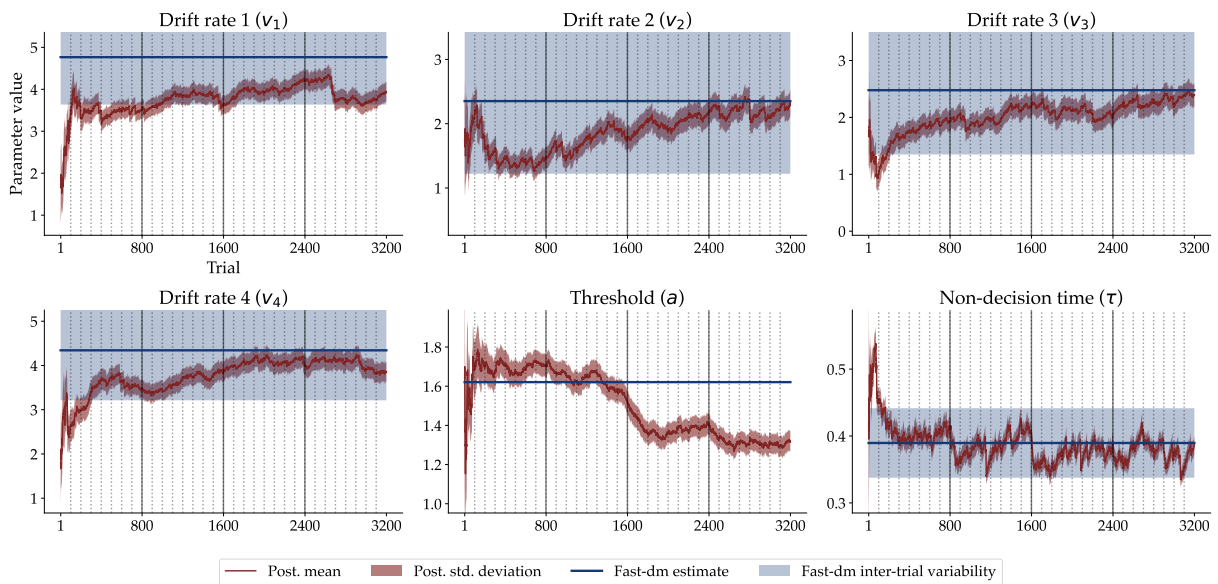


Figure A.33: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

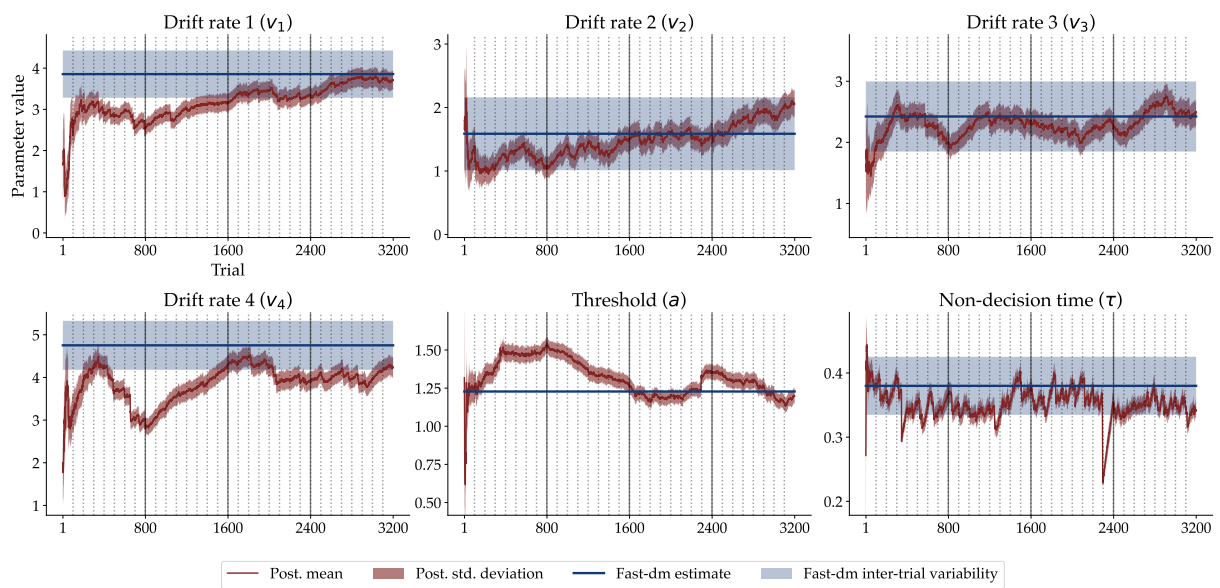


Figure A.34: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

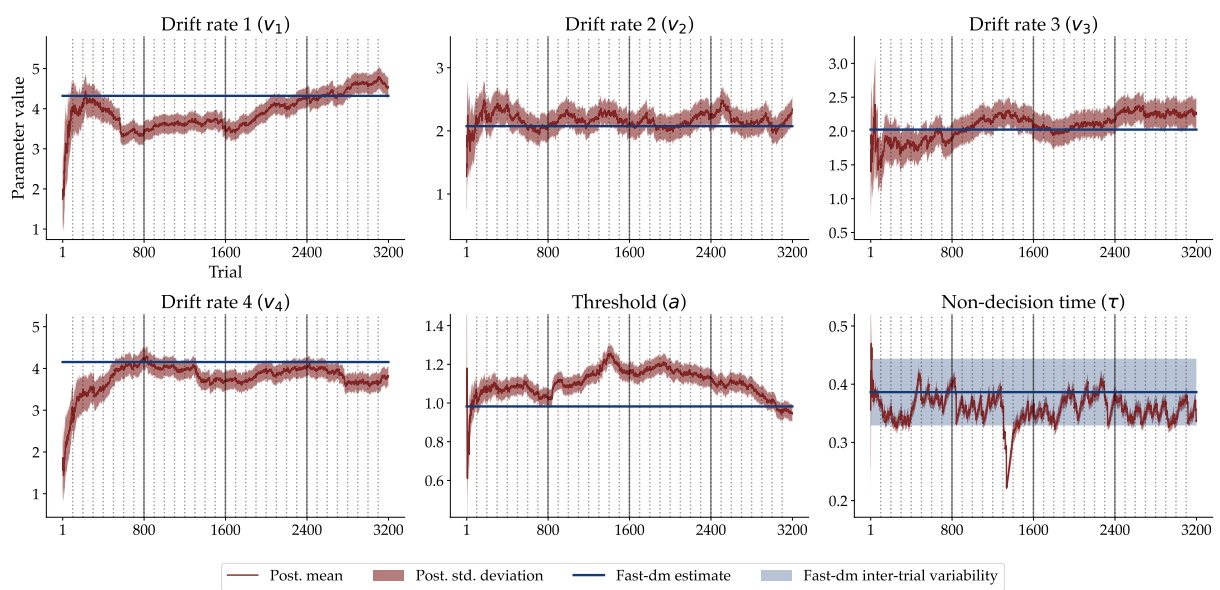


Figure A.35: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

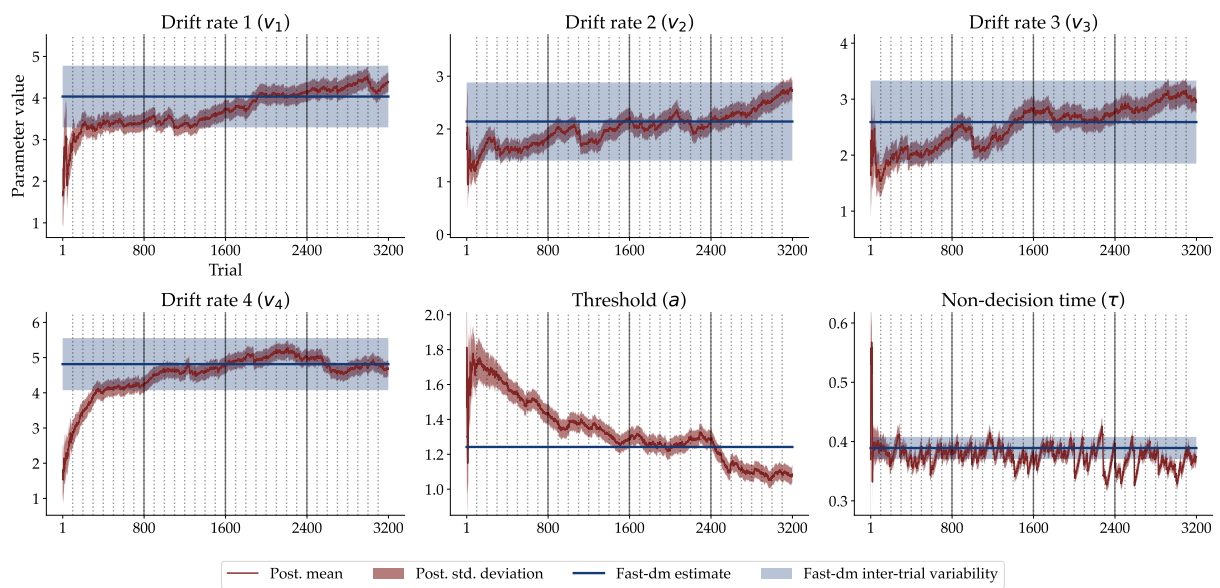


Figure A.36: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

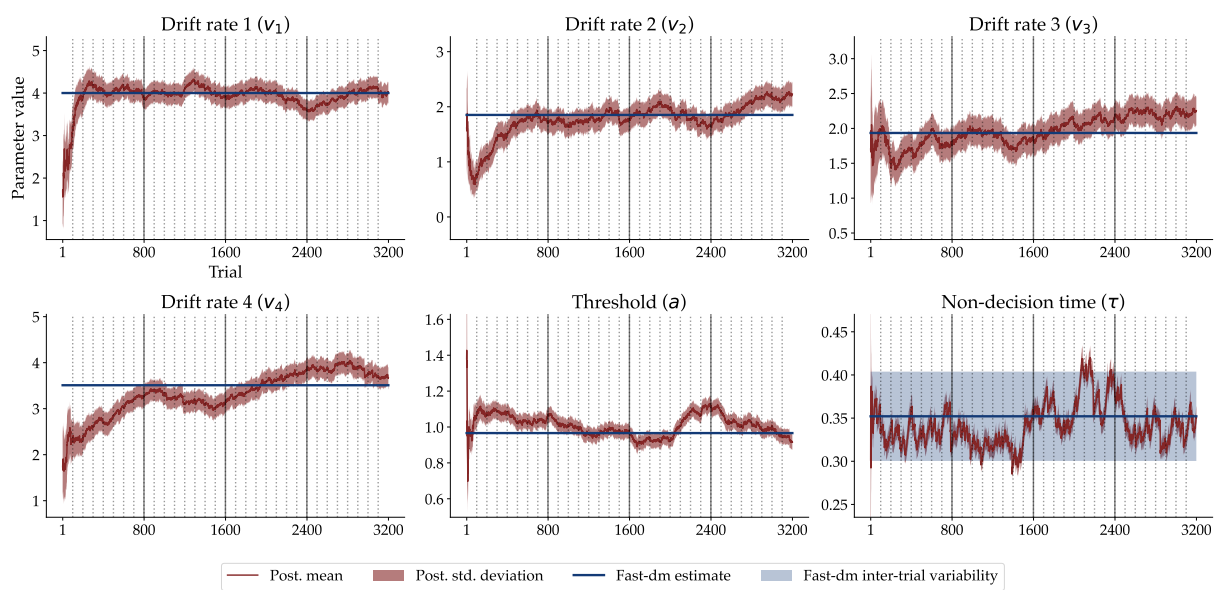


Figure A.37: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

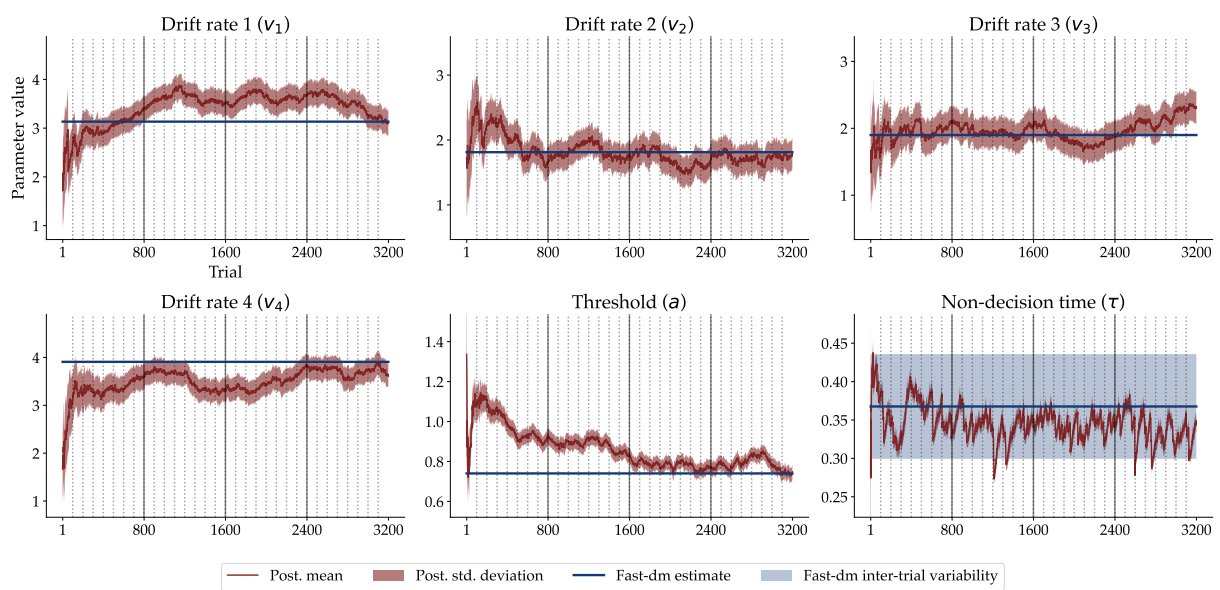


Figure A.38: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

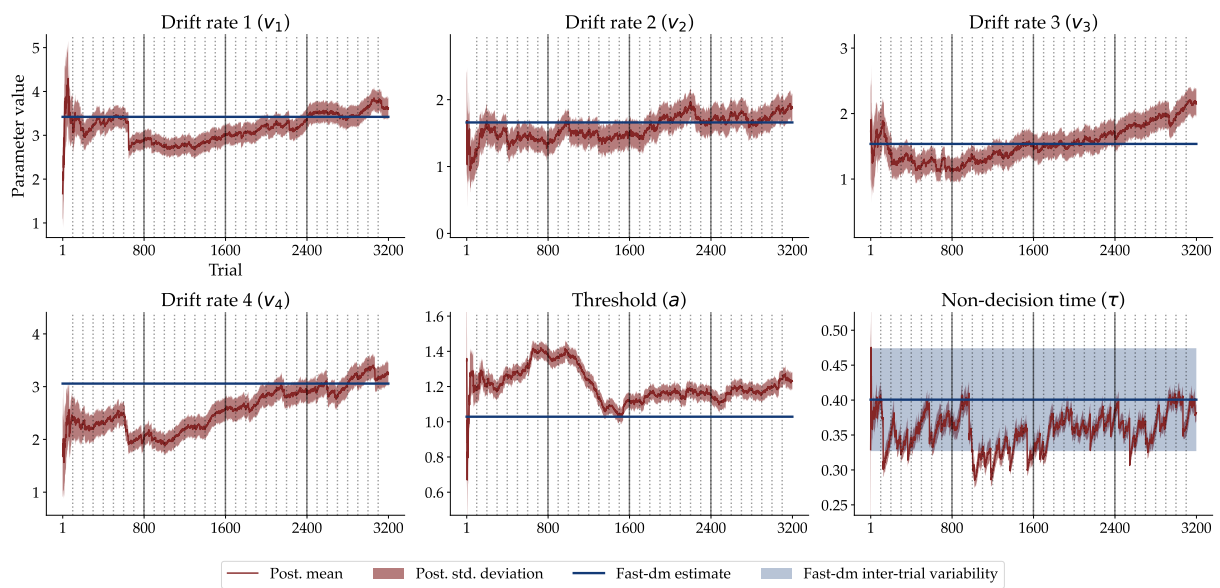


Figure A.39: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

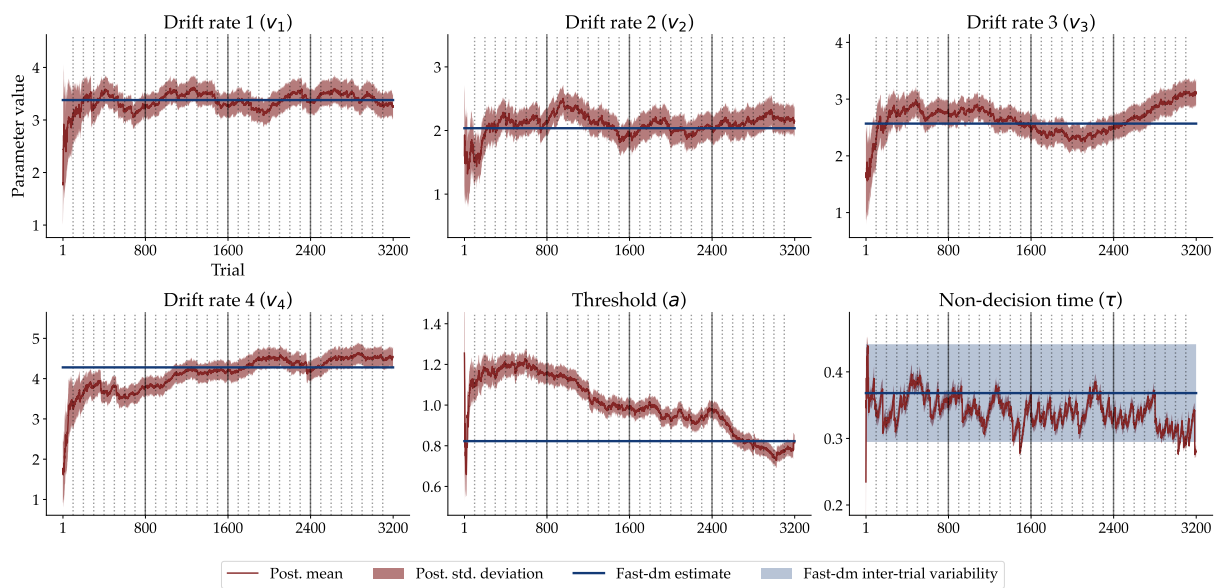


Figure A.40: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

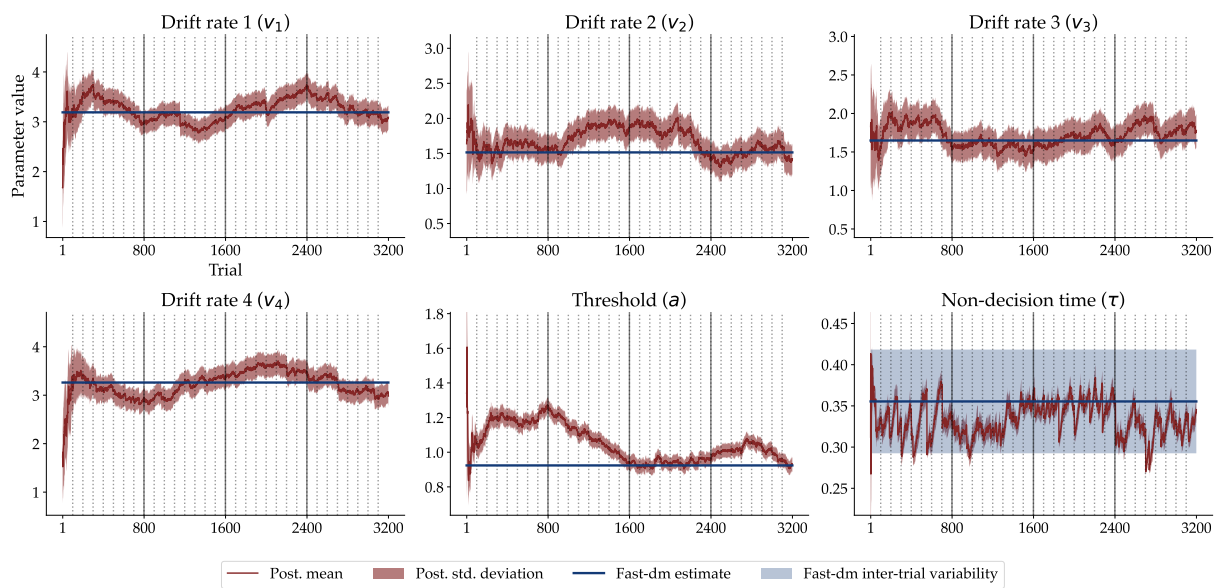


Figure A.41: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

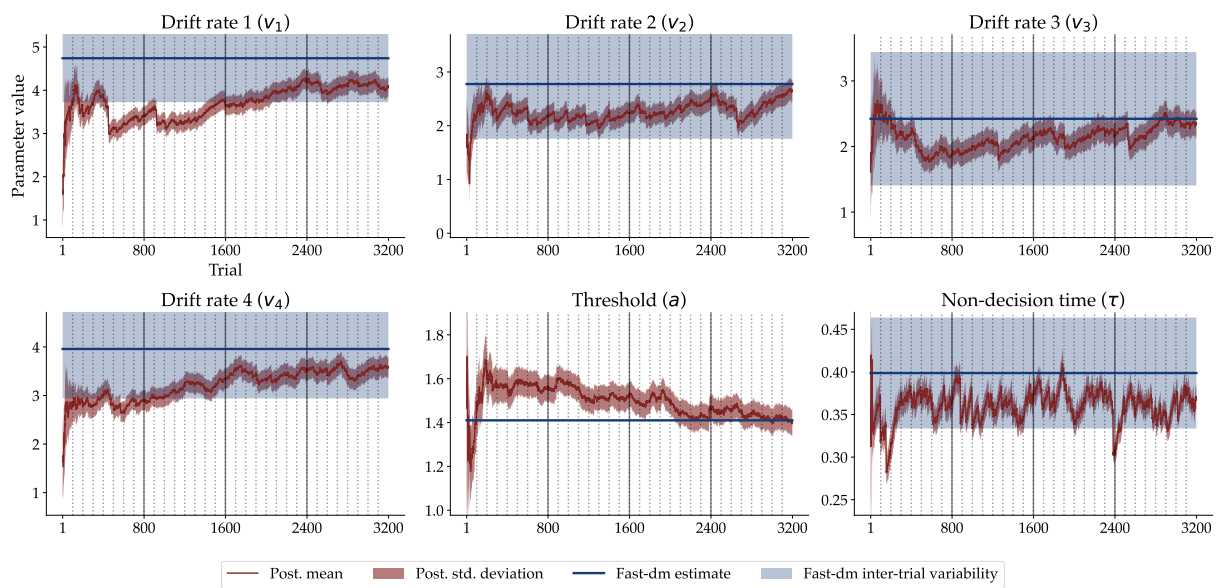


Figure A.42: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

Average Parameter Dynamics

Figure A.43 shows the parameter dynamic averaged across all participants.

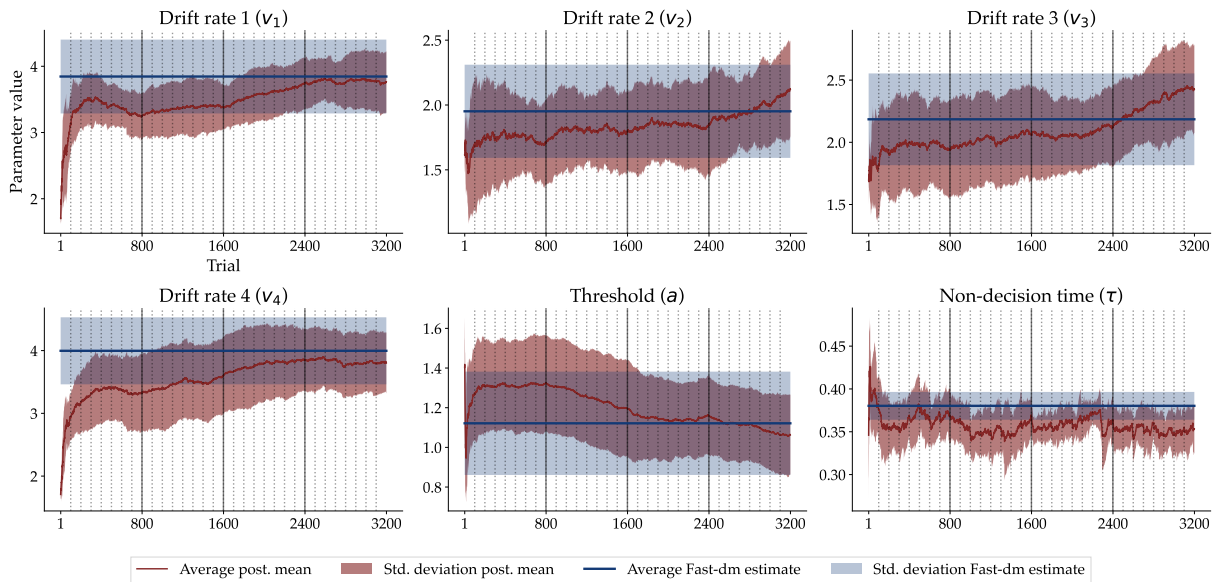


Figure A.43: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of averaged across all participant in solid red lines. The shaded red areas correspond to the ± 1 standard deviation of the posterior means of all individuals. The point estimates of the static DDM parameters averaged across all participants and the corresponding standard deviations are shown in solid blue lines and shaded blue areas, respectively.

Gaussian Random Walk Transition Model

We wanted to test if our neural estimation method can also estimate dynamic models with a simpler high-level transition model than a Gaussian process (GP). To this end, we fit a dynamic DDM with a Gaussian random walk as a transition model to the empirical data set described in the **Human data application** section:

$$\theta_t = T(\theta_{t-1}, \eta, z_t) = \theta_{t-1} + \eta z_t \quad \text{with} \quad z_t \sim \mathcal{N}(0, 1)$$

We use a Beta prior distribution parameterized with α and β for the standard deviations η_j of the Gaussian random walk transition model. The same prior distribution is used for all $j = 6$ low-level parameter transitions:

$$\eta_j \sim \text{Beta}(1, 25)$$

We trained the same neural network architecture as described in the main text for 50 epochs, 1000 batches per epoch, and a batch size of 8. The following figures show the results from simulation-based calibration (SBC), the model fit and inferred parameter dynamics for the same exemplar participant shown in the main text. Additionally, we depict the estimated parameter dynamics averaged across all individuals for comparison. These results are very similar to those obtained with the GP-DDM, which uses a Gaussian process as a transition model. However, the model with the Gaussian process transition model produces sharper predictions on unseen data. Note, that the dynamics implied by the random walk transition model are less sharper (i.e., contain more uncertainty) than those implied by the GP transition model.

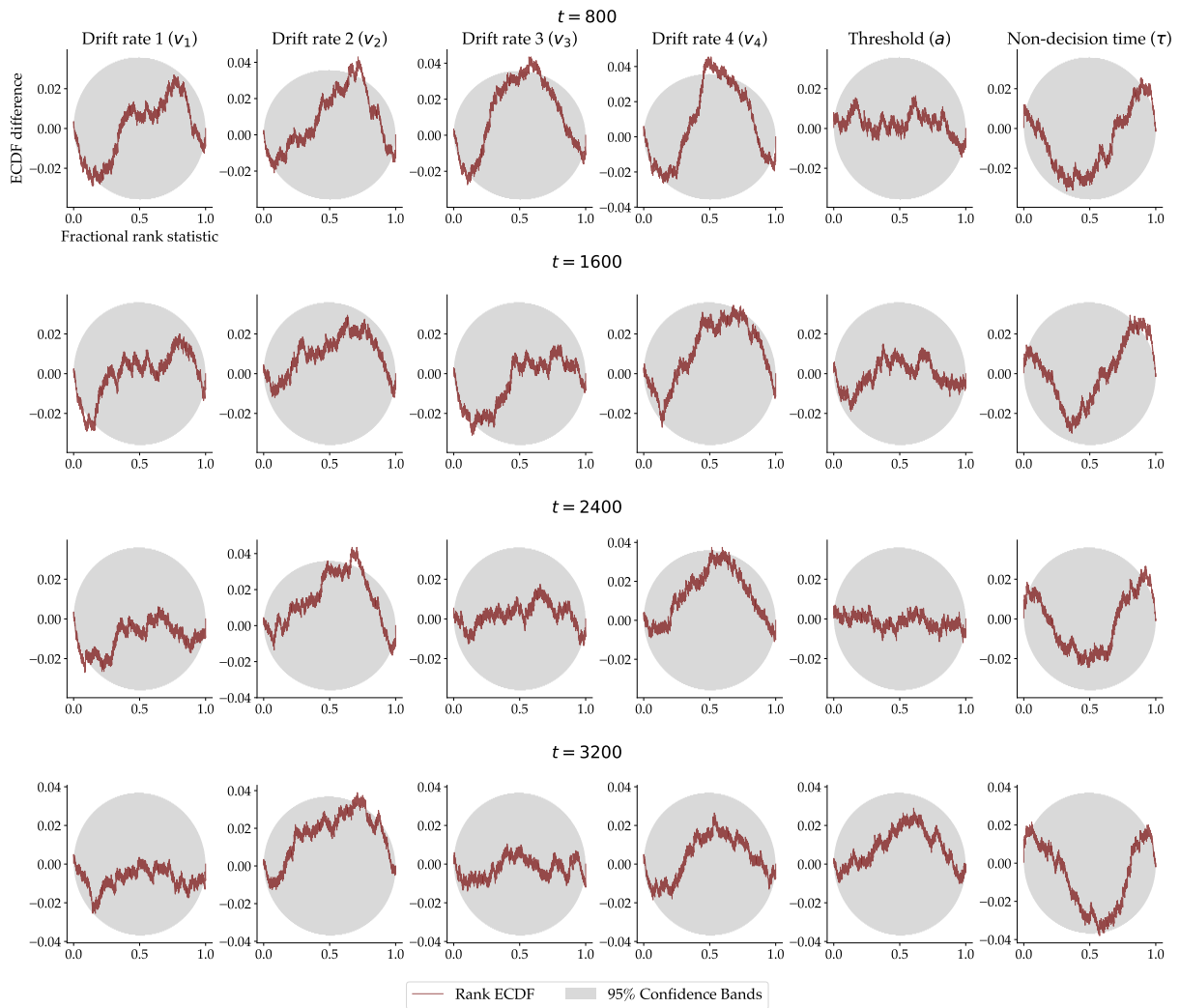


Figure A.44: **ECDF difference plot** 95% simultaneous confidence bands (gray) for the empirical cumulative distribution function (ECDF; red) for all 6 parameters at four selected time points (800, 1600, 2500, 3200) separately. We used the same settings as for the GP-DDM analysis.

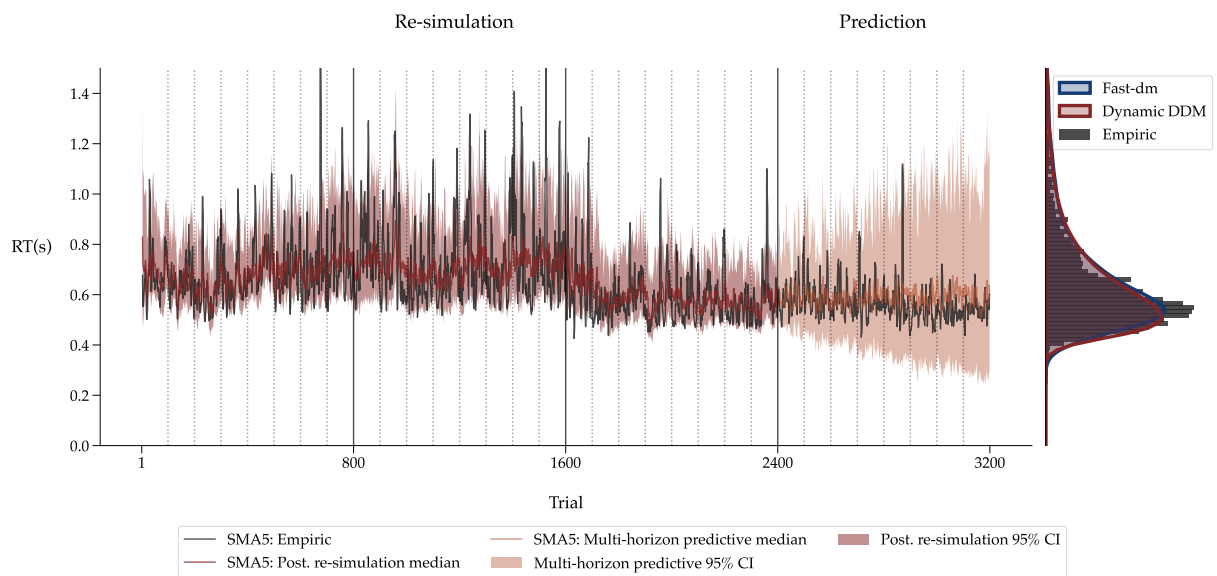


Figure A.45: **Left panel** The empirical RT time series of a single individual in black. From trial 1 to 2500, the median posterior re-simulation (aka *retrodictive check*) using the dynamic DDM is shown in red. The models' multi-horizon prediction is depicted for the remaining trials in orange. The shaded areas for the posterior re-simulation and prediction correspond to the 95% credibility interval. All the time series were smoothed via a simple moving average (SMA) with a period of 5. The dotted vertical lines indicate the end of an experimental block, and the solid vertical lines the end of an experimental session. **Right panel** The raw RT distribution is plotted as a histogram in black. The re-simulated RT distributions from the dynamic DDM and reference re-simulations from the static DDM using `Fast-dm` are shown as kernel density estimates (KDEs) in red and blue, respectively.

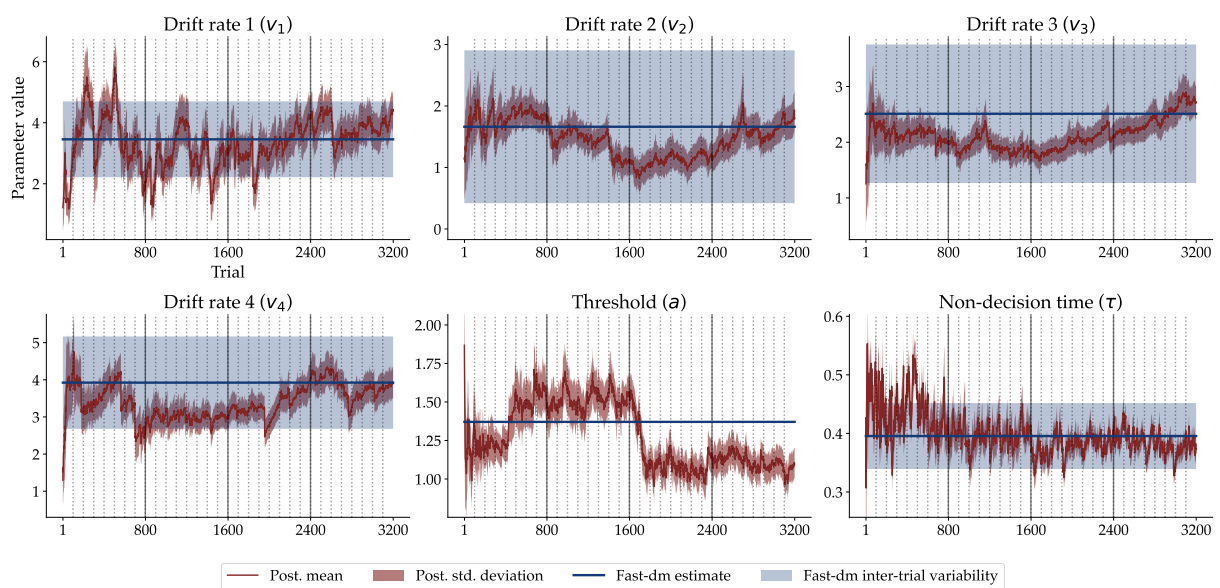


Figure A.46: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of an individual participant. The point estimates of the static DDM parameters and the corresponding inter-trial variabilities are shown in solid blue lines and blue shaded areas, respectively.

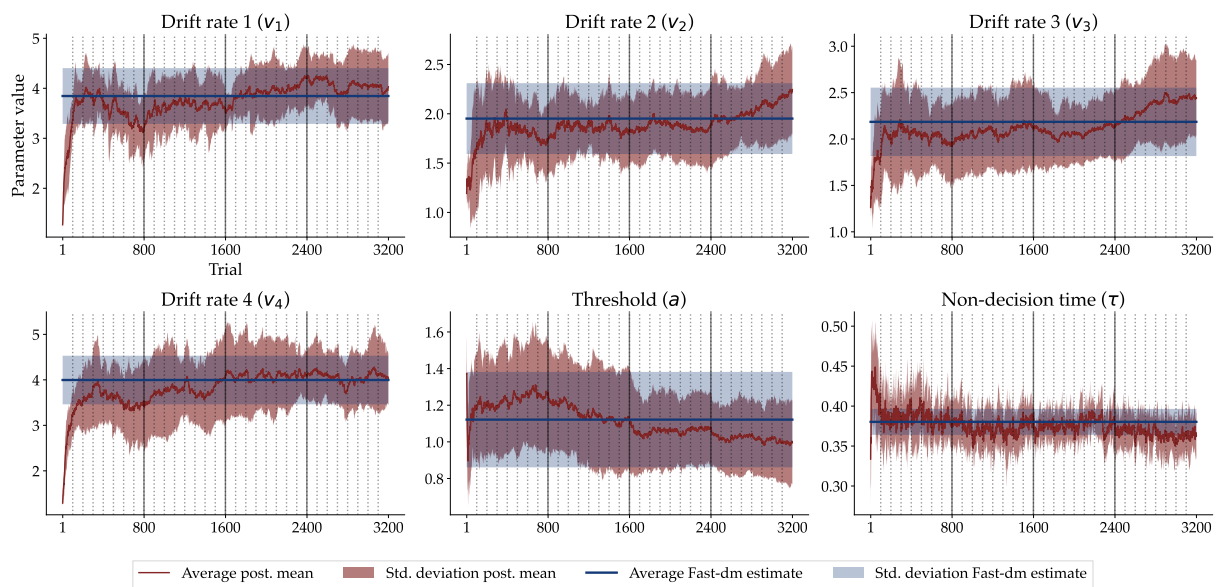


Figure A.47: The trial-wise posterior mean and ± 1 standard deviation for all six parameters, namely the four drift rates $v_1 - v_4$ (one for each experimental condition), the threshold a , and the non-decision time τ of averaged across all participant in solid red lines. The shaded red areas correspond to the ± 1 standard deviation of the posterior means of all individuals. The point estimates of the static DDM parameters averaged across all participants and the corresponding standard deviations are shown in solid blue lines and shaded blue areas, respectively.

APPENDIX C1 - MANUSCRIPT III

Manuscript III: Validation and Comparison of Non-Stationary Cognitive Models: A Diffusion Model Application

VALIDATION AND COMPARISON OF NON-STATIONARY COGNITIVE MODELS: A DIFFUSION MODEL APPLICATION

Lukas Schumacher*
Institute of Psychology
Heidelberg University

Martin Schnuerch
Institute of Psychology
University of Mannheim

Andreas Voss
Institute of Psychology
Heidelberg University

Stefan T. Radev
Department of Cognitive Science
Rensselaer Polytechnic Institute

ABSTRACT

Cognitive processes undergo various fluctuations and transient states across different temporal scales. Superstatistics are emerging as a flexible framework for incorporating such *non-stationary dynamics* into existing cognitive model classes. In this work, we provide the first experimental validation of superstatistics and formal comparison of four non-stationary diffusion decision models in a specifically designed perceptual decision-making task. Task difficulty and speed-accuracy trade-off were systematically manipulated to induce expected changes in model parameters. To validate our models, we assess whether the inferred parameter trajectories align with the patterns and sequences of the experimental manipulations. To address computational challenges, we present novel deep learning techniques for amortized Bayesian estimation and comparison of models with time-varying parameters. Our findings indicate that transition models incorporating both gradual and abrupt parameter shifts provide the best fit to the empirical data. Moreover, we find that the inferred parameter trajectories closely mirror the sequence of experimental manipulations. Posterior re-simulations further underscore the ability of the models to faithfully reproduce critical data patterns. Accordingly, our results suggest that the inferred non-stationary dynamics may reflect actual changes in the targeted psychological constructs. We argue that our initial experimental validation paves the way for the widespread application of superstatistics in cognitive modeling and beyond.

Introduction

The human brain operates in a perpetual state of activity, whether it is focused on a particular task or wandering in the inner world of thoughts. This activity reflects the non-stationary nature of neuronal dynamics, which are characterized by a complex interplay between transient, evoked states, and ongoing spontaneous fluctuations (Galadí et al., 2021; Melanson et al., 2017). The complex cognitive processes that emerge from this neuronal activity also tend to exhibit non-stationary dynamics (Craigmile et al., 2010; Sebastian Castro-Alvarez & Tendeiro, 2023; Van Orden et al., 2003; Wagenmakers et al., 2004). In other words, proverbial cognitive processes, such as attention, memory, and decision-making, are not constant over time, but instead undergo fluctuations, shifts, and alterations in their functions.

Lapses of attention are a canonical cause of such non-stationary dynamics. Even when actively engaged in a task, our focus can drift or momentarily falter (Weissman et al., 2006). Moreover, our capacity to sustain attention and concentrate may vary, influenced by factors such as fatigue, motivation, and external distractions (Esterman & Rothlein, 2019; Ratcliff & Van Dongen, 2011; Walsh et al., 2017). These fluctuations can have a significant impact on our cognitive functioning, but they are often overlooked or simplified in traditional models of cognition. And while these often assume cognitive processes to be stable and time-invariant, there has been a growing recognition that traditional models do not fully capture the complexity and variability of real-world cognition (Beer, 2023; Cochrane et al., 2023; Evans & Brown, 2017; Gunawan et al., 2022; Kucharský et al., 2021; Li et al., 2023; Schumacher et al., 2023). Common approaches to address variability in the components of cognitive models can be broadly classified into four categories: *stationary variability*, *trial binning*, *regression approach*, and *frontend-backend* models.

The first approach assumes random fluctuations around a stable mean, referred to as stationary variability (see [Figure 1A](#)). A prominent example of this approach is the “full” diffusion decision model (DDM), which allows for inter-

*For correspondence, please contact Lukas Schumacher (schuma.luk@gmail.com)

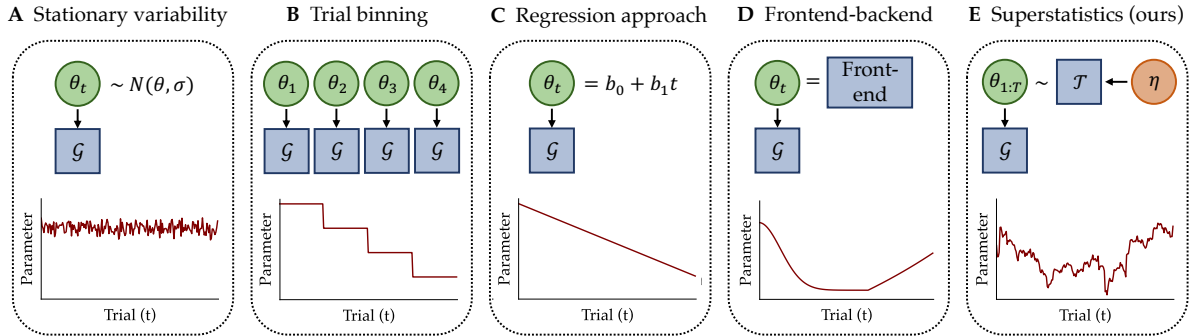


Figure 1: A conceptual illustration of the five main approaches to model temporal variation in the parameters θ of a cognitive model \mathcal{G} . **A** *Stationary variability*, also known as inter-trial variability, assumes that parameter values fluctuate around a stable mean. **B** *Trial binning* involves organizing the data into distinct bins and fitting a cognitive model \mathcal{G} to each bin individually. **C** *Regression approach* employs time (and sometimes additional contextual variables) as predictors for the parameters θ . **D** *Frontend-backend* models employ a mechanistic model, referred to as the frontend, to elucidate the dynamics of the parameter of the cognitive model (i.e., the backend). **E** *Superstatistics* involve a superposition of multiple stochastic processes operating on different temporal scales. They comprise a low-level observation model \mathcal{G} and a high-level transition model \mathcal{T} that specifies how the parameters θ_t evolve stochastically.

trial variability of its core parameters (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). However, stationary inter-trial variability mainly improves in-sample model fit and cannot identify systematic changes or sudden shifts in core model parameters. Moreover, the resulting model family still treats behavioral data as independent and identically distributed (IID) responses, making it unsuitable for investigating systematic changes in cognitive constructs.

Another approach for detecting systematic changes in cognitive model components is trial binning (Evans & Brown, 2017; Evans & Hawkins, 2019; Kahana et al., 2018). This method involves organizing data into discrete bins and then applying a stationary model to each of these data subsets separately (see Figure 1B). One can then examine variations in parameter estimates across these bins. The challenge in employing this approach is the selection of the number of time steps within each bin, which introduces an unwelcome trade-off between temporal resolution and estimation quality. For instance, if only a few time steps are chosen, the analysis can yield relatively fine-grained, but very uncertain estimates due to the low number of data points. A further shortcoming of trial binning is that estimates within a specific bin are not informed by data from neighboring bins. However, the appeal of dynamic modeling lies in the distinctive capability to utilize both past and future data to constrain the estimated parameter trajectories.

The third approach involves a generalized linear model (GLM) with time (and possibly other contextual factors) as a predictor of model parameters (Cochrane et al., 2023; Evans et al., 2018). The GLM approach is more appealing than trial binning, as it can detect linear or non-linear changes in model parameters without loss of resolution (see Figure 1C). However, the underlying regression function makes strong assumptions about the nature of the relationship between model parameters and time. Thus, even though a modeler will typically fit and compare a few plausible specifications (e.g., linear vs. exponential), it is often difficult to determine all plausible specifications *a priori*, and the overall flexibility of the GLM model as a process characterization remains severely limited (Gunawan et al., 2022).

Differently, the frontend-backend approach aims to account for changes in model parameters, while providing a mechanistic explanation for the dynamic nature of the target system (see Figure 1D). Here, the backend model pertains to the cognitive model which formalizes how the behavioral data is generated (e.g., a DDM). The frontend constitutes a mechanistic model, elucidating how the parameters of the backend model adapt over time, in different contexts and in response to additional factors (Brown et al., 2008; Fontanesi et al., 2019; Osth et al., 2018; Schumacher & Voss, 2023). This approach has several advantages, as it not only accommodates the dynamic nature of the parameters, but also provides a mechanistic description for their temporal variation through a set of static parameters and deterministic functions. For instance, there has been a recent trend to use reinforcement learning models as a frontend model to inform changes in DDM parameters due to reward-based learning (Fontanesi et al., 2019; McDougale & Collins, 2021; Miletić et al., 2021). Nevertheless, detailed frontend models are often challenging to develop, estimate, and compare.

Recently, we proposed an alternative approach that infers non-stationary parameter trajectories directly from the data, while imposing minimal constraints on how parameters change over time (Schumacher et al., 2023). Our approach leverages a framework known as *superstatistics* (Beck & Cohen, 2003; Beck, 2004; Mark et al., 2018), which involves a superposition of multiple stochastic processes operating on distinct time scales (see Figure 1E). At its core, this model comprises a low-level observation model and a high-level transition model. The former describes how data at a specific time point is generated, akin to the backend model. Like the frontend model, the transition model characterizes how the parameters change over time. However, from a superstatistics perspective, the transition model is inherently a stochastic process, exemplified, for instance, by a Gaussian random walk or a regime switching process.

The superstatistics approach effectively addresses the limitations of prior methodologies. Unlike stationary models, superstatistical models can readily generate non-stationary variations in the parameters of the low-level model, facilitating gradual or sudden transitions between different states. Furthermore, parameter estimates are contingent on past data points, thereby treating the data no longer as IID. In contrast to the trial-binning approach, models within the superstatistics framework leverage the entirety of available data, mitigating concerns about insufficient data points for parameter estimation. Different from GLM approaches, our superstatistics method imposes minimal assumptions on potential parameter trajectories, making it significantly less restrictive.

In contrast to frontend-backend models, superstatistics do not offer mechanistic explanations for parameter dynamics but provide greater flexibility in their estimation. Although mechanistic explanations are central to psychological research, there are cases where suitable explanations are lacking or are applicable only to specific parameters. Therefore, we consider these two approaches as complementary. The superstatistical framework takes a bottom-up, exploratory approach, functioning as a tool for generating hypotheses. In subsequent stages, one could potentially formulate plausible frontend models based on insights from parameter trajectories inferred with a superstatistical model. Additionally, superstatistical models can serve as benchmarks for testing and validating competing frontend-backend models by comparing resulting parameter trajectories from both methods.

Having laid out the potential benefits of the superstatistics framework and its applicability in the realm of cognitive process models (Schumacher et al., 2023), a pivotal question arises: Do the inferred parameter trajectories genuinely reflect shifts in the cognitive constructs they aim to represent, or are they merely a modeling artefact? To address this inquiry, we embark on an experimental validation study. In this study, we manipulate the experimental context in a manner that allows us to confidently anticipate how individuals and, consequently, their inferred cognitive constructs, will respond. In other words, if the inferred parameter time series mirror the alterations in the experimental context, we garner substantial evidence that these trajectories indeed reflect changes in the psychological constructs.

Throughout, we employ the well-established 4-parameter DDM (Ratcliff, 1978) as a low-level observation model. The DDM is a mathematical model that simultaneously accounts for response time (RT) and choice data obtained from two-alternative decision tasks. Fundamentally, it posits that, in forced-choice binary decision task, individuals accumulate evidence for the decision alternatives until a certain threshold is met, triggering a decision. Each of the DDM's four core parameters corresponds to a specific psychological construct: (i) the drift rate v signifies the average speed of information uptake; (ii) the threshold a serves as a proxy for decision caution; (iii) the relative starting point β represents *a priori* decision preferences; and (iv) the additional constant τ accounts for the duration of all processes taking place prior and following a decision, such as stimulus encoding or motor action (but see Verdonck et al., 2021).

A primary reason for our choice of the DDM as the observation model lies in its rigorous prior validation (Arnold et al., 2015; Lerche & Voss, 2019; Voss et al., 2004). These prior studies have convincingly demonstrated that the DDM's parameters are valid reflections of the intended psychological constructs. Moreover, the manipulation of experimental conditions leading to systematic alterations in specific DDM parameters is well-documented and comprehensively understood (Ratcliff & McKoon, 2008). For example, varying the difficulty of an experimental task alters the drift rate parameter, whereas providing verbal instructions to prioritize either speed or accuracy during task-solving leads to observable shifts in the threshold parameter and sometimes also in the non-decision time (Lerche & Voss, 2018).

In this study, we focus on the aforementioned experimental manipulations targeting the drift rate and the threshold parameters. We employed a *color discrimination task*, which was also utilized in the validation study by Voss et al. (2004). During this task, individuals must decide whether there are more blue or more orange pixels in a patch of pixels. The difficulty of the task can be easily manipulated by adjusting the ratio of blue and orange pixels. The farther the ratio is from 1:1, the easier the task becomes. Additionally, we manipulated the emphasis on speed or accuracy by verbally instructing participants to prioritize one over the other.

Systematic changes in cognitive model parameter can appear in different ways, ranging from changing slowly and gradually to more rapid and large shifts. In our experiment, we focus on two different types. Firstly, task difficulty changes frequently to the next easier or harder level, imitating gradual changes. Secondly, the speed-accuracy emphasis changes less regularly after each trial block, resembling sudden shifts. The primary aim of our experiment is to investigate whether the parameter trajectories inferred with a non-stationary DDM (NSDDM) match these changing patterns of the experimental conditions. Specifically, we expect the drift rate parameter to mirror the gradual changes of the task difficulty. Additionally, the threshold parameter should show sudden shifts when the priority switches between speed and accuracy. It is crucial to understand that in this application, the NSDDM does not have information about the experimental context and has to infer the parameter trajectory solely on the behavioral data.

When dealing with different types of fluctuations, another crucial question arises: What kind of transition model is most suitable for capturing the expected dynamics? To address this question, we implement different NSDDMs that solely differ in their transition model for the drift rate and the threshold parameter. Specifically, we compare four different transition models: (i) a Gaussian random walk; (ii) a mixture between a Gaussian random walk and uniformly distributed regime changes; (iii) a Lévy flight; and (iv) a regime switching function, where parameters either remain the same as in the previous time step or shift uniformly. These four transition models differ in their complexity (i.e., number of high-level parameters) and their ability to account for different types of temporal shifts.

Performing Bayesian model comparison and parameter estimation with superstatistical models can be computationally challenging (Schumacher et al., 2023). Therefore, we employ simulation-based inference (SBI, Cranmer et al., 2020) as implemented in the `BayesFlow` framework (Radev et al., 2023). `BayesFlow` enables us to carry out a principled Bayesian workflow utilizing simulation-based calibration (SBC, Säilynoja et al., 2022; Talts et al., 2020) and other validation methods (Gelman et al., 2020; Schad et al., 2021) that would otherwise be excessively time-consuming. The contributions of the present study can be summarized as follows:

1. We perform an experimental validation of different non-stationary instantiations of the diffusion decision model.
2. We propose an amortized method for Bayesian model comparison of non-stationary models via deep ensembles.
3. We showcase the potential of amortized Bayesian inference for increasing the aspirations of cognitive modeling.

Materials and Methods

Participants

A total of 14 participants (9 female, 5 male) were recruited for the experiment. The participants had an average age of 23.14 years ($SD = 1.29$, Range = [22, 26]). Every individual provided informed consent to participate in the study, and the research protocol received approval from the local ethics committee. The entire study was conducted in accordance with the ethical principles outlined in the Helsinki Declaration.

Task

The participants completed a total of 800 trials in a color discrimination task, including 32 practice trials. In each trial, individuals were presented with a rectangular patch containing blue and orange pixels and had to determine whether there were more blue or orange pixels. Prior to the patch presentation, a fixation cross was displayed for 300 ms. All stimuli were presented on a gray background.

Task difficulty was manipulated by varying the proportion of blue/orange pixels in the patch. The following ratios were utilized: 50.5:49.5; 52.25:47.75; 53.5:46.5; and 55:45. Half of the trials featured orange as the dominant color, while the other half featured blue. The difficulty level remained constant for either 8 or 16 trials before transitioning to the next level of difficulty.

In addition to manipulating task difficulty, participants received two types of instructions which changed every 48 trials. In the “accuracy” condition, individuals were instructed to prioritize accuracy in their responses. Conversely, in the “speed” condition, participants were directed to emphasize speed while maintaining a reasonable level of accuracy. Feedback was provided after each trial to make participants aware of their performance: a green cross for correct responses, a red minus for incorrect responses, and a red clock for responses slower than 700 ms in the speed condition.

Superstatistics Framework

To represent non-stationary changes in DDM parameters, we adopt a superstatistics framework (Beck & Cohen, 2003; Mark et al., 2018). Within this framework, each generative model comprises (at least) a *low-level observation model* \mathcal{G} characterized by time-dependent local parameters $\theta_t \in \mathbb{R}^K$ that vary according to a *high-level transition model* \mathcal{T} with static high-level parameters $\eta \in \mathbb{R}^D$. These models simulate parameters and observable data $x_t \in \mathcal{X}$ according to the following general recurrent system

$$\begin{aligned} \theta_t &= \mathcal{T}(\theta_{0:t-1}, \eta, \xi_t) & \text{with } \xi_t &\sim p(\xi | \eta), \theta_0 \sim p(\theta) \\ x_t &= \mathcal{G}(x_{1:t-1}, \theta_t, z_t) & \text{with } z_t &\sim p(z | \theta_t), \end{aligned} \quad (1)$$

where \mathcal{T} represents an arbitrary high-level transition function parameterized by η , and \mathcal{G} is a (non-linear) transformation that encapsulates the functional assumptions of the low-level model. The random variates ξ_t and z_t govern the stochastic nature of the two model components through noise outsourcing. The initial parameter configuration θ_0 adheres to a prior distribution $\theta_0 \sim p(\theta)$ encoding the available information about feasible starting parameter values.

The above formulation is very abstract and general, highlighting the flexibility of the superstatistics framework. Moreover, it does not assume that the corresponding transition or likelihood densities, given by

$$\mathbb{T}(\theta_t | \eta, \theta_{0:t-1}) = \int p(\theta_t, \xi | \eta, \theta_{0:t-1}) d\xi \quad (\text{implied transition density}) \quad (2)$$

$$p(x_t | \theta_t, x_{1:t-1}) = \int p(x_t, z | \theta_t, x_{1:t-1}) dz \quad (\text{implied likelihood density}), \quad (3)$$

are tractable or available in closed-form, situating our approach in the context of simulation-based inference (SBI, Cranmer et al., 2020). Here, we build on SBI with neural networks (Ardizzone et al., 2018; Greenberg et al., 2019; Radev, Mertens, et al., 2020) as a principled approach to perform fully Bayesian inference by using only samples from the generative system defined by Equation 1. Importantly, our estimation methods overcome key limitations of previous approaches related to the curse of dimensionality (Mark et al., 2018).

Low-Level Model

In this work, we use the same standard DDM implementation as a low-level observation model \mathcal{G} for all NSDDMs. The low-level dynamics of the evidence accumulation process are described by the following stochastic ordinary differential equation:

$$dx_n = v dt_s + z \sqrt{dt_s} \quad \text{with } z \sim \mathcal{N}(0, 1) \quad (4)$$

Accordingly, the evidence x_n on a given trial n follows a random walk with drift v and Gaussian noise z , where t_s represents time on a continuous time scale. The core assumption of the DDM is that evidence is accumulated with a fixed rate v until one of two thresholds, a or 0 , is reached, and the corresponding decision D_n is made:

$$D_n = \begin{cases} 1, & \text{if } x_n \geq a \\ 0, & \text{if } x_n \leq 0 \end{cases}. \quad (5)$$

Furthermore, the DDM incorporates an additive constant τ , which represents the time allocated to all non-decisional processes (i.e., stimulus encoding and motor action). Consequently, the DDM encompasses three distinct free parameters, namely $\theta = (v, a, \tau)$. We fixed the starting point of the evidence accumulation process at $a/2$, since, in our case, the two boundaries of the accumulation process correspond to correct and incorrect responses, respectively. Thus, it is unwarranted to estimate any potential *a priori* bias towards either of these boundaries (Voss et al., 2013).

High-Level Models

We formulate and compare four different high-level transition models, denoted as $\mathbb{T}_1, \dots, \mathbb{T}_4$, which govern the trial-by-trial changes in local DDM parameters $\theta_{1:T}$. These transition models vary in terms of their flexibility in allowing changes to the low-level parameters and their underlying complexity, including the number of high-level parameters

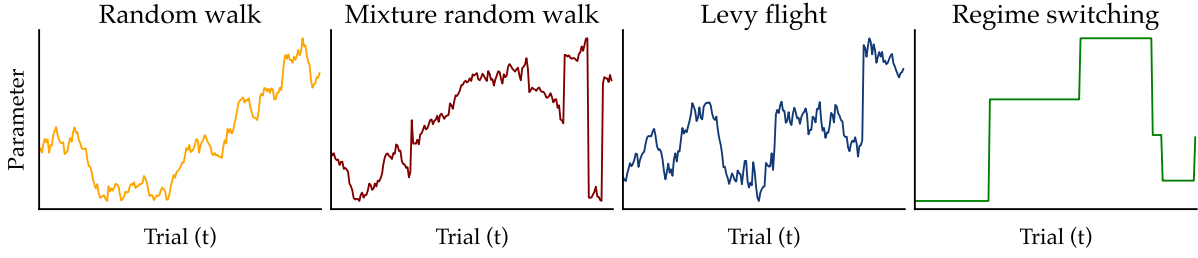


Figure 2: An example illustration of the four high-level (transition) models considered in our study, governing the temporal variation of a hypothetical low-level model parameter.

involved (see Figure 2 for an exemplar trajectory generated by each transition model). To ensure that the low-level parameters remain within plausible ranges, we impose both lower and upper bounds on their trajectories. Specifically, we set the upper bounds for the parameters v , a , and τ to 8, 6, and 4, respectively. Additionally, since negative parameter values are not meaningful for our DDM specification, we set the lower bounds for all parameters to 0. For all transition models, we assume independence between the trajectories of the local DDM parameters.

Random Walk The first transition model (\mathbb{T}_1) convolves the low-level model’s parameters with a Gaussian distribution, resulting in a gradual change that follows a random walk:

$$\mathbb{T}_1(\theta_{k,t} | \theta_{k,t-1}, \sigma_k) = \mathcal{N}(\theta_{k,t} | \theta_{k,t-1}, \sigma_k) \quad (6)$$

According to this transition model, the current value of each parameter $\theta_{k,t}$ is only influenced by its previous value $\theta_{k,t-1}$, generating more or less auto-correlated and gradual changes.

Mixture Random Walk The second transition model (\mathbb{T}_2) corresponds to a mixture distribution between a random walk (cf. Equation 6) and uniformly distributed shifts:

$$\mathbb{T}_2(\theta_{k,t} | \theta_{k,t-1}, \rho_k, \sigma_k, a_k, b_k) = \rho_k \mathcal{N}(\theta_{k,t} | \theta_{k,t-1}, \sigma_k) + (1 - \rho_k) \mathcal{U}(a_k, b_k) \quad (7)$$

where ρ indicates the probability of the type of change (gradual change or shift) as a mixing coefficient for the two states. The upper and lower bounds of the uniform distribution, denoted as a and b , are set to cover plausible parameter ranges and are not treated as free parameters themselves.

Lévy-Flight The Lévy flight transition model (\mathbb{T}_3) is similar to the Gaussian random walk. However, instead of assuming normally distributed noise, it assumes an alpha-stable transition for each component of θ :

$$\mathbb{T}_3(\theta_{k,t} | \theta_{k,t-1}, \sigma_k, \alpha_k) = \text{Alpha-Stable}(\theta_{k,t} | \theta_{k,t-1}, \sigma_k, \beta = 0, \alpha_k) \quad (8)$$

where $0 < \alpha \leq 2$ governs the heaviness of the noise distribution’s tails. If $\alpha_k = 2$ then the distribution is equivalent to a Gaussian distribution. Notably, as the value of α decreases, the distribution’s tails get heavier, allowing for larger shifts in the parameter values. When simulating from the Lévy flight transition model, we use a scale of $\sigma_k/\sqrt{2}$, such that the corresponding Gaussian distribution for $\alpha_k = 2$ has a standard deviation of σ_k .

Regime Switching Finally, the regime switching transition model (\mathbb{T}_4) is a simpler version of the mixture random walk. The parameter’s trajectory adheres to one of two possibilities: it either maintains its previous value or undergoes a uniform shift:

$$\mathbb{T}_4(\theta_{k,t} | \theta_{k,t-1}, \rho_k, a_k, b_k) = \rho_k \delta(\theta_{k,t} - \theta_{k,t-1}) + (1 - \rho_k) \mathcal{U}(a_k, b_k), \quad (9)$$

where $\delta(\cdot)$ is the Dirac delta distribution indicating that the parameter either does not change at all with probability ρ or undergoes a sudden change with probability $1 - \rho$.

Strictly speaking, some of the above transition models can effectively be transformed into others by employing specific high-level parameter configurations. For instance, the mixture random walk with $\sigma = 0$ reduces to the regime

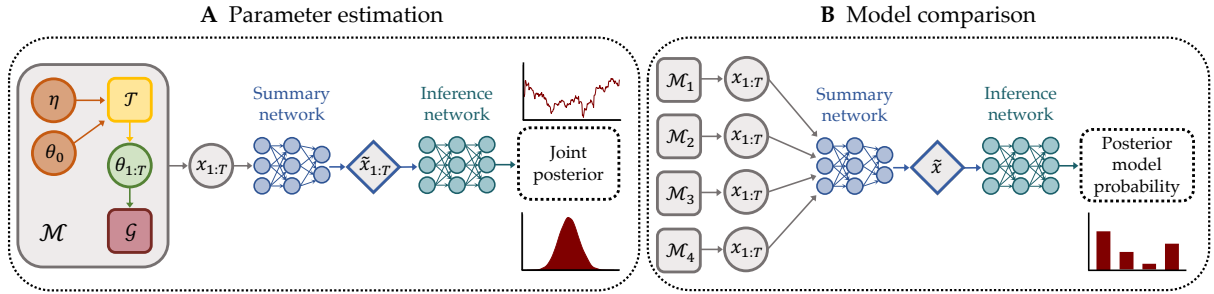


Figure 3: A conceptual illustration of our amortized Bayesian inference training setup. **A Parameter estimation** A recurrent *summary* network processes the synthetic time series $x_{1:T}$ and learns maximally informative temporal summary statistics $\tilde{x}_{1:T}$. An *inference network* (i.e., normalizing flow) learns to approximate the joint posterior distribution of time-varying low-level parameters $\theta_{1:T}$ and static high-level parameters η given the learned summaries. **B Model comparison** A transformer *summary* network consumes time series simulated from competing models and learns maximally informative summary vectors \tilde{x} . An *inference network* (i.e., a probabilistic classifier) learns to approximate posterior model probabilities (PMPs) given the summary vectors. Once trained, the networks can be efficiently validated using principled Bayesian methods and applied to the observed data.

switching transition function. Conversely, when $\rho = 0$ it reduces to a simple Gaussian random walk. Also, the Lévy flight transition model with $\alpha = 2$ turns into a random walk transition function. The mixture random walk and the Lévy flight transition function both have two high-level parameters and can thus be regarded as more complex and more flexible than the other two transition models, which only have a single high-level parameter. Notably, the random walk transition model is the only one that cannot generate relatively large sudden shifts in parameter values.

Model Comparison Setup

One of the major aims of this study is to compare four NSDDMs sharing the same low-level diffusion model but differing in their assumptions about the type of stochastic variation of the drift rate (v) and threshold (a) parameters. All four NSDDMs employ the same Gaussian random walk model \mathbb{T}_1 for the non-decision time parameter (τ). We base this decision on previous research (Schumacher et al., 2023) and the rationale of our experimental manipulations, which should not imply sudden large shifts in the τ parameter. For \mathcal{M}_1 , the drift rate and threshold parameter also follow a Gaussian random walk, resulting in three high-level parameters, $\eta = (\sigma_v, \sigma_a, \sigma_\tau)$. In \mathcal{M}_2 , both v and a follow a mixture between a Gaussian random walk and uniform shifts (\mathbb{T}_2), which results in a total of five high-level parameters, $\eta = (\sigma_v, \sigma_a, \sigma_\tau, \rho_v, \rho_a)$. In contrast, \mathcal{M}_3 introduces a trajectory for the drift rate and threshold parameters characterized by a Lévy flight (\mathbb{T}_3), which has five free high-level parameters, $\eta = (\sigma_v, \sigma_a, \sigma_\tau, \alpha_v, \alpha_a)$. Lastly, for \mathcal{M}_4 , the two parameters v and a either remain the same as in the previous time point or shift uniformly (\mathbb{T}_4). This model has a total of three high-level parameters, $\eta = (\sigma_\tau, \rho_v, \rho_a)$. A listing of the weakly informative prior distributions assigned to the model parameters can be found in the **Appendix**.

Amortized Bayesian Inference

Amortized Bayesian inference (ABI) is a flexible framework for estimating, comparing, and validating complex models through simulation-based training of specialized neural networks (Radev et al., 2023). ABI consists of (i) a training phase where the networks learn a surrogate distribution, and (ii) an inference phase where the networks infer the target quantities (e.g., model parameters or model posterior probabilities) in real-time for any new data set supported by the model(s). The neural networks are trained purely on simulations from the generative model and do not require an explicit likelihood or numerical integration. Thus, ABI re-casts expensive Bayesian inference into a neural network prediction task, such that sampling from the target posterior and model refits happen almost instantaneously.

Amortized Parameter Estimation Our deep learning approach for jointly estimating time-varying and static parameters follows Schumacher et al. (2023), who extend ideas from ABI with static parameters (Gonçalves et al., 2020; Radev, Mertens, et al., 2020) to non-stationary Bayesian models. Accordingly, our goal is not only to infer the tra-

jectories of all three model parameters $\{\theta_t\}_{t=1}^T$, but also to estimate the posterior distribution for the static high-level parameters η of the transition model. Thus, we are interested in recovering the full joint posterior $p(\theta_{1:T}, \eta | x_{1:T})$ from the observed time series $\{x_t\}_{t=1}^T$:

$$p(\theta_{1:T}, \eta | x_{1:T}) \propto p(\eta, \theta_0) p(x_1 | \theta_1) \prod_{t=2}^T p(x_t | \theta_t, x_{1:t-1}) \prod_{t=1}^T \mathbb{T}(\theta_t | \eta, \theta_{0:t-1}) \quad (10)$$

where $p(\eta, \theta_0)$ is the joint prior over high-level parameters and initial low-level parameter values. The joint prior typically factorizes as $p(\eta, \theta_0) = p(\eta)p(\theta_0)$, assuming that η and θ_0 are independent in the absence of any information. Even though our SBI method is applicable to any model of the general form in Eq. 10, our low-level (**Low-Level Model**) and high-level (**High-Level Models**) specifications lead to a simplified formulation

$$p(\theta_{1:T}, \eta | x_{1:T}) \propto p(\eta, \theta_0) \prod_{t=1}^T p(x_t | \theta_t) \prod_{t=1}^T \mathbb{T}(\theta_t | \eta, \theta_{t-1}). \quad (11)$$

The simplified formulation follows from the fact that our transition models share the Markov property and the DDM likelihood depends on time only through the current parameter θ_t in the latent trajectory $\theta_{1:T}$.

Following the typical ABI offline training setting (see Figure 3A for a conceptual illustration), we generate a *data set of simulated data sets*, $\mathcal{D} = \{\eta^{(b)}, \theta_{1:T}^{(b)}, x_{1:T}^{(b)}\}_{b=1}^B$, and use the simulated data to train a specialized neural network, $F_\psi(\theta_{1:T}, \eta; x_{1:T})$, which approximates the full joint posterior (i.e., a normalizing flow, see Papamakarios et al., 2021). In particular, we minimize the following loss in expectation over the full non-stationary generative model (i.e., the right hand-side of Eq. 10)

$$\mathcal{L}(\psi) = \mathbb{E}_{(\eta, \theta_{1:T}, x_{1:T}) \sim \mathcal{D}} [-\log q_\psi(\theta_{1:T}, \eta | x_{1:T})], \quad (12)$$

where we approximate the expectation over $p(\theta_0)p(\eta, \theta_{1:T}, x_{1:T})$ via our training set \mathcal{D} and regularize against overfitting with standard techniques, such as dropout and weight decay. It is also possible to run the simulator(s) indefinitely and perform online training using on-the-fly simulation (Radev, Mertens, et al., 2020). In fact, this approach should be preferred for *fast simulators*, as it makes overfitting hardly possible. Thus, online learning is the approach we pursue for estimating the parameters of our NSDDMs.

In the context of dynamic Bayesian models, we have many choices on how to *factorize* the joint posterior (Särkkä, 2013). The two most common choices are to approximate the *filtering distribution* or the *smoothing distribution* (Mark et al., 2018). The filtering distribution corresponds to an online analysis, where the low-level parameters θ_t at time step t are only informed by past data points. Differently, the smoothing distribution conditions the posterior of θ_t on all past and future data points, and provides potentially sharper estimates. Thus, in this study, we exclusively target the approximate smoothing distribution due to its superior parameter recoverability in an offline analysis.² In practice, we employ unidirectional or bidirectional long-short term memory (LSTM) networks (Gers et al., 2000) with many-to-many input-output relationships as a backbone for approximating the filtering or the smoothing distribution, respectively. We then train four separate neural approximators, such that each network becomes an “expert” in inferring the smoothing distribution of the corresponding NSDDM. The **Appendix** contains more details on the neural network settings and training hyperparameters.

Amortized Model Comparison To conduct a comparative analysis of the four NSDDMs, we focus on Bayes factors (BFs) and posterior model probabilities (PMPs), which can be classified as *prior predictive* methods embodying Occam’s razor (Kass & Raftery, 1995; MacKay, 2003). The efficacy of these measures has been demonstrated in a wide range of psychological modeling studies (Heck et al., 2023). Nevertheless, an ongoing debate surrounds the preference between the two (Tendeiro & Kiers, 2019; van Ravenzwaaij & Wagenmakers, 2022). Since BFs and posterior odds (i.e., ratios between PMPs) are equivalent when all models are assumed to be equally likely *a priori*, we estimate and analyse both quantities in our study.

Following the common Bayesian terminology (MacKay, 2003), we can refer to the four competing models through an index set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$. The aim of prior predictive Bayesian model comparison is to find the simplest

²Note, that Schumacher et al. (2023) focused exclusively on the filtering distribution in their benchmarking experiments.

most plausible model within \mathcal{M} . To this end, we can compute PMPs for each of the competing models

$$p(\mathcal{M}_j | x_{1:T}) = \frac{p(x_{1:T} | \mathcal{M}_j) p(\mathcal{M}_j)}{\mathbb{E}_{p(\mathcal{M})} [p(x_{1:T} | \mathcal{M})]}, \quad (13)$$

where $p(\mathcal{M})$ refers to the prior distribution over the discrete model space. The marginal likelihood $p(x_{1:T} | \mathcal{M}_j)$ plays a crucial role in Equation 13, and can be expressed by integrating out all parameters of the joint model,

$$p(x_{1:T} | \mathcal{M}_j) = \int p(\eta, \theta_0) \prod_{t=1}^T p(x_t | \theta_t, \mathcal{M}_j) \prod_{t=1}^T \mathbb{T}_j(\theta_t | \eta, \theta_{t-1}) d\eta d\theta_0, \dots, d\theta_T. \quad (14)$$

Importantly, since the marginal likelihood averages the likelihood over the joint prior, it automatically incorporates a probabilistic Occam’s razor, favoring models with constrained prior predictive flexibility. When comparing a pair of competing models, \mathcal{M}_j and \mathcal{M}_i , we can compute the ratio between their respective marginal likelihood,

$$\text{BF}_{ji} = \frac{p(x_{1:T} | \mathcal{M}_j)}{p(x_{1:T} | \mathcal{M}_i)}. \quad (15)$$

This ratio is referred to as the Bayes factor (BF). Consequently, a $\text{BF}_{ji} > 1$ signifies a relative preference for model j over model i based on the given data $x_{1:T}$ (Kass & Raftery, 1995).

Unfortunately, the marginal likelihood is notoriously hard to approximate (Gronau et al., 2017) and even doubly intractable for mechanistic models with unknown or unnormalized likelihoods. To circumvent this intractability, we follow the neural method of Elsemüller, Schnuerch, et al. (2023) and Radev, D’Alessandro, et al. (2020) which enables amortized Bayesian model comparison for arbitrary computational models (see Figure 3B for a graphical illustration). This method involves the simultaneous training of two neural networks with different roles: a *summary network* and an *inference network*. The summary network learns maximally informative summary statistics from the raw data (e.g., behavioral time series). The inference network approximates the PMPs for the candidate models, $q_\phi(\mathcal{M} | x_{1:T})$ given the outputs of the summary network. Here, we subsume all trainable network parameters under ϕ and refer to the composition of the two networks as an *evidential network*.

The training data for the evidential network consists of all simulations from the candidate models together with the corresponding model index, $\mathcal{D}(\mathcal{M}) = \{x_{1:T}^{(b)}, \mathcal{M}_j^{(b)}\}_{b=1}^{B'}$, where B' denotes the total number of simulations from all models. Together, the two networks minimize the standard cross-entropy loss,

$$\mathcal{L}(\phi) = \mathbb{E}_{(\mathcal{M}_j, x_{1:T}) \sim \mathcal{D}(\mathcal{M})} \left[- \sum_{j=1}^J \mathbb{I}_{\mathcal{M}_j} \log q_\phi(\mathcal{M}_j | x_{1:T}) \right], \quad (16)$$

and we approximate the expectation over $p(\eta, \theta_{1:T}, x_{1:T})$ by our training set $\mathcal{D}(\mathcal{M})$, and $\mathbb{I}_{\mathcal{M}_j}$ denotes an indicator function (i.e., one-hot encoding) for the true model index. In principle, we could use online learning for amortized model comparison as well, but we found *offline training* to yield sufficiently accurate results.

More recently, Elsemüller, Olischläger, et al. (2023) demonstrated the importance of gauging the sensitivity of amortized neural approximators, especially in the context of model comparison. The authors suggest to train an ensemble of multiple evidential networks, instead of relying on a single network. Accordingly, we can measure the (lack of) agreement between ensemble members and obtain a hint at the robustness of the approximate PMPs. Here, we trained an ensemble of ten evidential networks and computed the mean and standard deviation of the estimated PMPs across all ten networks. For more details regarding the neural network architecture and training settings, we refer the reader to the **Appendix**.

Code Availability Complete code for reproducing the results reported in this manuscript is available in the project’s GitHub repository <https://github.com/bayesflow-org/Non-Stationary-DDM-Validation>.

Results

Model Comparison

As a first step, we assess the closed-world (i.e., *in silico*) performance of our model comparison method in terms of computational faithfulness and accuracy of model recovery. To assess the former, we perform simulation-based

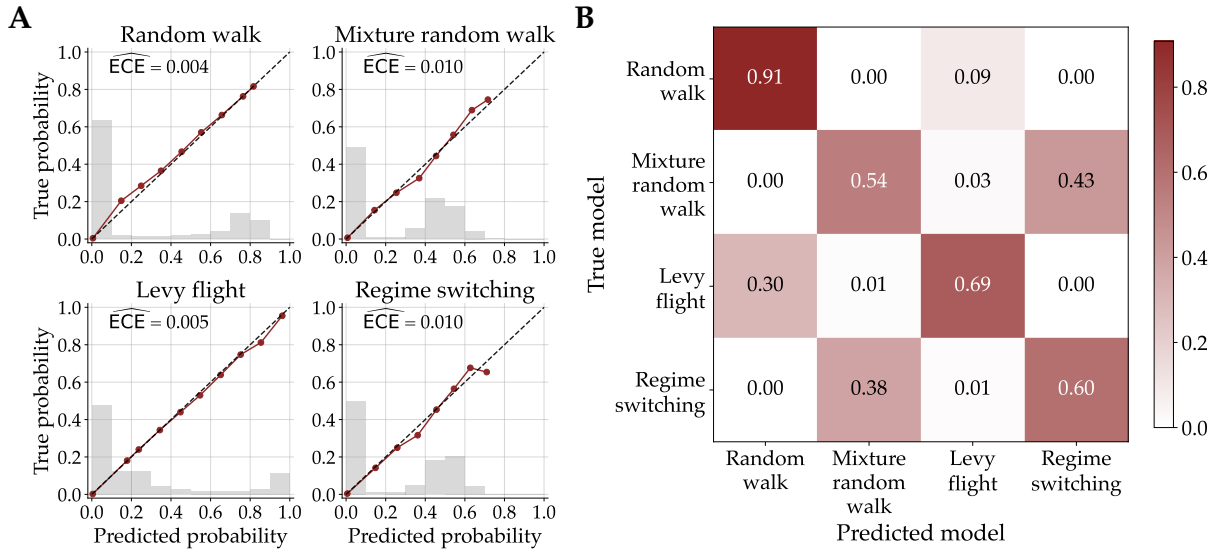


Figure 4: In silico model comparison and sensitivity results. **A** Calibration curves of all four NSDDMs aggregated across the neural approximator ensemble. Additionally, the expected calibration error (\widehat{ECE}) is annotated within each subfigure. The gray histograms depict the relative frequencies of the predicted model probabilities. **B** Confusion matrix between true data generating model and predicted model. The proportion values were averaged across the ten neural approximator within the ensemble.

calibration (SBC; Säilynoja et al., 2022; Talts et al., 2020) based on 10 000 synthetic data sets per model. Figure 4A shows the calibration curves for each NSDDM averaged across the ten evidential networks in our deep ensemble. We observe excellent calibration with very minimal expected calibration errors (\widehat{ECE}) across all models. Thus, we conclude that the approximate posterior probabilities are well-calibrated in the closed-world setting.

Next, we assess the accuracy of our model comparison networks in terms of their ability to correctly identify the ground-truth data-generating model. To this end, we apply the deep ensemble to the 40 000 synthetic data sets we have already simulated for assessing calibration. In Figure 4B, we present the resulting confusion matrix, which illustrates the agreement between true and predicted models averaged across the ten approximators. Among the four models, the random walk DDM is the only one that rarely gets confused with the other models. A possible explanation is that it is the only transition model not capable of generating sudden shifts in parameter values. The remaining models are susceptible to more frequent misclassifications. For example, the mixture random walk DDM is correctly identified only 54% of the time, and it is often confused with the regime switching model, occurring 43% of the time. Notably, the Lévy flight DDM is prone to mimicry with the random walk DDM (on average 30% of the time).

It is essential to emphasize that these results do not imply a deficiency in your model comparison method, but rather underscore the fact that certain pairs of models, such as the mixture random walk and the regime switching DDM, can generate remarkably similar data patterns. For instance, a significant portion of the prior distribution’s mass for the α parameter of the Lévy flight transition model centers around 2. If $\alpha \approx 2$, then the Lévy alpha-stable distribution closely resembles a Gaussian distribution, with equality in the case of $\alpha = 2$. Consequently, simulating the Lévy flight DDM would often yield data patterns that could have just as plausibly originated from the simpler random walk DDM.

Similarly, a substantial portion of the prior mass for the σ priors of the mixture random walk transition model clusters around 0, which subsequently transforms it into a regime switching transition model, resulting in large overlap in synthetic data sets. Interestingly, the mixture random walk and the Lévy flight DDM are seldom confused, even though both models can produce subtle local changes and large sudden shifts. This implies that these two transition models generate qualitatively similar but quantitatively easy to distinguish parameter trajectories. In summary, the observation of occasional model confusion is not a limitation of our method; rather, it underscores our method’s effectiveness in discerning when two models generate highly similar data, making them less straightforward to differentiate from each

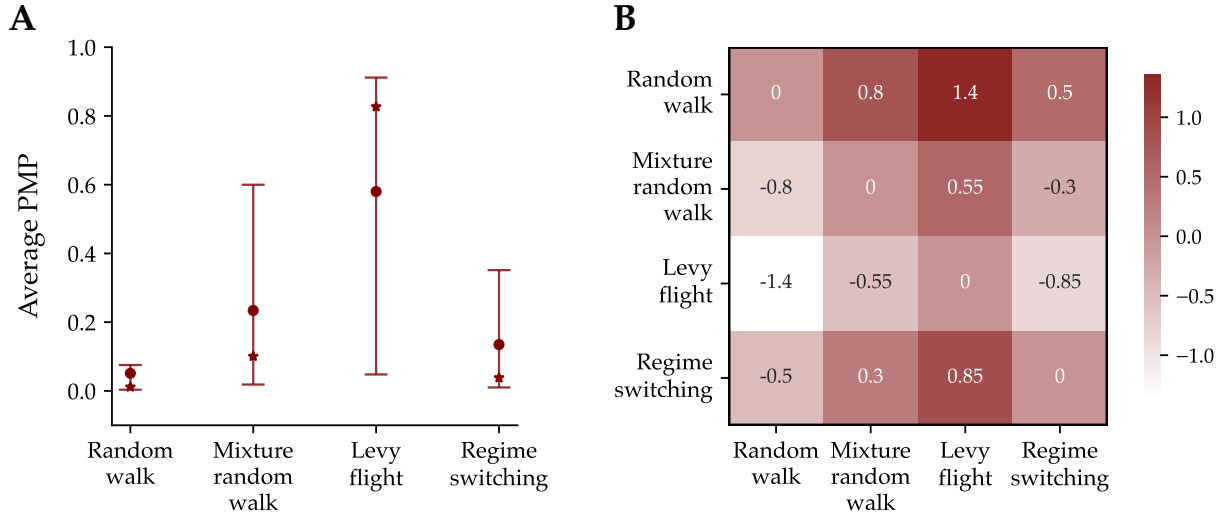


Figure 5: Empirical model comparison results. **A** Aggregate posterior model probabilities (PMP) across the ensemble and the 14 individual participants. Points depict the mean, stars the median, and the error bars indicate the 75% credibility interval (CI). **B** Heatmap of average \log_{10} Bayes factors (BF). Both metrics agree on favoring the Lévy flight DDM over the other models.

other. Moreover, the amortization property of our method enables us to easily conduct such simulation studies prior to analyzing real data – estimating 40 000 posterior model probabilities would have been infeasible for any other method.

After successfully validating our model comparison method, we apply the deep ensemble to the empirical data of the 14 participants. Each approximator in the ensemble was used to infer posterior model probabilities (PMP) for each model, considering each individual’s data separately. Subsequently, we calculated the mean (points), median (stars), and 75% credibility interval (CI) for the PMPs for all approximators the 14 individuals (Figure 5A). The analysis reveals that the Lévy flight DDM is the most plausible model with an average PMP of approximately 60%. It was the most plausible model for 9 out of the 14 participants. In contrast, the mixture random walk model collects an average PMP of less than 30%. Nevertheless, it was estimated to be the most plausible model for 5 participants. The random walk DDM and regime switching DDM were consistently less plausible than the other models and did not emerge as superior for any of the participants.

In addition to PMPs, we computed \log_{10} Bayes factors (BF). Figure 5B depicts a heatmap of BFs for all one-to-one comparisons between our four NSDDMs, averaged across the participants and the evidential networks of the ensemble. Following Kass and Raftery (1995), an absolute value of $\log_{10}(\text{BF}) > 2$ indicates *decisive* evidence, absolute values between 1 to 2 signify *strong*, and between 0.5 to 1 *substantial* evidence. An absolute value of $\log_{10}(\text{BF}) < 0.5$ is labeled as *not worth more than a bare mention*. The BF patterns in Figure 5B align with the PMP findings, implying *strong* evidence for the Lévy flight DDM over the random walk DDM and *substantial* evidence over the other NSDDMs. Also, both the mixture random walk and the regime switching DDM have *substantial* evidence over the random walk model. Interestingly, there is little evidence favoring the mixture random walk DDM over the regime-switching model, suggesting comparable performance.

These findings offer two substantive insights. First, the ability of transition models to generate sudden shifts in parameters seems essential, as seen in the random walk DDM’s lower plausibility. Moreover, the regime switching DDM, allowing for occasional shifts, but neglecting small gradual changes, performed less effectively than the more complex models. This result underscores the importance of accommodating both gradual as well as sharp changes in model parameters for achieving optimal fit. Consequently, the more complex NSDDMs, particularly the mixture random walk DDM and Lévy flight DDM, emerged as more plausible than their simpler counterparts, despite the implicit penalty for prior complexity imposed by Bayesian model comparison.

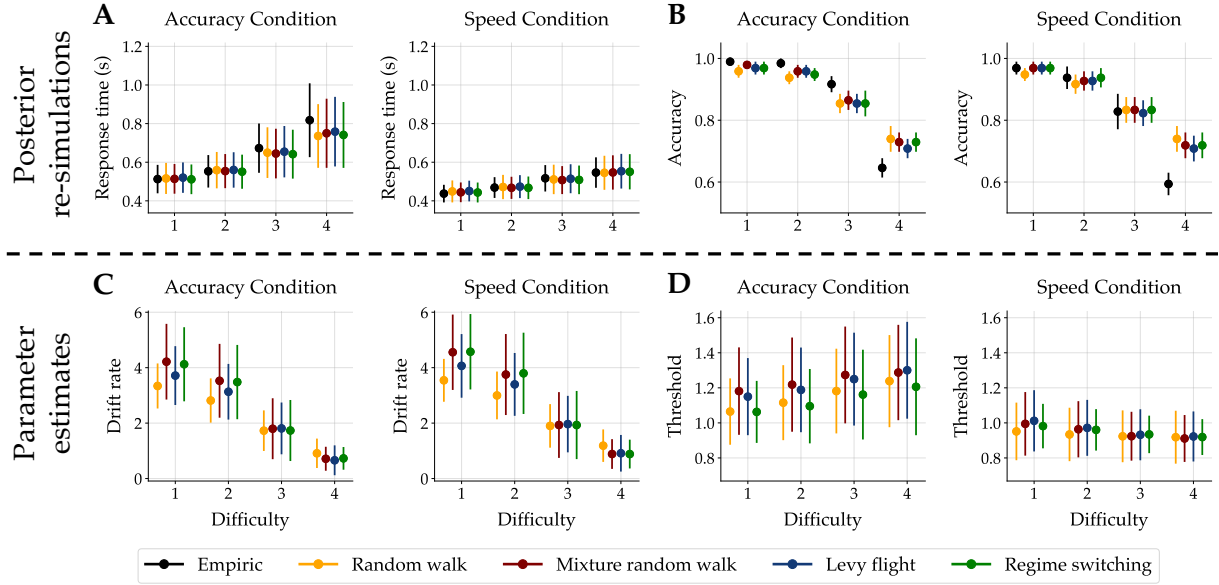


Figure 6: Aggregated results from all models fitted to the empirical data. The top row illustrates posterior re-simulations as a measure of the model’s generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated RTs for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individuals and re-simulations.

Posterior Re-simulation

Subsequently, we fit all four variants of the NSDDM to each of the 14 data sets, evaluating the absolute goodness-of-fit of each model. To achieve this, we conducted 500 re-simulations with randomly sampled posterior parameter trajectories for each individual data set. In Figure 6A, we present the median and median absolute deviation (MAD) of response times (RT) across all individuals and re-simulations. We provide these aggregates for each NSDDM, categorized by task difficulty level and the two experimental conditions. Notably, an initial observation reveals that the experimental manipulations were effective on average: empirical median RTs increased with task difficulty, and individuals tended to respond faster in the speed condition compared to the accuracy condition. Remarkably, all four variants of the NSDDM demonstrated an outstanding fit to these empirical data patterns. Solely, RTs in the accuracy condition with the highest task difficult level consistently are underestimated by all NSDDM variants.

The empirical and re-simulated proportion of correct choices (accuracy) are aggregated and presented in the same way as the RTs (see Figure 6B). Again, the empirical data mirror the anticipated patterns resulting from our experimental manipulations. As expected, accuracy diminishes with increasing task difficulty. Individuals are generally less accurate in the *speed* condition compared to the *accuracy* condition. Although NSDDMs successfully reproduce the general patterns in the choice data, we observe notably worse re-simulation compared to that of the RTs data. In both accuracy and speed conditions, re-simulated accuracies exhibit a less pronounced decline as a function of difficulty than observed in the empirical data. Further, the difference in accuracy between the two experimental conditions is less pronounced in the re-simulated data compared to the behavioral data. Notably, the random walk DDM underperforms relative to the other three NSDDMs in this analysis.

It is important to highlight that, unlike conventional approaches, the models did not receive any information regarding the specific experimental context an individual faced at any given moment. From these analyses, we conclude that all NSDDM implementations successfully capture the general patterns in the empirical RT data. Individual participant analyses, detailed in the **Appendix**, affirm the same conclusions.

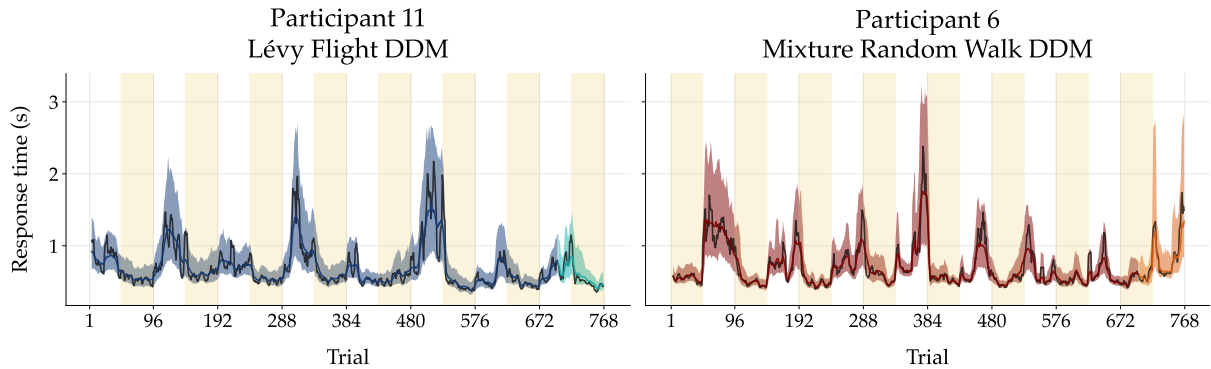


Figure 7: Model fit to response time (RT) time series. The empirical RT time series of two exemplar individuals are shown in black. From trial 1 to 700, the posterior re-simulations (aka retrodictive checks) using the best fitting non-stationary diffusion decision model (NSDDM) for the specific individual are shown in blue and red, respectively. In this instance, the left column showcases results from a Lévy flight DDM, while the right column displays parameter trajectories from a mixture random walk DDM. For the remaining trials, one-step-ahead posterior predictions from the NSDDMs are depicted in cyan and orange, respectively. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

In addition to the analysis of the absolute model fit on the aggregate level, we evaluated the fit across the RT time series. For every participant, we generated 250 posterior re-simulations for the first 700 trials with the corresponding best fitting NSDDM. The remaining 68 data points were left out for predictive analysis. Employing a *one-step-ahead* prediction approach, we iteratively forecasted the subsequent data point, followed by a re-fitting of the model in each step.

Figure 7 illustrates the empirical and re-simulated RT time series for two exemplary participants. Results for the remaining 12 participants can be seen in the **Appendix**. The colored lines depict the median and the shaded bands represent to 90% credibility intervals (CI) across the 250 re-simulations. Both the empirical data (solid black lines) and the re-simulated/predicted RTs were smoothed using a simple moving average (SMA) with a period of 5. Yellow shaded regions highlight trials where speed was emphasised over accuracy, whereas blank white areas denote instances where the opposite emphasis was applied. Overall, RTs were slower and more variable in the accuracy condition. Notably, the NSDDM not only closely replicated the empirical time series but also effectively predicted future data points. This suggests that the model does not overfit the data.

Parameter Estimates

At the heart of the current validation study are the inferred parameters, prompting a crucial question: Do these parameter dynamics align with the sequence of experimental manipulations? We address this question by examining both the time-averaged and time-varying estimates.

Aggregate Analysis We initially examine the parameter estimates averaged across individuals for each difficulty level and condition separately. This provides a comprehensive overview of average effects on model parameters in different experimental contexts, at first, without delving into the temporal aspect. The bottom panel of Figure 6 illustrates the posterior medians and MADs collapsed onto the different experimental contexts for the drift rate (Figure 6C) and threshold parameter (Figure 6D).

Analyzing the aggregated drift rate estimates reveal an anticipated pattern. On average, the drift rate decreases as task difficulty increases, observed in both the accuracy and speed conditions. Additionally, slightly higher overall values are estimated in the speed condition compared to the accuracy condition. While all four NSDDMs yield fairly similar

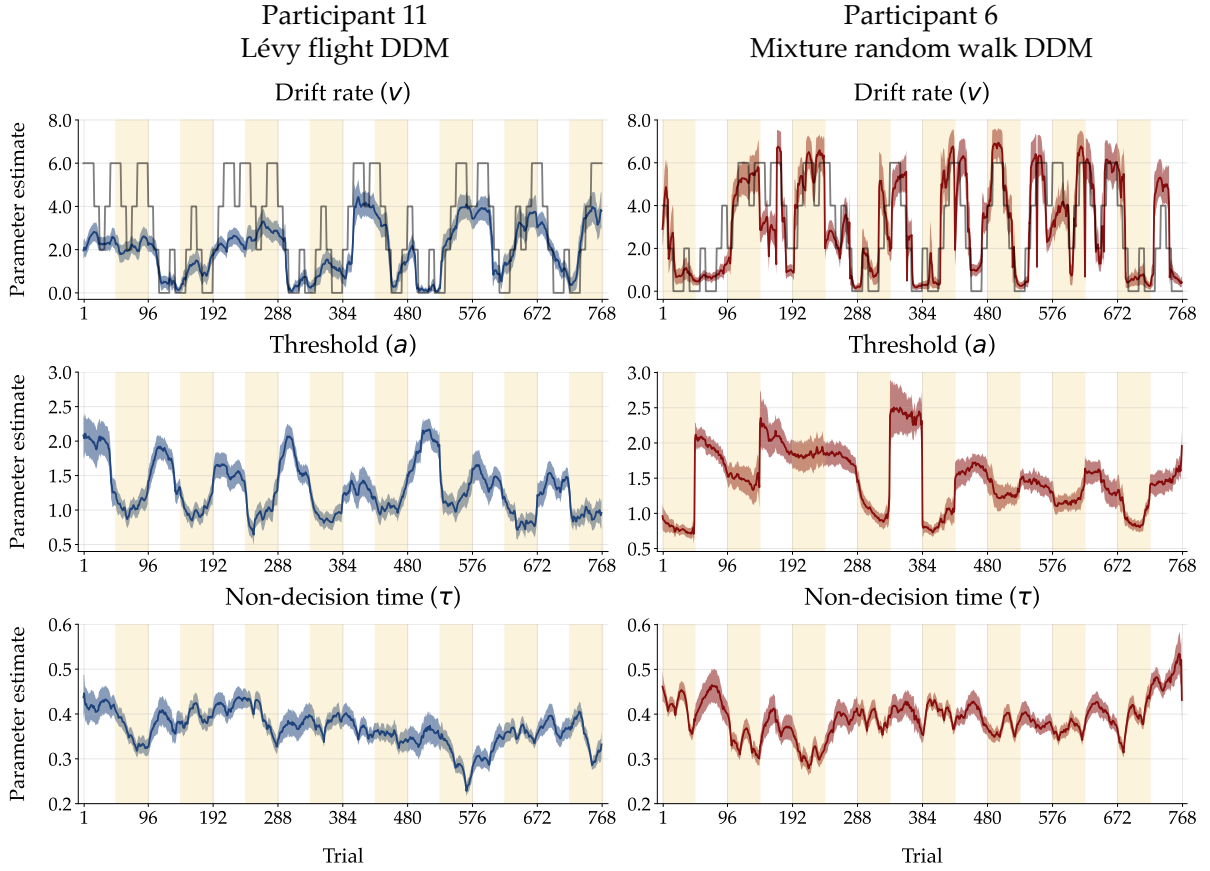


Figure 8: Estimated parameter trajectories of two exemplar individuals corresponding to the respective best-fitting non-stationary diffusion decision model (NSDDM). In this instance, the left column showcases results from a Lévy flight DDM, while the right column displays parameter trajectories from a mixture random walk DDM. Each low-level parameter (drift rate, threshold, and non-decision time) is displayed on a separate row. The solid lines are color-coded (blue for the Lévy flight DDM and red for the mixture random walk DDM) to represent the posterior medians, while the shaded regions mark the median absolute deviation (MAD). The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied. The sequences of task difficulty levels are depicted with black lines and overlaid with the drift rate in the top panels.

parameter values, the distinctions in average parameter values between difficulty levels are less pronounced when estimated with the random walk DDM.

With the second experimental manipulation - namely, the instruction to emphasize speed or accuracy - we aimed to manipulate the participants' decision caution, which is assumed to be captured by the threshold parameter. Examining the aggregated estimates of the threshold parameter in Figure 6D, we observe generally increased values in the accuracy condition compared to the speed condition. Interestingly, in the accuracy condition, the threshold parameter also slightly increases with growing task difficulty — a pattern not observed in the speed condition. A comparison between the estimates of the four NSDDMs reveals that the mixture random walk DDM and the Lévy flight DDM yield higher threshold estimates in the accuracy condition compared to the other two NSDDMs. Conversely, all four NSDDMs seem to converge in their threshold parameter estimates in the speed condition.

Parameter Trajectories For a more fine-grained analysis, particularly considering temporal aspects, we present the complete inferred parameter trajectories of the three low-level parameters of a NSDDM for two exemplary individuals in Figure 8. The Appendix contains the inferred parameter trajectories of the remaining 12 participants. Each partici-

participant's trajectory is depicted with the posterior median (solid lines) and the median absolute deviation (MAD, shaded bands) across all 768 experimental trials, estimated with the model with the highest posterior model probability for that specific individual. The trajectory of participant 11 corresponds to a Lévy flight DDM, whereas the trajectory of participant 6 comes from a mixture random walk DDM. Shaded blocks along the timeline denote the experimental condition at a given trial, with yellow indicating an emphasis on speed.

The top panel illustrates the estimated trajectories of the drift rate parameter alongside the sequences of task difficulty levels (depicted by black line). Here, 0 corresponds to the most difficult level, while 6 represents the easiest. It is important to note that the absolute values of the difficulty conditions hold no intrinsic meaning. As observed, the drift rates of both participants align with the overarching trend of the difficulty condition sequence. They decrease when the difficulty is high and increase as the task becomes easier.

Regarding the trajectory of the threshold parameter (middle panel), we anticipated that a shift from an accuracy instruction to a speed instruction would lead to a decrease in the threshold parameter, and *vice versa*. This hypothesized pattern is clearly evident when examining the estimated threshold parameter trajectories of the two participants in the middle panel of Figure 8. For instance, the threshold parameter estimated for participant 11 oscillates around an approximate value of 1 in the speed condition. Moreover, it consistently rises whenever a switch in the accuracy condition takes place. Intriguingly, the parameter's value during accuracy emphasis is not as uniform compared to the speed condition. In some blocks, it fluctuates around 2, while in others it hovers around 1.5 or even lower. Similarly, participant 6 displays pronounced shifts in the threshold parameter when a change in the condition occurs, with these shifts being more pronounced in the first half of the experiment and diminishing in the second half.

Finally, the bottom panel of Figure 8 illustrates the trajectory of the non-decision time parameter. Although our experimental manipulations did not systematically target the dynamics of this parameter, it is sometimes assumed that the manipulation of speed and accuracy instructions may also influence it (Arnold et al., 2015; Voss et al., 2004). While both individuals exhibit some fluctuations in τ , no systematic differences between the two conditions are apparent.

Upon reviewing the parameter trajectories of the remaining participants in the **Appendix**, similar patterns emerge. In summary, both the inferred means and trajectories of the drift rate and threshold parameters align with the sequence of experimental manipulations, as predicted by our design. Moreover, our NSDDMs were able to estimate these trajectories directly from the behavioral data, getting no explicit information whatsoever about the experimental context. Thus, our validation study suggests that NSDDMs are able to detect genuine changes in cognitive constructs.

Discussion

Psychology and cognitive science are witnessing a growing interest in incorporating dynamic aspects into mechanistic models that seek to formalize and explain cognitive processes. In a previous study, we explored a method to estimate plausible trajectories of cognitive process model parameters directly from behavioral data (Schumacher et al., 2023). Nevertheless, an empirical validation of this modeling approach was lacking. Thus, the current study sought to bridge this gap by experimentally examining the validity of the inferred diffusion decision model (DDM) parameter dynamics.

Experimental validation

The present study posed to the following core question: Can non-stationary DDMs (NSDDM) effectively detect experimentally induced changes in cognitive constructs from behavioral data alone? If this holds true, our findings can provide the first substantial evidence for the validity of the superstatistics framework as applied to cognitive models. Notably, our results demonstrated that the NSDDMs indeed reliably identified the sequence of two experimental manipulations, despite the absence of any contextual information. Moreover, posterior re-simulation revealed an outstanding fit to the response data, both on an aggregate level as well as on the level of the raw time series. This performance stands as compelling evidence supporting the validity of NSDDMs.

The trajectory of the drift rate parameter for all individuals closely mirrored the sequence of the task difficulty manipulation. Specifically, the drift rate parameter decreased when task difficulty increased, and conversely, increased as task difficulty decreased. This not only confirms the anticipated impact of the manipulation, but also highlights the NSDDMs' ability to discern these variations directly from the behavioral data, agnostic to additional contextual information.

Interestingly, drift rates increased throughout the experiment, although this was not the case for trials with the highest task difficulty. This observation suggests a practice effect among participants, where task performance generally improved with experience, except under the most challenging condition. Practice effects are a widely recognized phenomenon in various decision-making and memory paradigms (Forstmann et al., 2008; Healey & Kahana, 2014, 2016; Wagenmakers et al., 2008; Wynton & Anglim, 2017). In fact, practice effects have been studied with various dynamic cognitive modeling approaches (Evans et al., 2018; Evans & Hawkins, 2019; Gunawan et al., 2022; Kahana et al., 2018). A notable contribution to this field comes from Gunawan et al. (2022), who conducted a comprehensive re-analysis of three datasets derived from widely cited articles. Their study compared three dynamic models: (i) a smooth polynomial trend, (ii) a non-smooth autoregressive process, and (iii) a regime switching model instantiated by a hidden Markov model (HMM) with two different states.

In their study, Gunawan et al. (2022) employed a low-level model similar to the DDM, namely the linear ballistic accumulator model (LBA; Brown & Heathcote, 2008). However, their transition models, specifically the polynomial trend and the autoregressive process, differed in that they allowed LBA parameters to change only from block-to-block, neglecting trial-to-trial parameter fluctuations (except for the HMM). Their findings indicated that the HMM outperformed the other two dynamic model instantiations. This superiority can possibly be attributed to the model's capacity to flexibly change parameters from trial-to-trial, in contrast to changes occurring only from block-to-block. Even though the trial-by-trial specification of the HMM captures the microstructure of the decision-making process, it is still less flexible than the models we examined in the current study. HMMs assume a pre-defined number of possible states, whereas this is not the case with the implementation of our regime switching model. The advantage of not fixing the number of distinct states beforehand is particularly evident when the exact latent quantity is unknown prior to investigation. Moreover, results from our model comparison clearly favored transition models that account for both, gradual changes as well as sudden shifts. This suggests that regime-switching models may fall short in certain fields of application. Nevertheless, both models have their merits, and the choice between them should be guided by the specific research question at hand and formal model comparison.

As our study focused on experimentally validating parameter trajectories estimated with NSDDMs, we deliberately refrained from further analysing practice effects. However, we suggest that our flexible framework could be a promising alternative for investigating practice effects. Unlike pure regime switching models, it has the capacity to reveal a mixture of practice-related changes, ranging from abrupt shifts to gradual changes. When exploring substantive research questions, such as practice effects, with superstatistical models, it is imperative to depart from the approach taken in the current study. That is, one should always incorporate contextual information from the experimental setting when estimating parameter trajectories. Here the question arises, how to incorporate this information? In a previous study, we simply assumed separate low-level parameters for each experimental condition (Schumacher et al., 2023). This approach is particularly appropriate when conditions randomly change from trial to trial. However, future research could explore alternative ways of including experimental context information with the goal to further inform the parameters.

Concerning the second experimental manipulation, that is, the emphasis on speed or accuracy, their effect on the threshold parameter is more diverse across individuals. While a majority of participants demonstrated shifts in the threshold parameter in response to instructional changes, the consistency and magnitude of these changes varied significantly among individuals. Some participants exhibited only a few adjustments in the threshold parameters, seemingly overlooking the change in instruction on certain occasions. In contrast, others consistently heightened their threshold parameter during accuracy-focused tasks, followed by a subsequent decrease when transitioning to speed-oriented conditions. Meanwhile, some participants displayed rather unsystematic changes in decision caution, suggesting that the participants reacted differently to the speed-accuracy manipulation.

Kucharský et al. (2021) introduced a dynamic LBA incorporating a hidden Markov transition model with two states, akin to the model proposed by Gunawan et al. (2022). Their focus centered on scrutinizing the speed-accuracy trade-off, exploring the hypothesis that individuals dynamically switch between different operating states under varying instruction conditions. By fitting their model to previously collected data, they provided evidence that individuals tend to oscillate between two stable states: a deliberative, stimulus-driven mode emphasizing accuracy and sacrificing speed, and a guessing mode characterized by random and relatively faster choices.

However, our approach for estimating parameter trajectories reveals a more intricate scenario, challenging the assumed binary operational shift. Contrary to expectations, individuals manifest more than two discernible states. At times,

they exhibit an extreme adaptation to a change in condition, while at other times, they display little or no reaction to the altered condition. This complexity underscores the necessity for more flexible transition models, as employed in our study. Failing to utilize such adaptive models could potentially obscure the complex unfolding of individuals' cognition and behavior over time.

Model comparison

When implementing non-stationary models, a modeler encounters a myriad of options, ranging from various transition models to decisions about which parameter follows which transition model. In this study, we limited our choices to a small subset of the possibility space. Based on our experimental manipulations we anticipated that the DDM parameters, particularly the threshold parameter, would not only undergo gradual changes, but also manifest more abrupt shifts in response to changing conditions. Consequently, we tested different implementations accommodating such shifts (mixture random walk, Lévy flight, regime switching) against a transition model that does not, namely, the simple Gaussian random walk.

The inferred posterior model probabilities (PMPs) and Bayes factors (BFs) consistently favored the Lévy flight and occasionally the mixture random walk transition models. However, in terms of the absolute goodness-of-fit, as assessed through posterior re-simulations, the performance of all four NSDDMs showed remarkable similarity. This leads to two notable conclusions. First, even the models with lower PMPs demonstrated a good fit to the data, likely owing to the inherent flexibility of the superstatistical framework. Second, our Bayesian model comparison method could reliably detect the most favorable model even when the absolute differences were marginal.

Limitations

Psychological research is usually interested in some group or overall estimate of parameters. Thus, it would have been informative to compute and inspect "average" parameter trajectories. Unfortunately, our experiment was designed in a way that the difficulty and the speed-accuracy instruction manipulation was randomized across participants. This made it impossible to average the individual trajectories directly. Instead, we collapsed the estimates by the different experimental conditions and provided an aggregate view across individuals. Although this is certainly a limitation of this study, we argue that the current analysis is sufficient to address our specific research question.

Moreover, despite using many default settings from the `BayesFlow` software (Radev et al., 2023), the configuration and training of neural approximators for both parameter inference and model comparison for non-stationary models can still be a challenge. A basic understanding of deep learning principles and simulation-based inference is an essential prerequisite. These requirements may pose obstacles to the adoption of our method, highlighting the necessity for improved software and tutorials addressing these intricacies.

Outlook

Going forward, we see the relevance of our superstatistics framework as twofold. First, superstatistics could become a powerful tool in the methodological toolkit of the researcher interested in temporal changes in cognitive constructs. It is a general framework and provides large flexibility. Thus far, we only used the DDM as a low-level observation model. However, there are many other cognitive process models that could benefit from such a framework. For instance, reinforcement learning model parameters, such as the learning rate or the softmax temperature parameter, likely change over time (Li et al., 2023). Second, even when the temporal evolution of cognitive parameters is not a central research question, the adoption of non-stationary models may bring advantages over their stationary counterparts (Schumacher et al., 2023). Our analysis of estimated trajectories vividly illustrates discernible changes in parameters. Assuming stationarity would have led to misleading substantive conclusions.

With great flexibility comes a great plethora of choices. In this study, we compared different transition models guided by the contrast between gradual and sudden changes. However, there are more degrees of freedom when implementing superstatistical models, or Bayesian models in general (Gelman et al., 2020). Elsemüller, Olischläger, et al. (2023) advocates for the crucial role of sensitivity analysis, illustrating a potent methodology to facilitate informed decisions regarding factors such as the type and shape of prior distributions, neural network architectures, and other pivotal elements. We believe that using such an approach in the context of superstatistics could provide better guidelines for their implementation.

Up to this point, we focused on the estimates of the low-level parameter trajectories. Yet, it is crucial to note that we also obtain posterior distributions for the static high-level parameters. These estimates can also yield valuable insights into individuals' behavior and cognition. Depending on the chosen transition model, these estimates can offer indications of the frequency with which individuals transition between distinct operational states or the variability inherent in their cognitive constructs. Thus, analyzing these high-level parameters could constitute a compelling avenue for future research.

Conclusion

In conclusion, the experimental validation of non-stationary diffusion decision models presented in this study represents a significant step forward in the field of cognitive modeling. Our results provide compelling evidence that the estimated parameter trajectories genuinely reflect tangible changes in the targeted psychological constructs. We hope that our validation opens the door to widespread applications of non-stationary models in future modeling endeavors, offering a more nuanced understanding of cognitive processes across varying time scales.

Acknowledgments

We thank Steffen Ernst for his efforts in programming the experiment and collecting the data. L.S., M.S., and A.V. were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; grant number GRK 2277 "Statistical Modeling in Psychology").

Author contributions

L.S. and M.S. conceived the initial idea and the experimental design. L.S. and S.T.R. created and applied the models, and wrote the initial draft of the manuscript. A.V. supervised the project. All authors reviewed and refined the initial draft of the manuscript and agreed to its current version.

Data and Code Availability

All models, data, and scripts for reproducing the results of this work are publicly available in the project's GitHub repository <https://github.com/bayesflow-org/Non-Stationary-DDM-Validation>. The neural superstatistics method is implemented in the `BayesFlow` Python library for amortized Bayesian workflows (Radev et al., 2023).

References

- Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., & Köthe, U. (2018). Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*.
- Arnold, N. R., Bröder, A., & Bayen, U. J. (2015). Empirical validation of the diffusion model for recognition memory and a comparison of parameter-estimation methods. *Psychological Research, 79*(5), 882–898.
- Beck, C., & Cohen, E. G. D. (2003). Superstatistics. *Physica A: Statistical Mechanics and its Applications, 322*, 267–275.
- Beck, C. (2004). Superstatistics: Theory and applications. *Continuum mechanics and thermodynamics, 16*, 293–304.
- Beer, R. D. (2023). On the proper treatment of dynamics in cognitive science. *Topics in cognitive science*.
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153–178.
- Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115*(2), 396–425.
- Cochrane, A., Sims, C. R., Bejjanki, V. R., Green, C. S., & Bavelier, D. (2023). Multiple timescales of learning indicated by changes in evidence-accumulation processes during perceptual decision-making. *npj Science of Learning, 8*(1), 1–10.
- Craigmile, P. F., Peruggia, M., & Van Zandt, T. (2010). Hierarchical Bayes Models for Response Time Data. *Psychometrika, 75*(4), 613–632.
- Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences, 117*(48), 30055–30062.
- Else Müller, L., Olischläger, H., Schmitt, M., Bürkner, P.-C., Köthe, U., & Radev, S. T. (2023). Sensitivity-Aware Amortized Bayesian Inference.
- Else Müller, L., Schnuerch, M., Bürkner, P.-C., & Radev, S. T. (2023). A deep learning method for comparing bayesian hierarchical models.
- Esterman, M., & Rothlein, D. (2019). Models of sustained attention. *Current opinion in psychology, 29*, 174–180.
- Evans, N. J., & Brown, S. D. (2017). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review, 24*(2), 597–606.
- Evans, N. J., Brown, S. D., Mewhort, D. J. K., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review, 125*(4), 592–605.
- 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.
- Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review, 26*(4), 1099–1121.
- Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences, 105*(45), 17538–17542.
- Galadí, J. A., Silva Pereira, S., Sanz Perl, Y., Kringelbach, M. L., Gayte, I., Laufs, H., Tagliazucchi, E., Langa, J. A., & Deco, G. (2021). Capturing the non-stationarity of whole-brain dynamics underlying human brain states. *NeuroImage, 244*, 118551.
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.
- Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation, 12*(10), 2451–2471.
- Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., Chintaluri, C., Podlaski, W. F., Haddad, S. A., Vogels, T. P., et al. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife, 9*, e56261.
- Greenberg, D., Nonnenmacher, M., & Macke, J. (2019). Automatic posterior transformation for likelihood-free inference. *International Conference on Machine Learning, 2404–2414*.
- Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of mathematical psychology, 81*, 80–97.
- Gunawan, D., Hawkins, G. E., Kohn, R., Tran, M.-N., & Brown, S. D. (2022). Time-evolving psychological processes over repeated decisions. *Psychological review, 129*(3), 438.

- Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*(2), 575–596.
- Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, *123*(1), 23–69.
- Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Lepplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., ... Hoijsink, H. (2023). A review of applications of the Bayes factor in psychological research. *Psychological Methods*, *28*(3), 558–579.
- Kahana, M. J., Aggarwal, E. V., & Phan, T. D. (2018). The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1857–1863.
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.
- Kucharský, Š., Tran, N.-H., Veldkamp, K., Raijmakers, M., & Visser, I. (2021). Hidden Markov Models of Evidence Accumulation in Speeded Decision Tasks. *Computational Brain & Behavior*, *4*(4), 416–441.
- Lerche, V., & Voss, A. (2018). Speed–accuracy manipulations and diffusion modeling: Lack of discriminant validity of the manipulation or of the parameter estimates? *Behavior Research Methods*, *50*(6), 2568–2585.
- Lerche, V., & Voss, A. (2019). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychological Research*, *83*(6), 1194–1209.
- Li, J.-J., Shi, C., Li, L., & Collins, A. (2023). Dynamic noise estimation: A generalized method for modeling noise in sequential decision-making behavior. *bioRxiv*.
- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.
- Mark, C., Metzner, C., Lautscham, L., Strissel, P. L., Strick, R., & Fabry, B. (2018). Bayesian model selection for complex dynamic systems. *Nature Communications*, *9*(1), 1803.
- McDougle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, *28*(1), 20–39.
- Melanson, A., Mejias, J. F., Jun, J. J., Maler, L., & Longtin, A. (2017). Nonstationary Stochastic Dynamics Underlie Spontaneous Transitions between Active and Inactive Behavioral States. *eNeuro*, *4*(2).
- Miletić, S., Boag, R. J., Trutti, A. C., Stevenson, N., Forstmann, B. U., & Heathcote, A. (2021). A new model of decision processing in instrumental learning tasks (V. Wyart, J. I. Gold, & J. W. de Gee, Eds.). *eLife*, *10*, e63055.
- Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology*, *104*, 106–142.
- Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, *22*(1), 2617–2680.
- Radev, S. T., D’Alessandro, M., Mertens, U. K., Voss, A., Köthe, U., & Bürkner, P.-C. (2020). Amortized Bayesian Model Comparison With Evidential Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 4903–4917.
- Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2020). Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*, *33*(4), 1452–1466.
- Radev, S. T., Schmitt, M., Schumacher, L., Else Müller, L., Pratz, V., Schälte, Y., Köthe, U., & Bürkner, P.-C. (2023). Bayesflow: Amortized bayesian workflows with neural networks. *Journal of Open Source Software*, *8*(89), 5702.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.
- Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural computation*, *20*(4), 873–922.
- Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356.
- Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.
- Ratcliff, R., & Van Dongen, H. P. A. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences*, *108*(27), 11285–11290.

- Säilynoja, T., Bürkner, P.-C., & Vehtari, A. (2022). Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison. *Statistics and Computing*, 32(2), 32.
- Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press.
- Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled bayesian workflow in cognitive science. *Psychological methods*, 26(1), 103.
- Schumacher, L., Bürkner, P.-C., Voss, A., Köthe, U., & Radev, S. T. (2023). Neural superstatistics for Bayesian estimation of dynamic cognitive models. *Scientific Reports*, 13(1), 13778.
- Schumacher, L., & Voss, A. (2023). Duration discrimination: A diffusion decision modeling approach. *Attention, Perception, & Psychophysics*, 85(2), 560–577.
- Sebastian Castro-Alvarez, R. R. M., Laura F. Bringmann, & Tendeiro, J. N. (2023). A time-varying dynamic partial credit model to analyze polytomous and multivariate time series data. *Multivariate Behavioral Research*, 0(0), 1–20.
- Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2020). Validating Bayesian Inference Algorithms with Simulation-Based Calibration.
- Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, 24(6), 774–795.
- Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, 132(3), 331–350.
- van Ravenzwaaij, D., & Wagenmakers, E.-J. (2022). Advantages masquerading as “issues” in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). *Psychological Methods*, 27(3), 451–465.
- Verdonck, S., Loossens, T., & Philiastides, M. G. (2021). The Leaky Integrating Threshold and its impact on evidence accumulation models of choice response time (RT). *Psychological Review*, 128(2), 203–221.
- Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology. *Experimental psychology*.
- Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, 32(7), 1206–1220.
- Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of $1/f^\alpha$ noise in human cognition. *Psychonomic Bulletin & Review*, 11(4), 579–615.
- 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.
- Walsh, M. M., Gunzelmann, G., & Van Dongen, H. P. A. (2017). Computational cognitive modeling of the temporal dynamics of fatigue from sleep loss. *Psychonomic Bulletin & Review*, 24(6), 1785–1807.
- Weissman, D. H., Roberts, K., Visscher, K., & Woldorff, M. (2006). The neural bases of momentary lapses in attention. *Nature neuroscience*, 9(7), 971–978.
- Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023). Transformers in Time Series: A Survey.
- Wynton, S. K. A., & Anglim, J. (2017). Abrupt strategy change underlies gradual performance change: Bayesian hierarchical models of component and aggregate strategy use. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(10), 1630–1642.

Appendix

S1 Appendix. Prior distributions

In the following we list the prior distributions we used for all four NSDDM's.

DDM Starting Values

For the starting values of the parameter trajectories we used half-normal distributions with a mean μ and a standard deviation σ denoted as $\mathcal{HN}(\mu, \sigma)$:

$$\begin{aligned} v_0 &\sim \mathcal{HN}(2.0, 2.0) \\ a_0 &\sim \mathcal{HN}(2.0, 1.5) \\ \tau_0 &\sim \mathcal{HN}(0.3, 1.0) \end{aligned}$$

Random Walk Transition Model

Half-normal distributions were used for the standard deviations of the Gaussian random walk transition model:

$$\begin{aligned} \sigma_v &\sim \mathcal{HN}(0.0, 0.1) \\ \sigma_a &\sim \mathcal{HN}(0.0, 0.1) \\ \sigma_\tau &\sim \mathcal{HN}(0.0, 0.01) \end{aligned}$$

We decided to use a relatively narrower prior on σ_τ because the non-decision time parameter is not expected to fluctuate as heavily as the other two parameters.

Mixture Random Walk Transition Model

The mixture random walk transition model used the same prior for the Gaussian random walk as described above. Additionally, Uniform distributions denoted as \mathcal{U} were used for the mixture proportion parameter ρ :

$$\begin{aligned} \rho_v &\sim \mathcal{U}(0.0, 0.2) \\ \rho_a &\sim \mathcal{U}(0.0, 0.1) \end{aligned}$$

Levy Flight Transition Model

The Levy flight transition model uses an alpha stable distribution instead of a Gaussian distribution for the transition. We used the same priors for the standard deviations as in the random walk and the mixture random walk. The alpha stable distribution has an additional parameter α , which determines the fatness of the tails. This parameter is bound between 1 and 2. Therefore, we used a Beta distribution denoted as \mathbf{B} and added 1 to the sampled values:

$$\begin{aligned} \tilde{\alpha}_v &\sim \mathbf{B}(1.5, 1.5) \\ \tilde{\alpha}_a &\sim \mathbf{B}(2.5, 1.5) \\ \alpha_v &= \tilde{\alpha}_v + 1 \\ \alpha_a &= \tilde{\alpha}_a + 1 \end{aligned}$$

Regime Switching Transition Model

The same prior distributions as for the mixture random walk were used for the mixture probabilities of the regime switching transition model.

S2 Appendix. Neural network architectures and training setups

In the following, we outline our implementation of the neural approximators and the training setup used for model comparison and parameter estimation.

Model comparison

For model comparison we trained an ensemble of ten neural approximators. Each approximator consists of a summary network and an inference network. The summary network is a many-to-one transformer architecture for time series encoding (Wen et al., 2023). The time series transformer has 128 template and 64 summary dimensions. For inference, we use a network that approximates posterior model probabilities (PMPs) as employed in Elsemüller, Schnuerch, et al. (2023).

We performed offline training for each of the ten neural approximators separately. The training data consisted of 25 000 simulations per model. Training was performed with 25 epochs and a batch size of 16 starting with an initial learning rate of 0.0005. The learning rate was adjusted with a cosine decay from its initial value to 0.

Parameter estimation

For parameter estimation we trained one neural approximator for each of the four NSDDM implementations. Each approximator consists of a hierarchical summary network as employed in Elsemüller, Schnuerch, et al. (2023) and two inference networks. Three bidirectional long-short term memory (LSTM) networks were used for the hierarchical summary network. The number of hidden units were 512, 256, and 128 respectively.

For inference, we use a composition of two invertible neural networks (Radev, Mertens, et al., 2020), one for the low-level and one for the high-level parameters. The network for the low-level parameters has 8 coupling layers with an interleaved *affine* and *spline* internal coupling design. The network for the high-level parameters only differs from the former in its number of coupling layers which is 6.

Since our simulators can be run fast, the training of the four neural approximators was performed online, with 75 epochs, 1 000 iterations per epoch, and a batch size of 16. Thus, each approximator was trained on $N = 1\,200\,000$ simulated data sets. The initial learning rate was set to 0.0005 and was reduced with a cosine decay function to 0.

S3 Appendix. Individual analyses

The following section shows the individual specific posterior re-simulations and parameter estimates for each difficulty level and both conditions separately. The visualizations are constructed in the vain of Figure 6 in the main text.

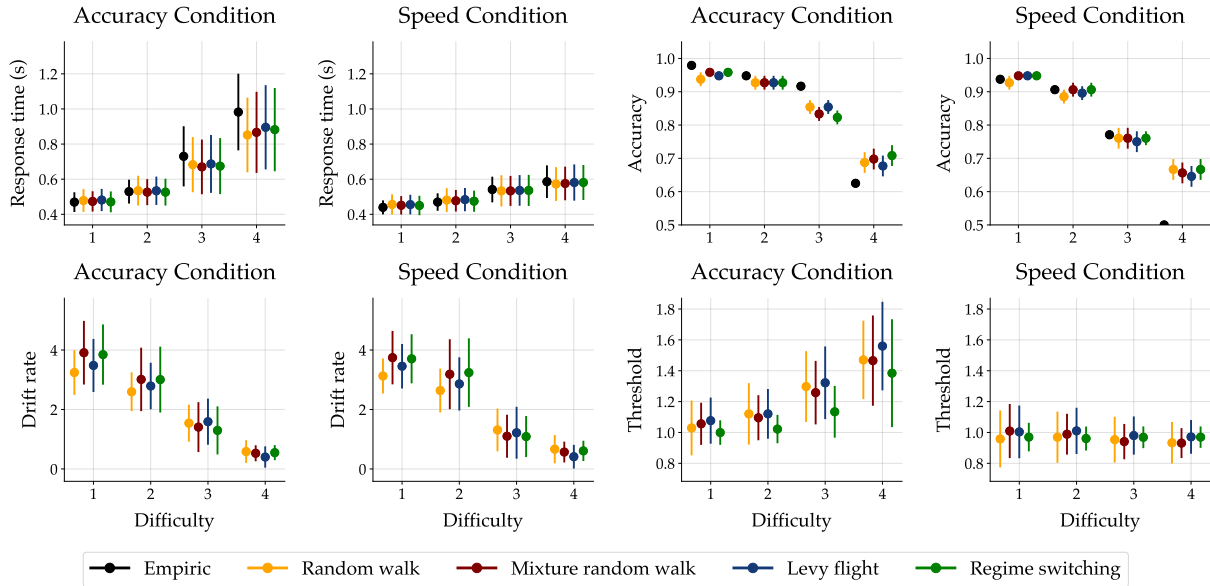


Figure 9: Aggregate results from all models fitted to the data from participant 1. The top row illustrates posterior re-simulations as a measure of the model’s generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

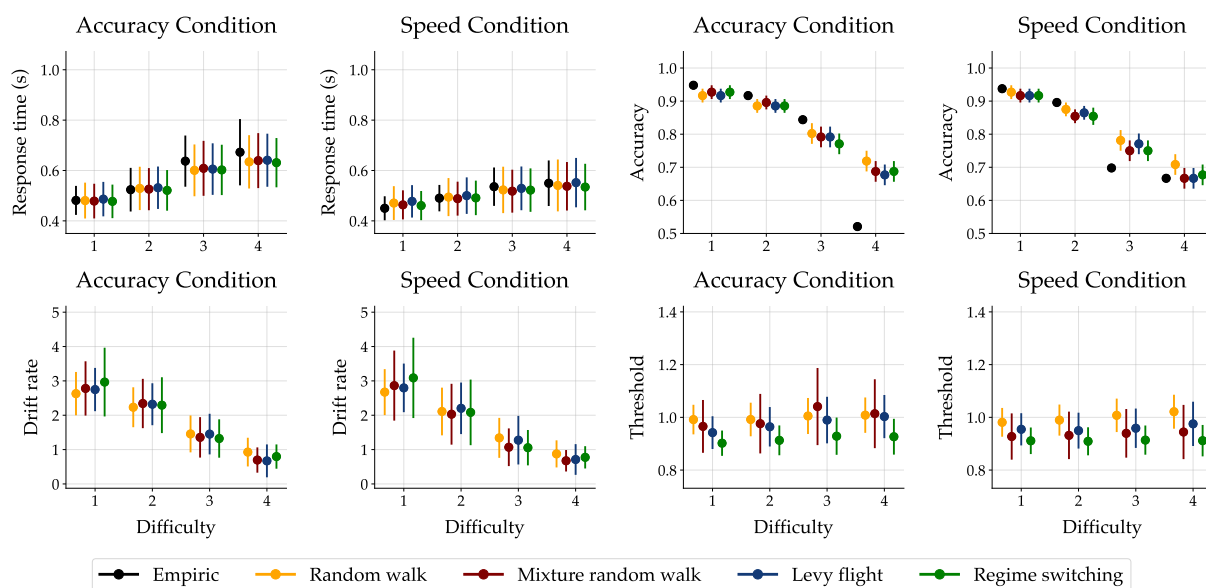


Figure 10: Aggregate results from all models fitted to the data from participant 2. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

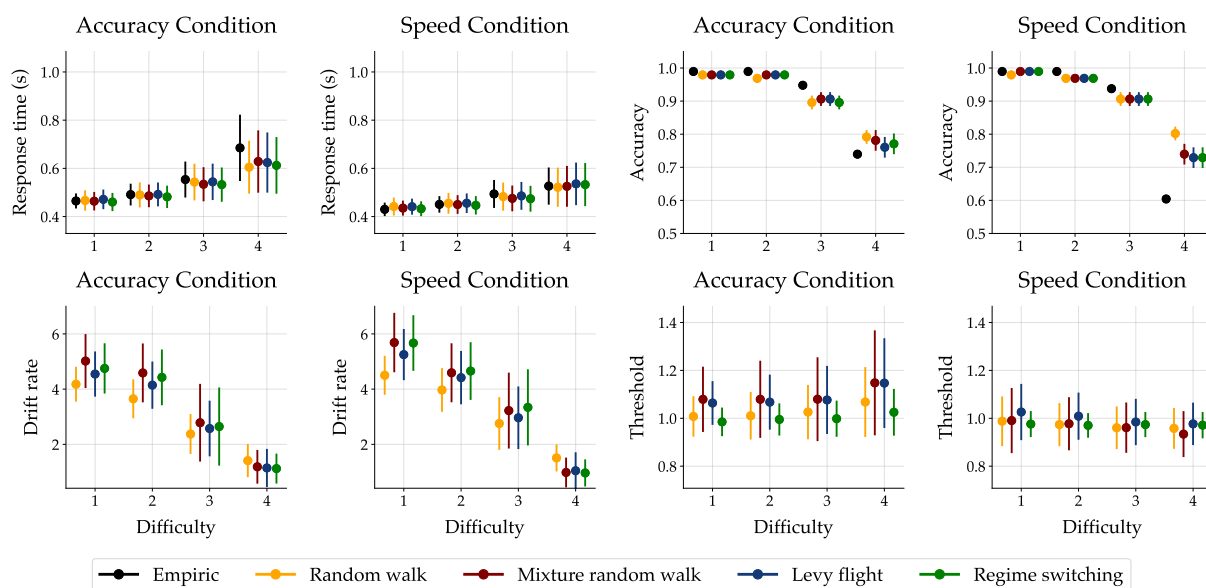


Figure 11: Aggregate results from all models fitted to the data from participant 3. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

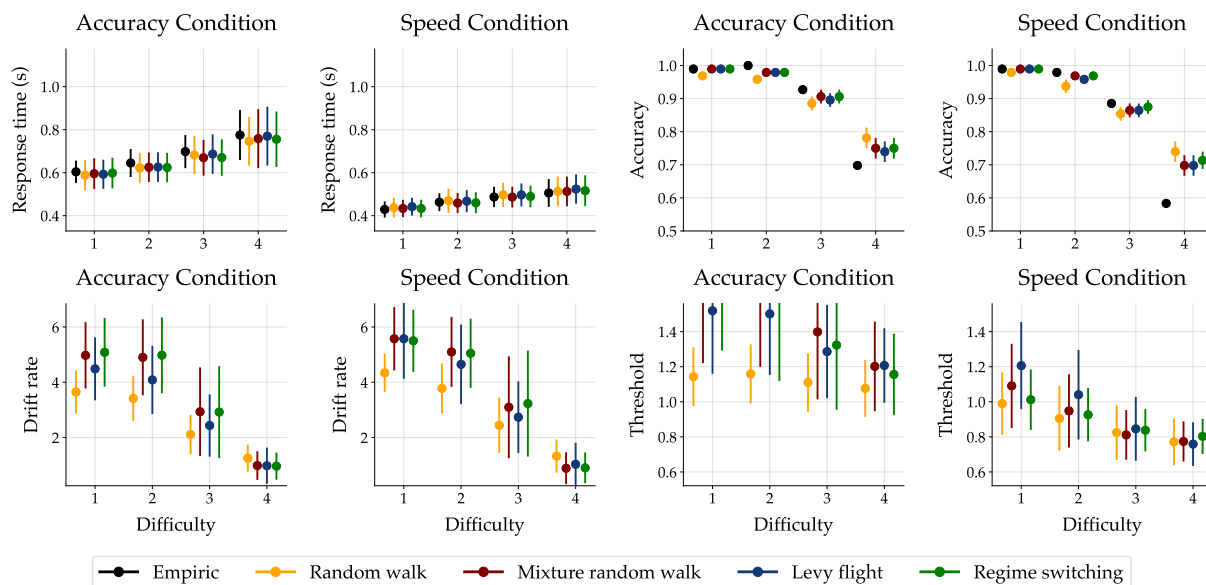


Figure 12: Aggregate results from all models fitted to the data from participant 4. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

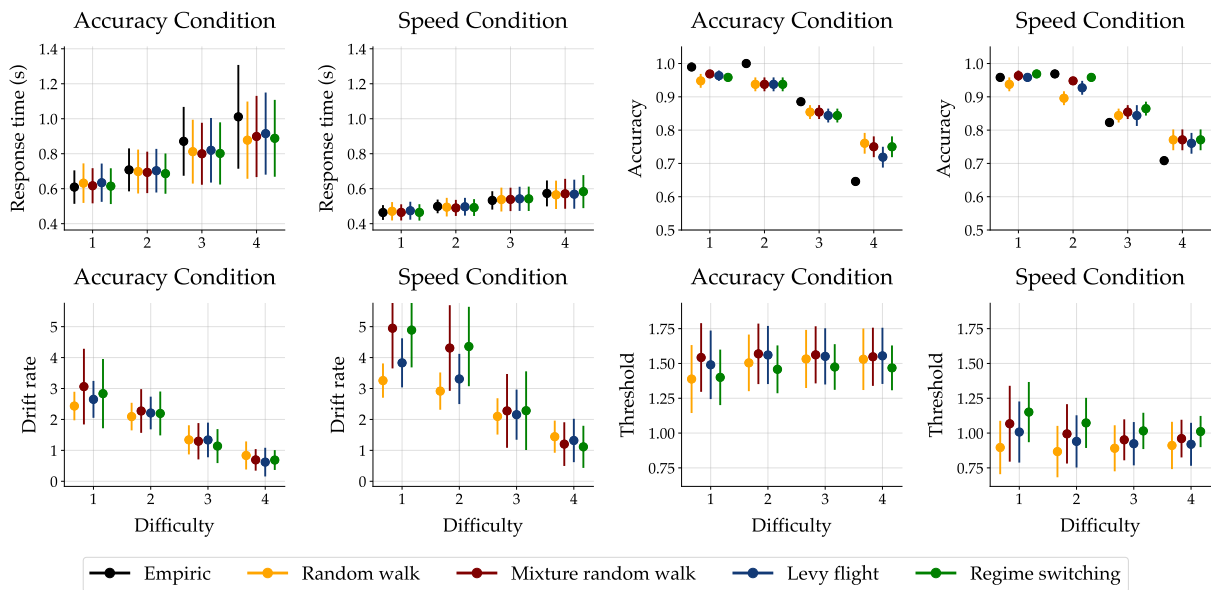


Figure 13: Aggregate results from all models fitted to the data from participant 5. The top row illustrates posterior re-simulations as a measure of the model’s generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

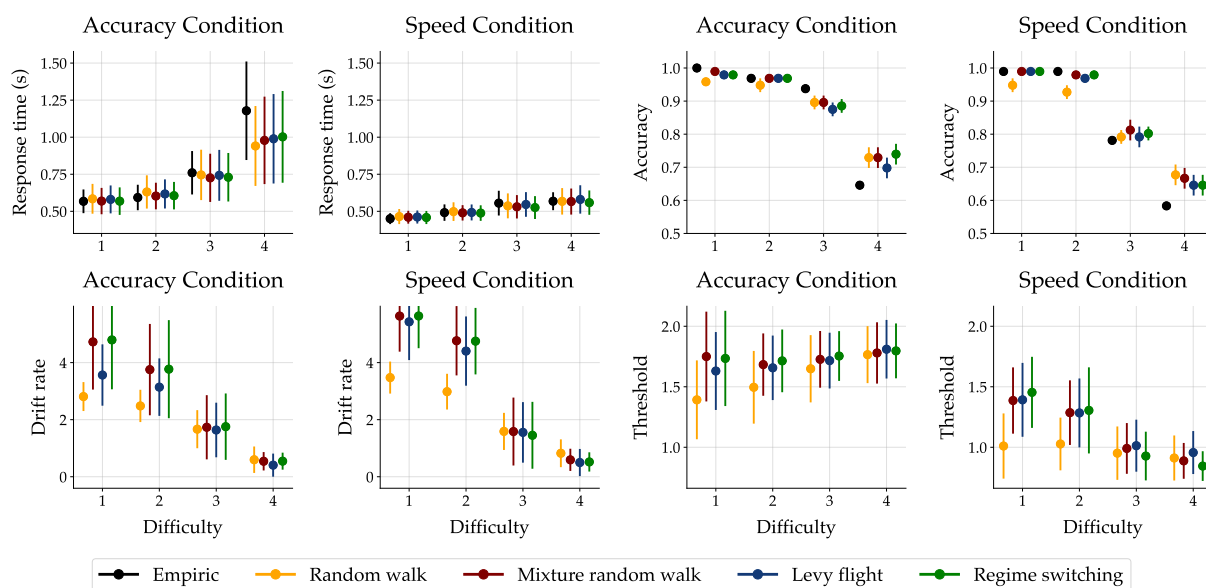


Figure 14: Aggregate results from all models fitted to the data from participant 6. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

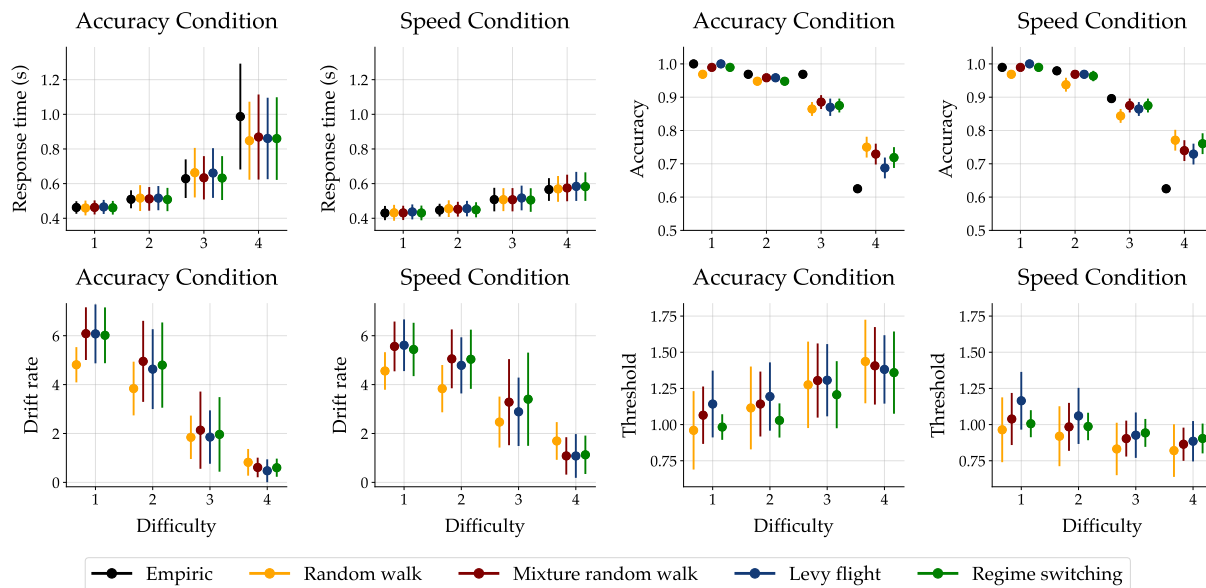


Figure 15: Aggregate results from all models fitted to the data from participant 7. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

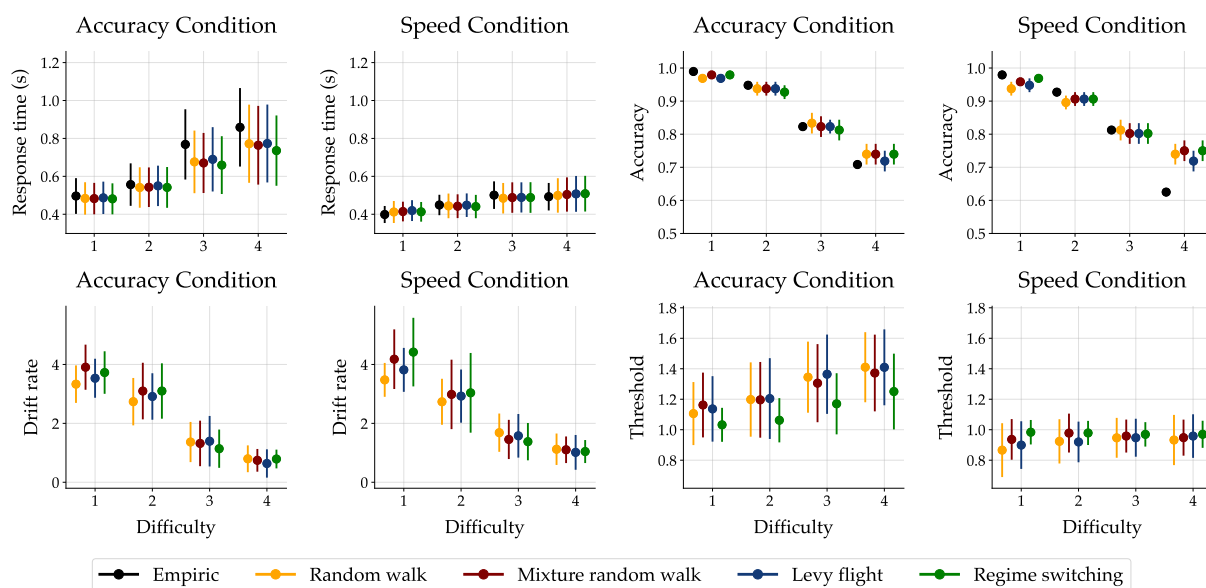


Figure 16: Aggregate results from all models fitted to the data from participant 8. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

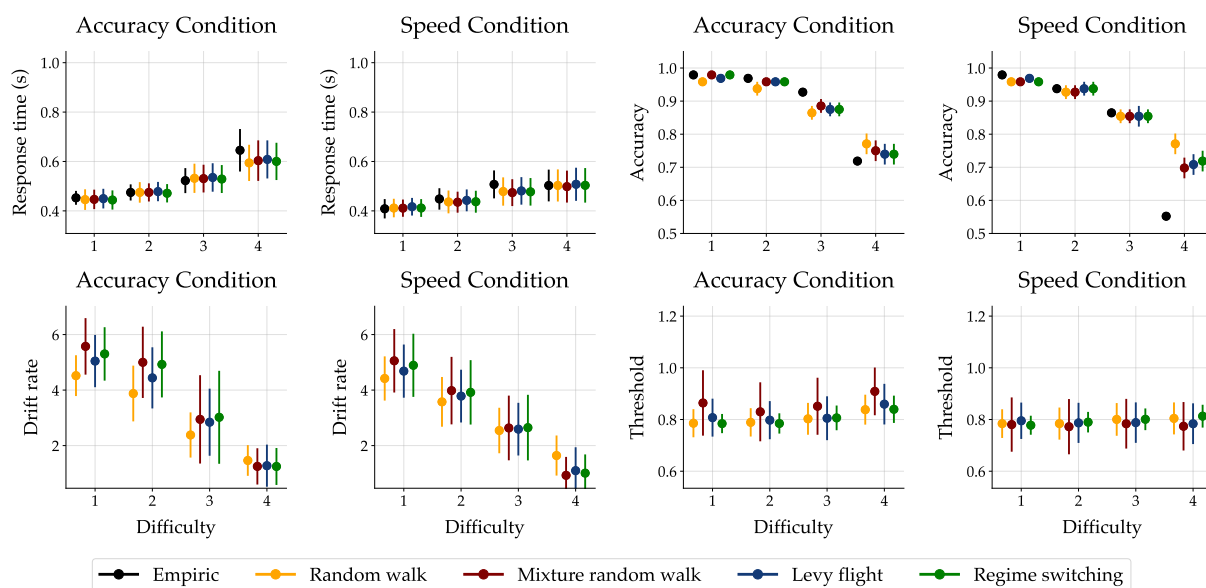


Figure 17: Aggregate results from all models fitted to the data from participant 9. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

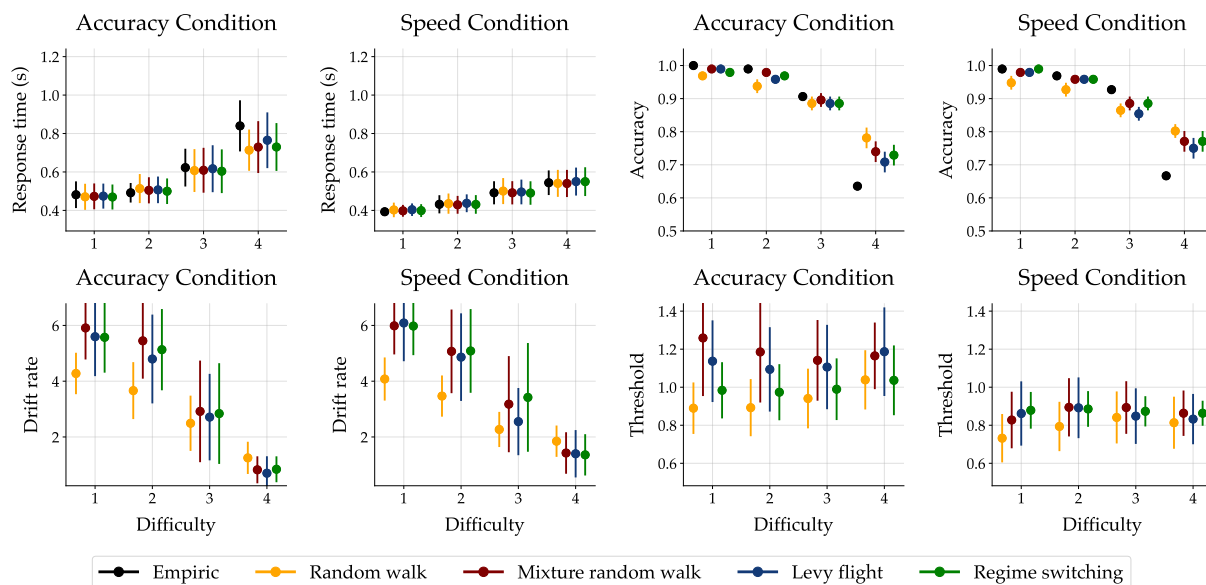


Figure 18: Aggregate results from all models fitted to the data from participant 10. The top row illustrates posterior re-simulations as a measure of the model’s generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

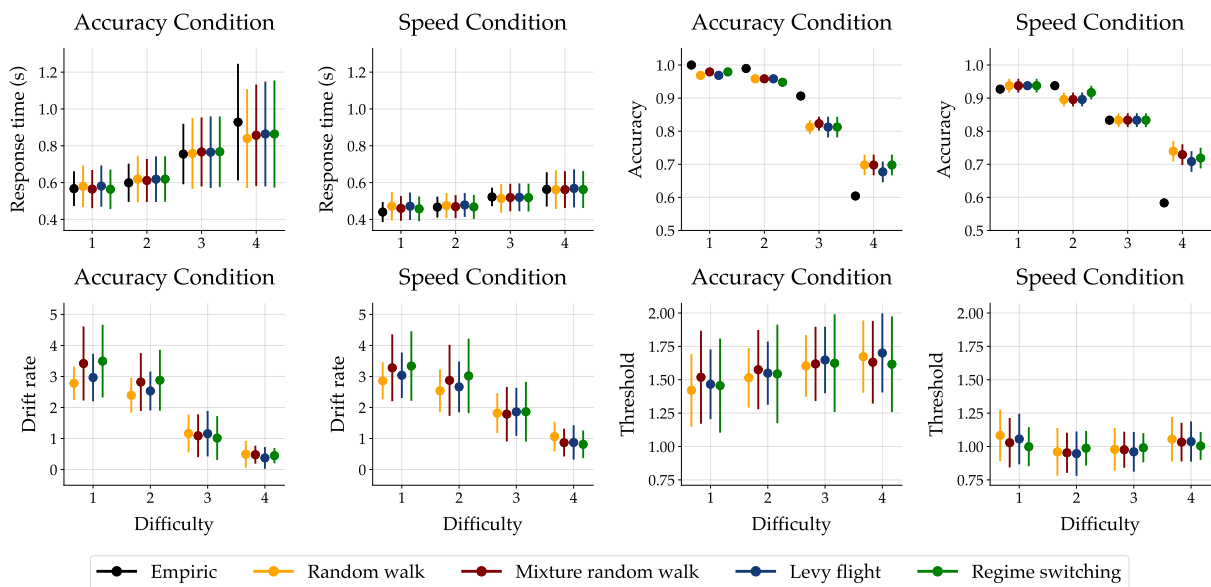


Figure 19: Aggregate results from all models fitted to the data from participant 11. The top row illustrates posterior re-simulations as a measure of the model’s generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

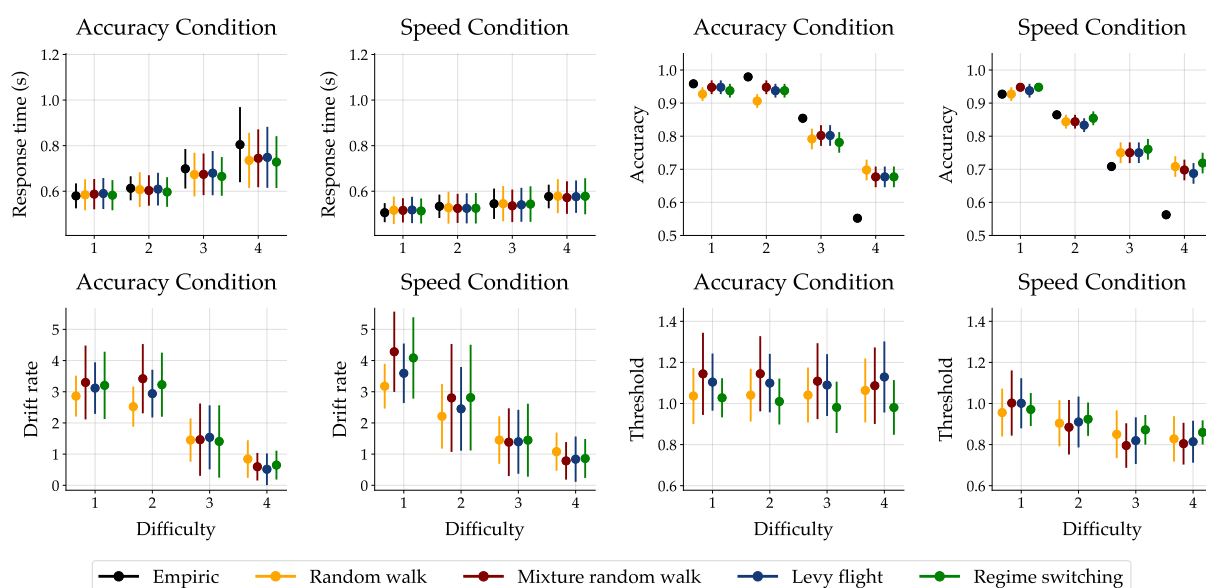


Figure 20: Aggregate results from all models fitted to the data from participant 12. The top row illustrates posterior re-simulations as a measure of the model’s generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

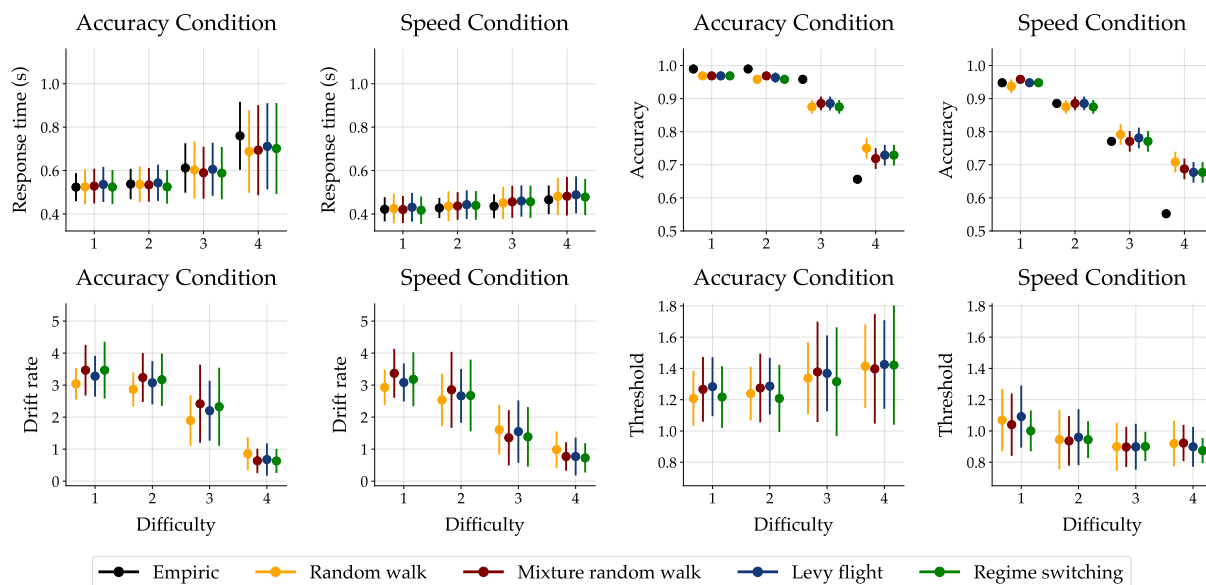


Figure 21: Aggregate results from all models fitted to the data from participant 13. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

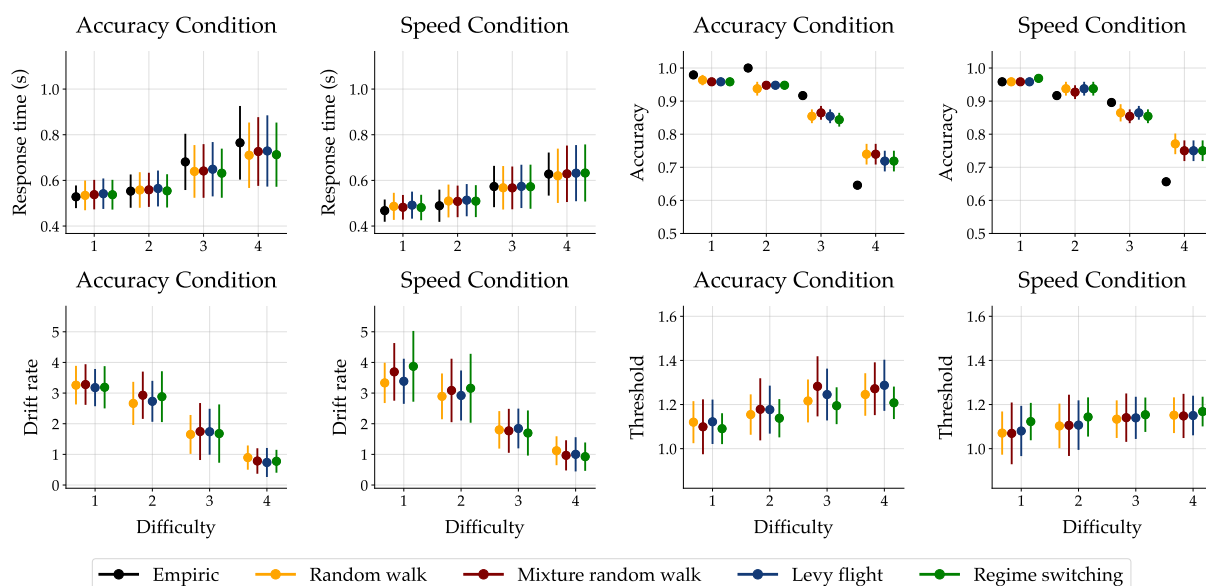


Figure 22: Aggregate results from all models fitted to the data from participant 14. The top row illustrates posterior re-simulations as a measure of the model’s generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter from the non-stationary diffusion decision models (NSDDM). **A** Empirical and re-simulated response times for each difficulty level and both conditions. **B** Empirical and re-simulated proportions of correct choices (accuracy) for each difficulty level and both conditions separately. **C** Posterior estimates of the drift rate parameter for each difficulty level and both conditions separately. **D** Posterior estimates of the threshold parameter for each difficulty level and both conditions separately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individual data and re-simulations.

S4 Appendix. Response time time series

In the following, we present the model fit to the whole response time time series for the remaining 12 participants.

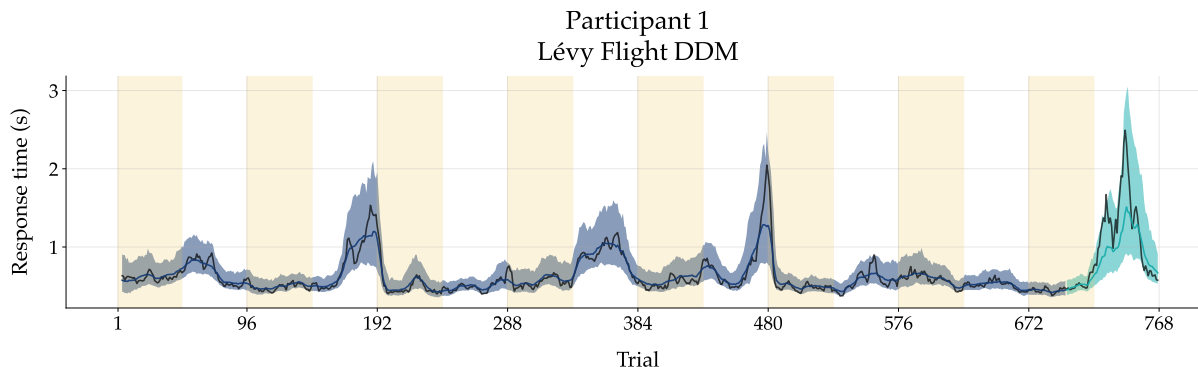


Figure 23: Model fit to response time (RT) time series. The empirical RT time series of participant 1 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in blue. In this instance, the results stem from a Lévy flight DDM. For the remaining trials, one-step-ahead predictions are depicted in cyan. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

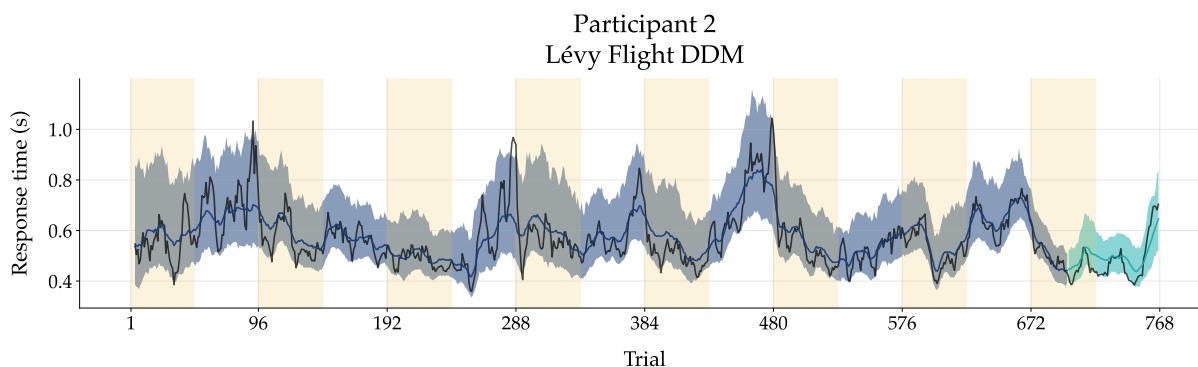


Figure 24: Model fit to response time (RT) time series. The empirical RT time series of participant 2 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in blue. In this instance, the results stem from a Lévy flight DDM. For the remaining trials, one-step-ahead predictions are depicted in cyan. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

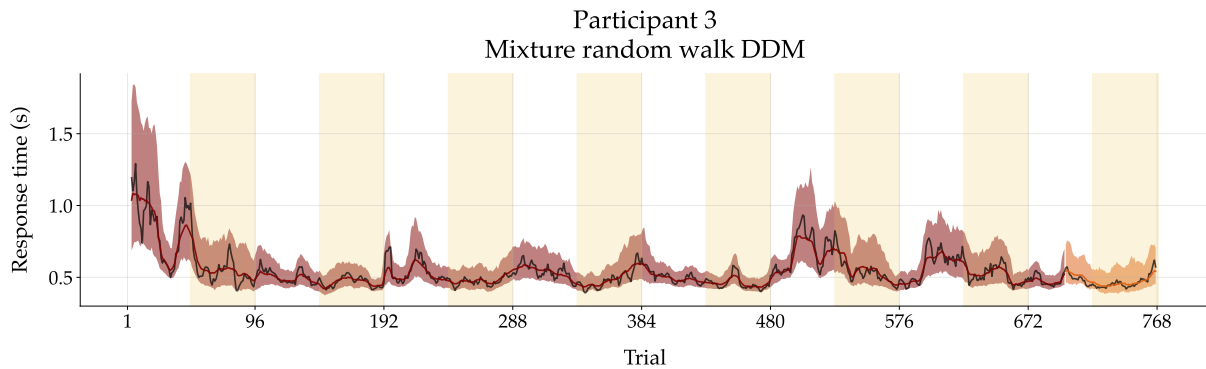


Figure 25: Model fit to response time (RT) time series. The empirical RT time series of participant 3 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in red. In this instance, the results stem from a mixture random walk DDM. For the remaining trials, one-step-ahead predictions are depicted in orange. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

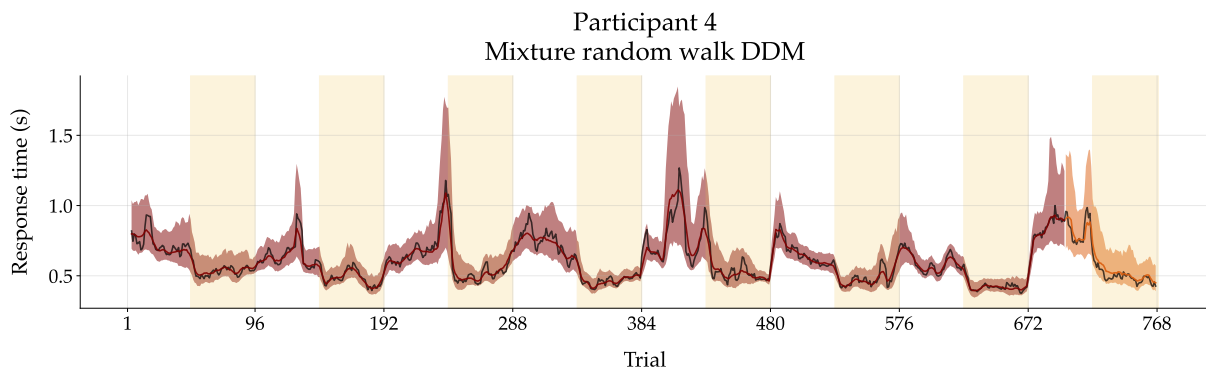


Figure 26: Model fit to response time (RT) time series. The empirical RT time series of participant 4 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in red. In this instance, the results stem from a mixture random walk DDM. For the remaining trials, one-step-ahead predictions are depicted in orange. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

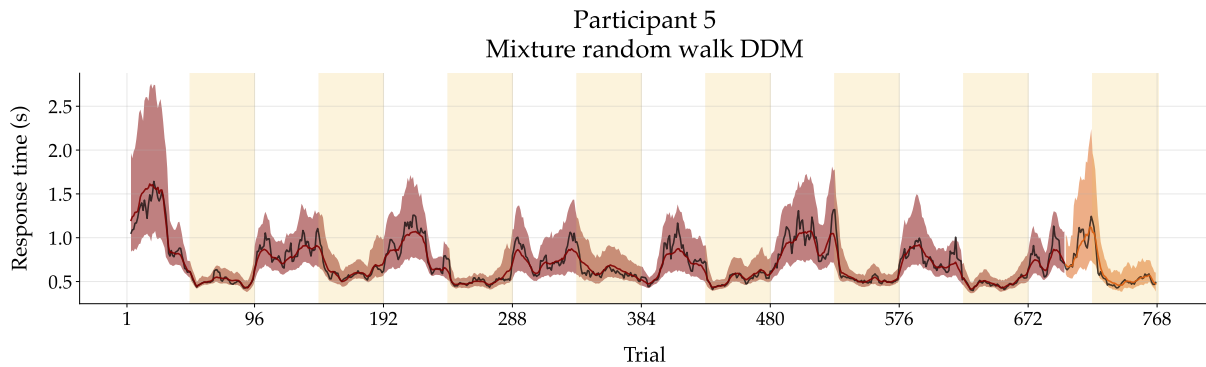


Figure 27: Model fit to response time (RT) time series. The empirical RT time series of participant 5 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in red. In this instance, the results stem from a mixture random walk DDM. For the remaining trials, one-step-ahead predictions are depicted in orange. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

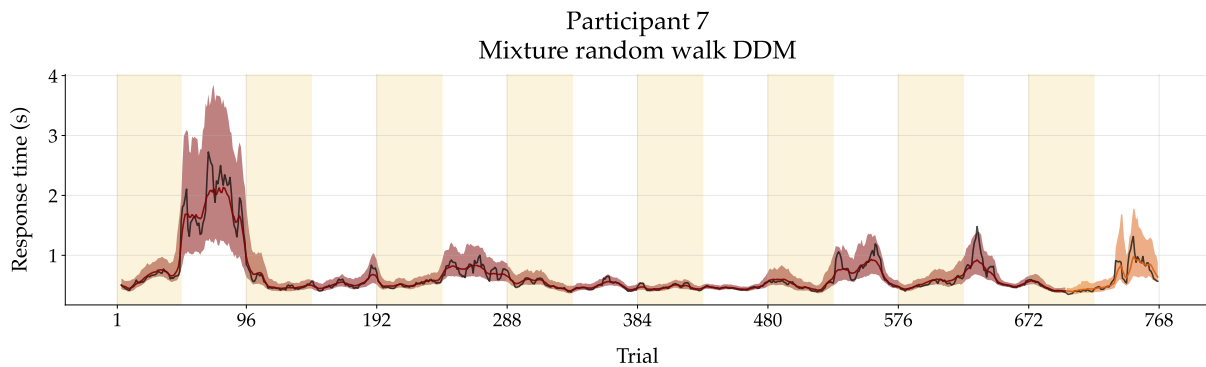


Figure 28: Model fit to response time (RT) time series. The empirical RT time series of participant 7 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in red. In this instance, the results stem from a mixture random walk DDM. For the remaining trials, one-step-ahead predictions are depicted in orange. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

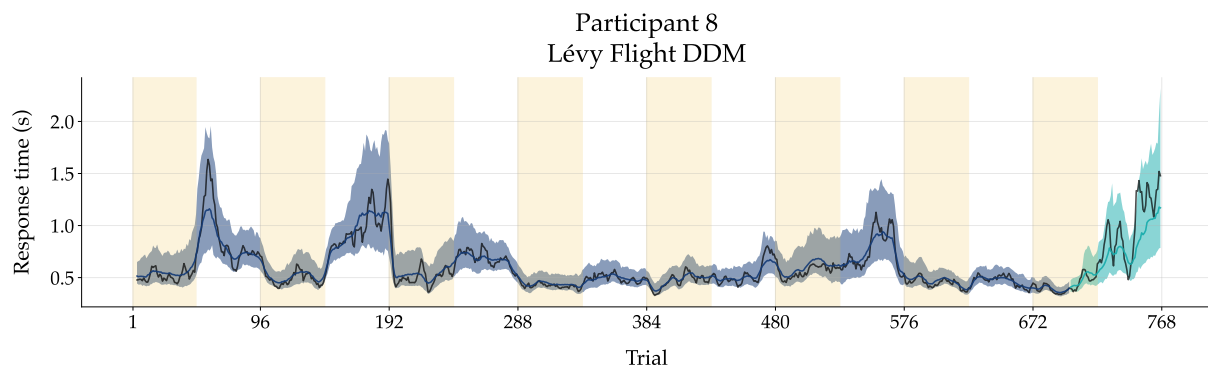


Figure 29: Model fit to response time (RT) time series. The empirical RT time series of participant 8 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in blue. In this instance, the results stem from a Lévy flight DDM. For the remaining trials, one-step-ahead predictions are depicted in cyan. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

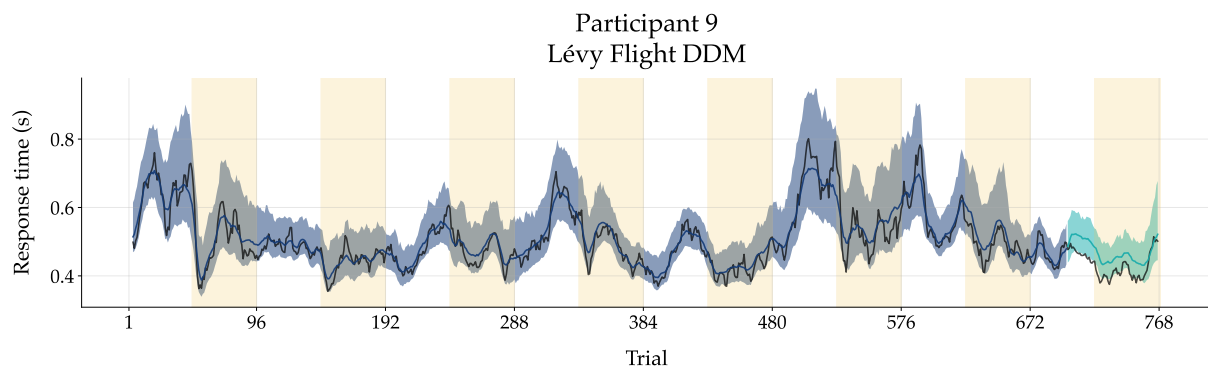


Figure 30: Model fit to response time (RT) time series. The empirical RT time series of participant 9 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in blue. In this instance, the results stem from a Lévy flight DDM. For the remaining trials, one-step-ahead predictions are depicted in cyan. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

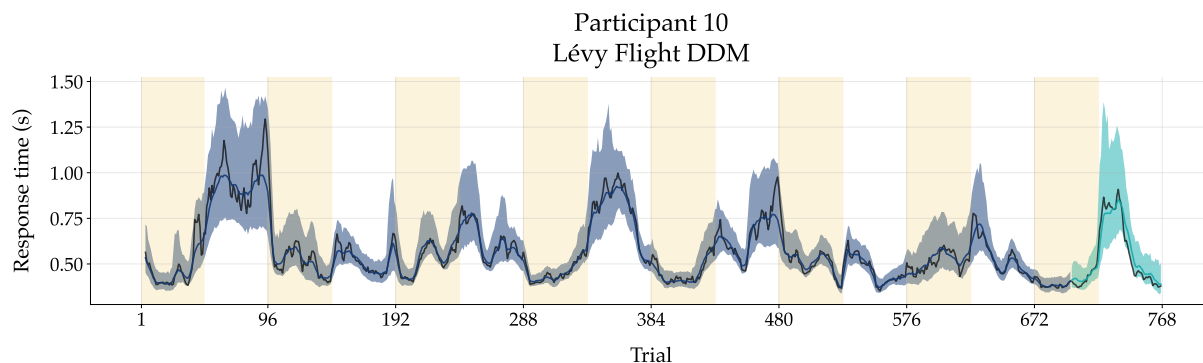


Figure 31: Model fit to response time (RT) time series. The empirical RT time series of participant 10 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in blue. In this instance, the results stem from a Lévy flight DDM. For the remaining trials, one-step-ahead predictions are depicted in cyan. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

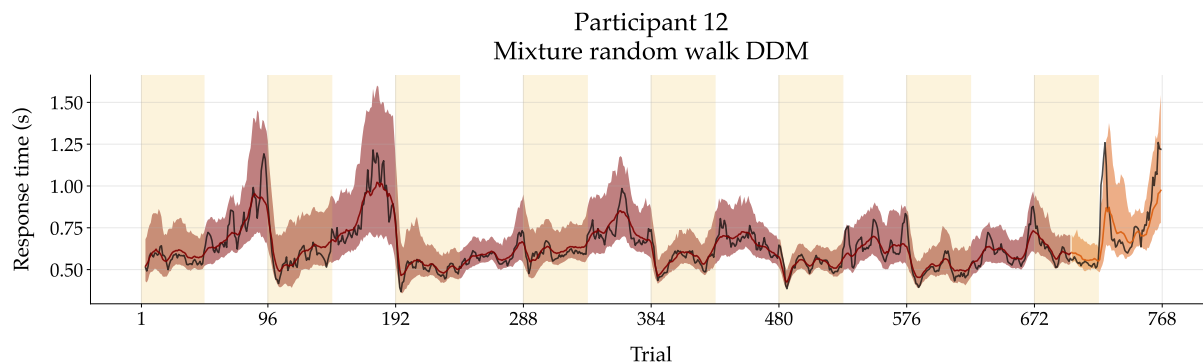


Figure 32: Model fit to response time (RT) time series. The empirical RT time series of participant 7 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in red. In this instance, the results stem from a mixture random walk DDM. For the remaining trials, one-step-ahead predictions are depicted in orange. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

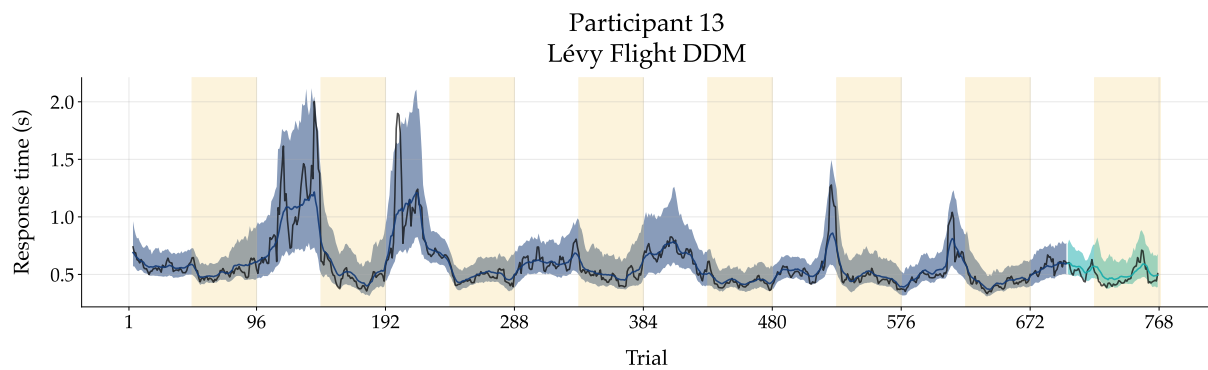


Figure 33: Model fit to response time (RT) time series. The empirical RT time series of participant 13 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in blue. In this instance, the results stem from a Lévy flight DDM. For the remaining trials, one-step-ahead predictions are depicted in cyan. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

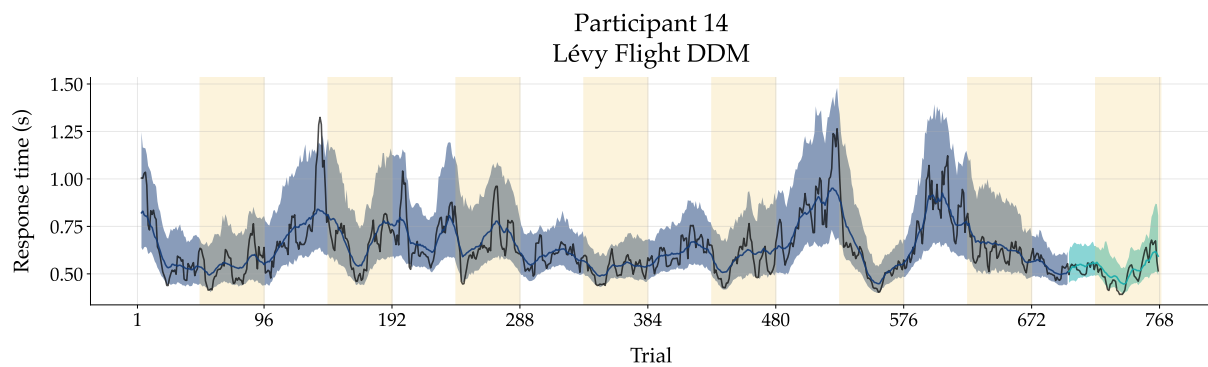


Figure 34: Model fit to response time (RT) time series. The empirical RT time series of participant 14 is shown in black. From trial 1 to 700, the posterior re-simulation (aka retrodictive check) using the best fitting non-stationary diffusion decision model (NSDDM) for this specific individual are shown in blue. In this instance, the results stem from a Lévy flight DDM. For the remaining trials, one-step-ahead predictions are depicted in cyan. Solid lines correspond to the median and shaded bands to 90% credibility intervals (CI). The empirical, re-simulated, and predicted RT time series were smoothed via a simple moving average (SMA) with a period of 5. The yellow shaded regions indicate trials where speed was emphasised over accuracy, while blank white areas denote instances where the opposite emphasis was applied.

S5 Appendix. Parameter trajectories

In the following, we present the inferred parameter trajectories for the remaining participants. For each visualisation the model with the highest posterior model probability for that specific individual was used.

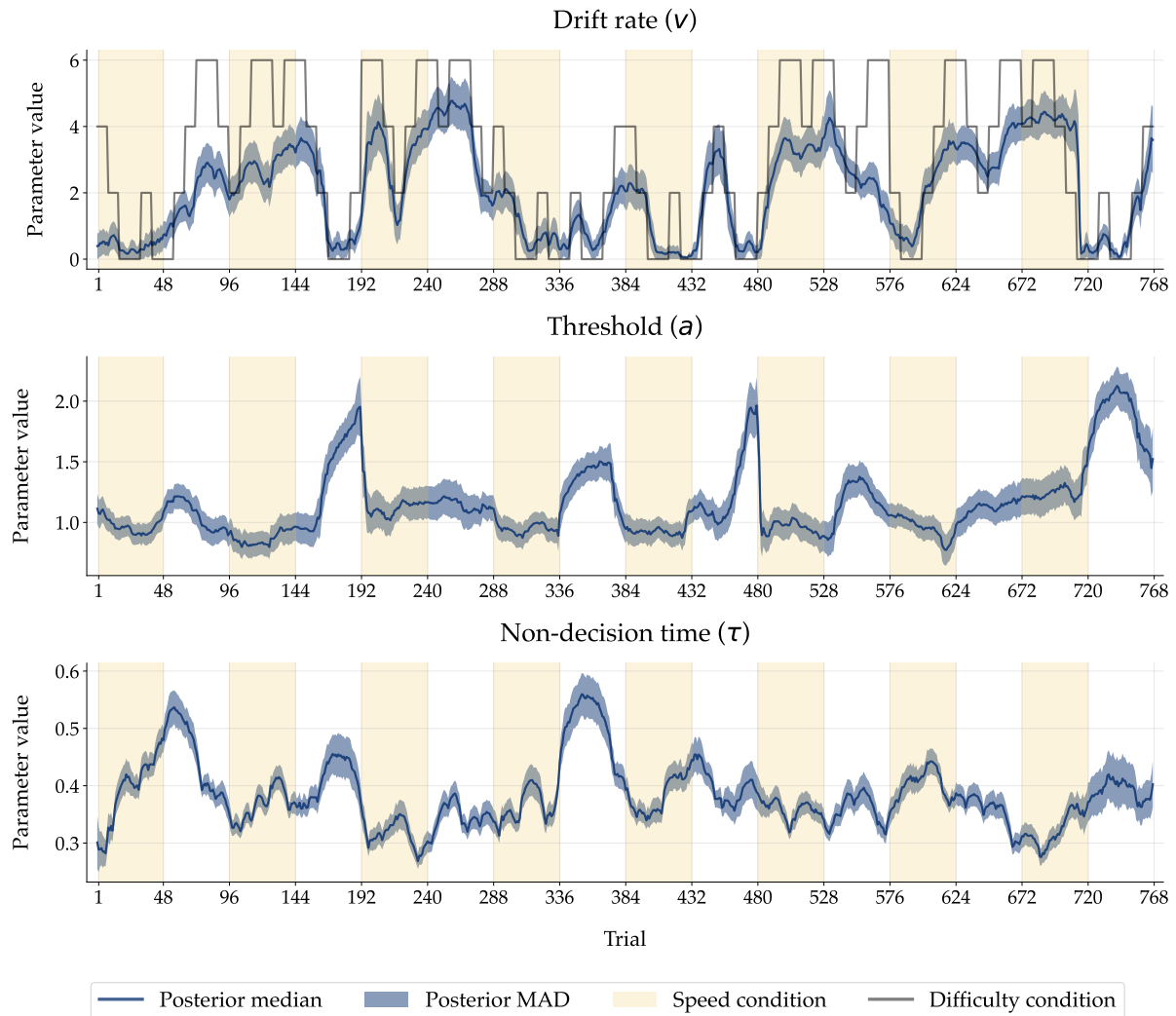


Figure 35: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 1 (a Lévy flight DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

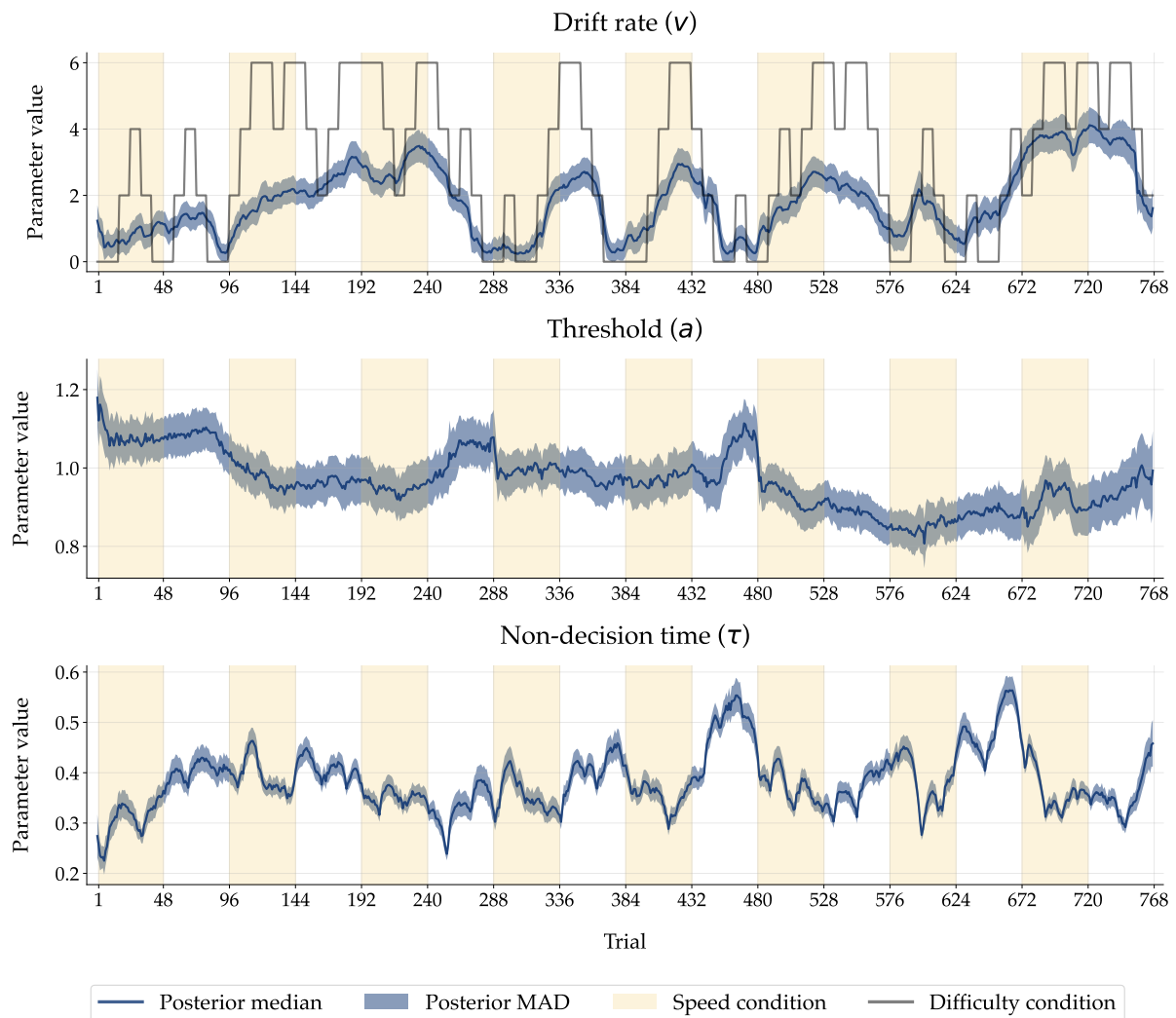


Figure 36: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 2 (a Lévy flight DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

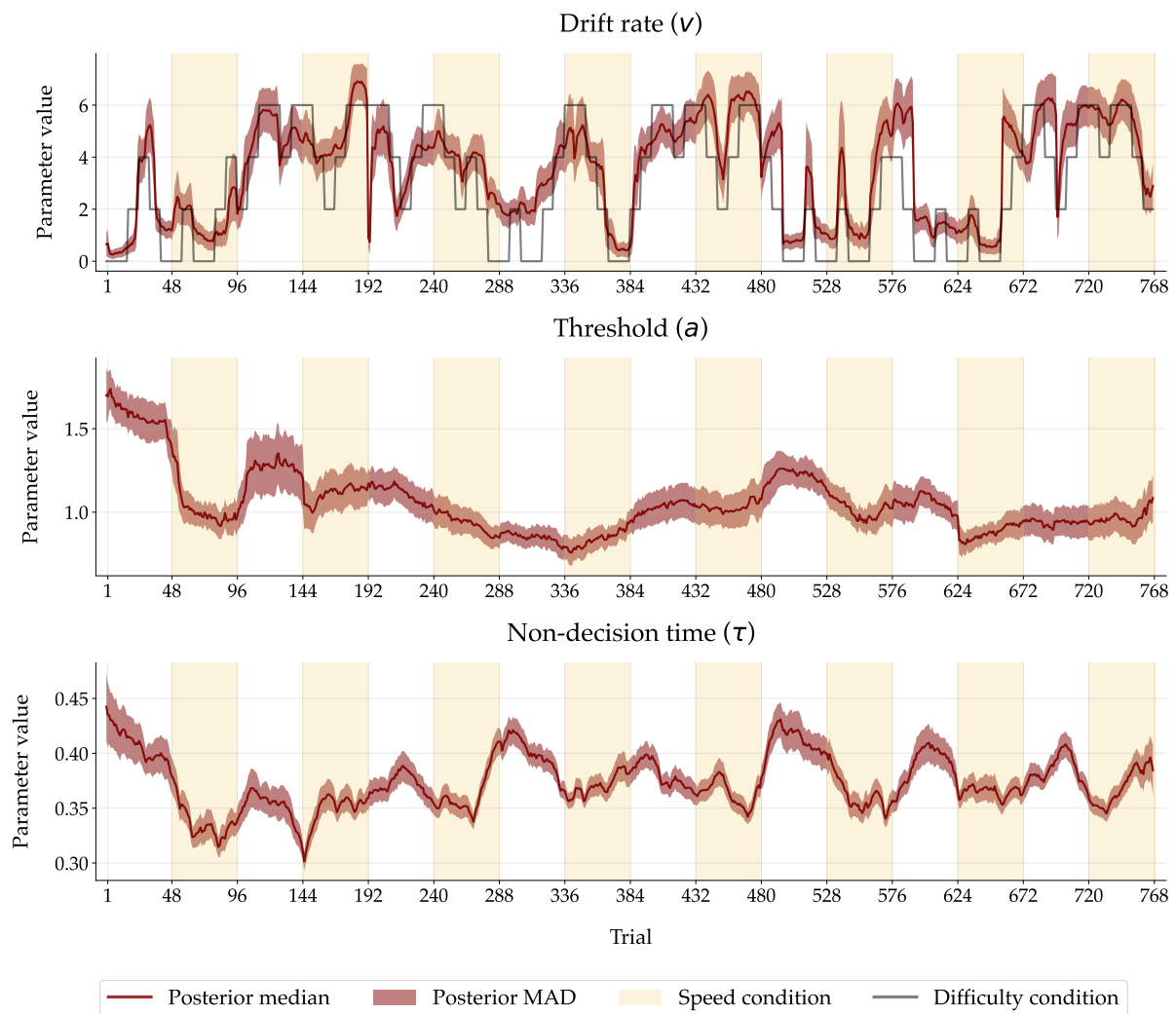


Figure 37: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 3 (a mixture random walk DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

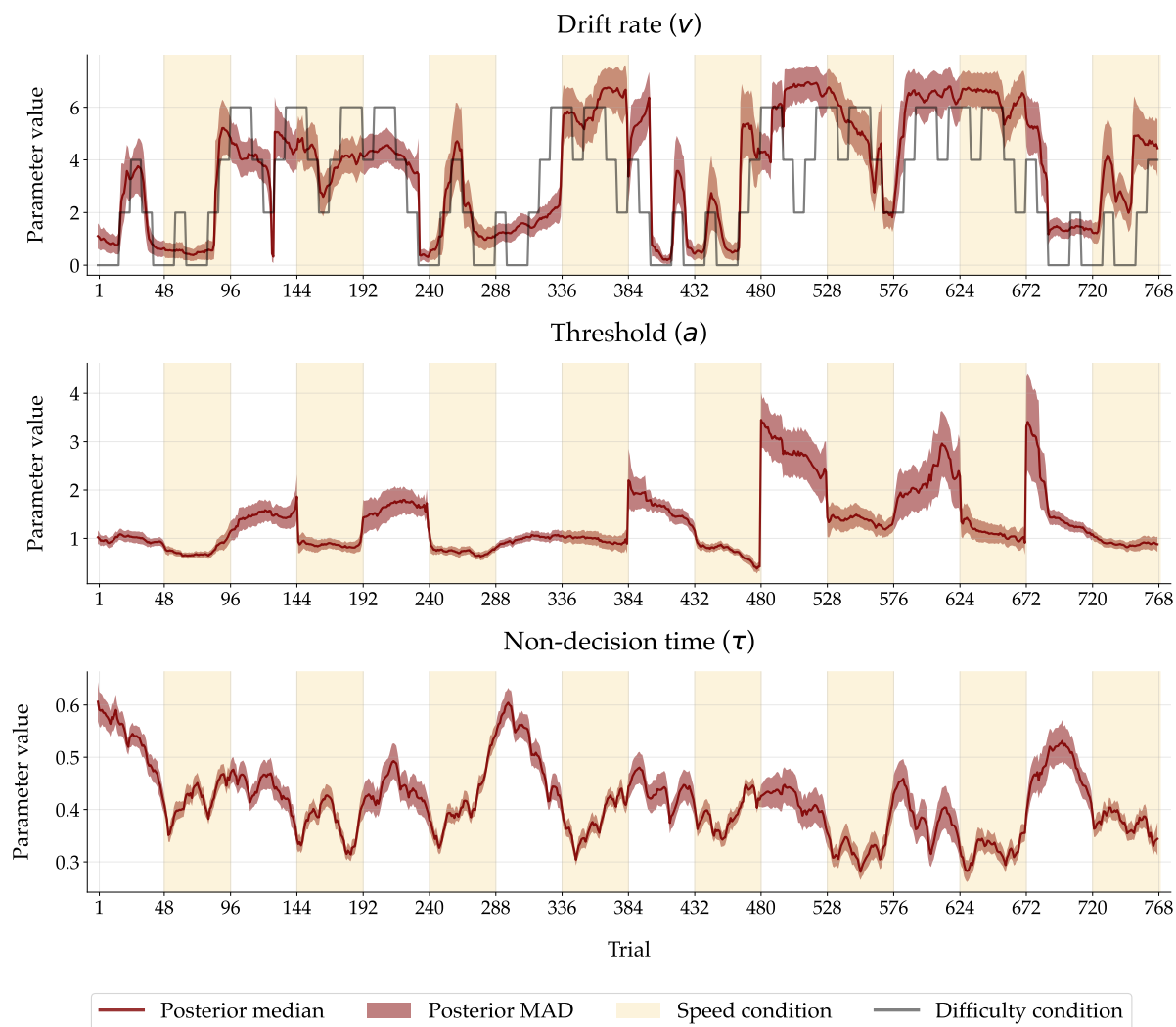


Figure 38: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 4 (a mixture random walk DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

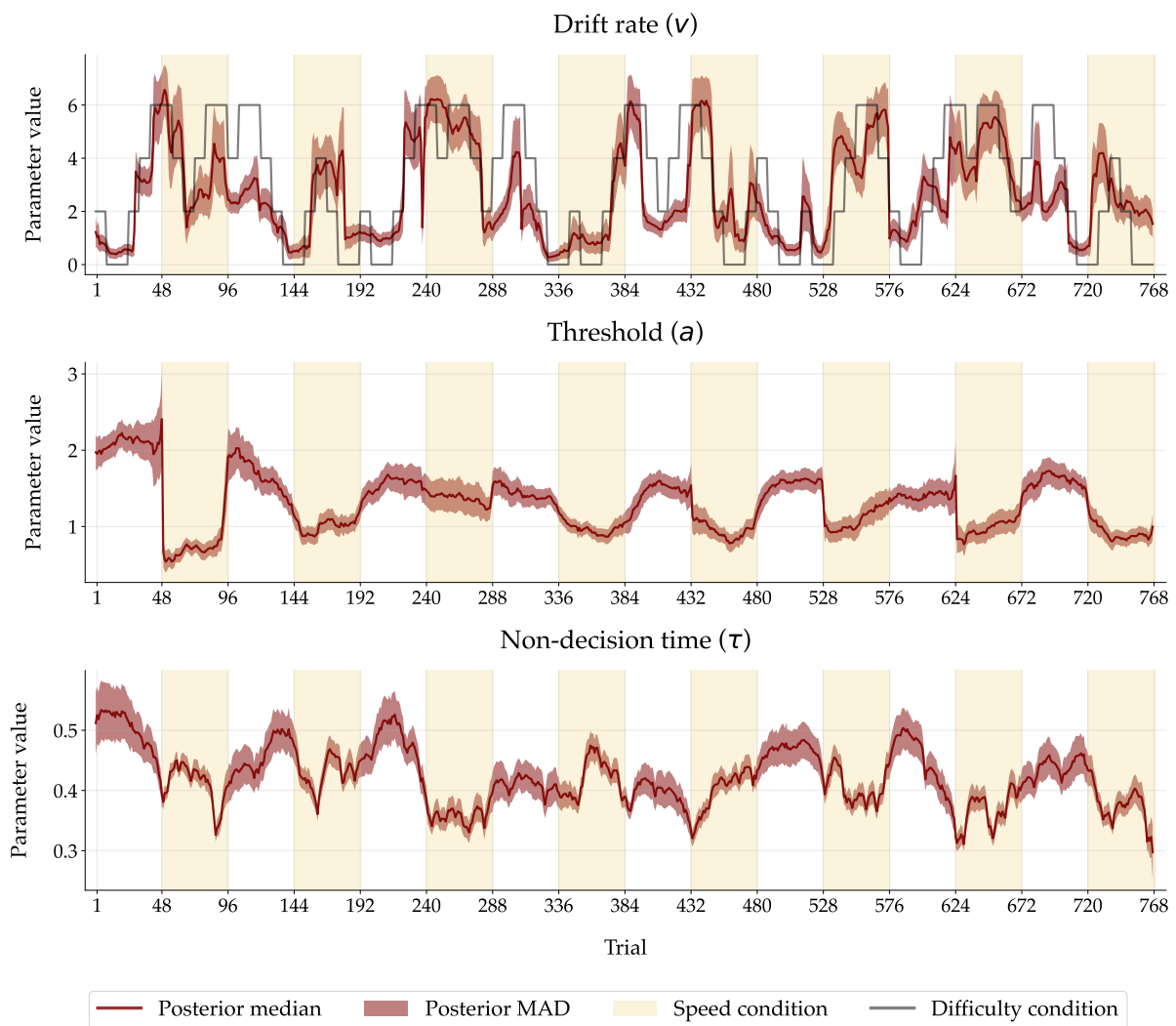


Figure 39: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 5 (a mixture random walk DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

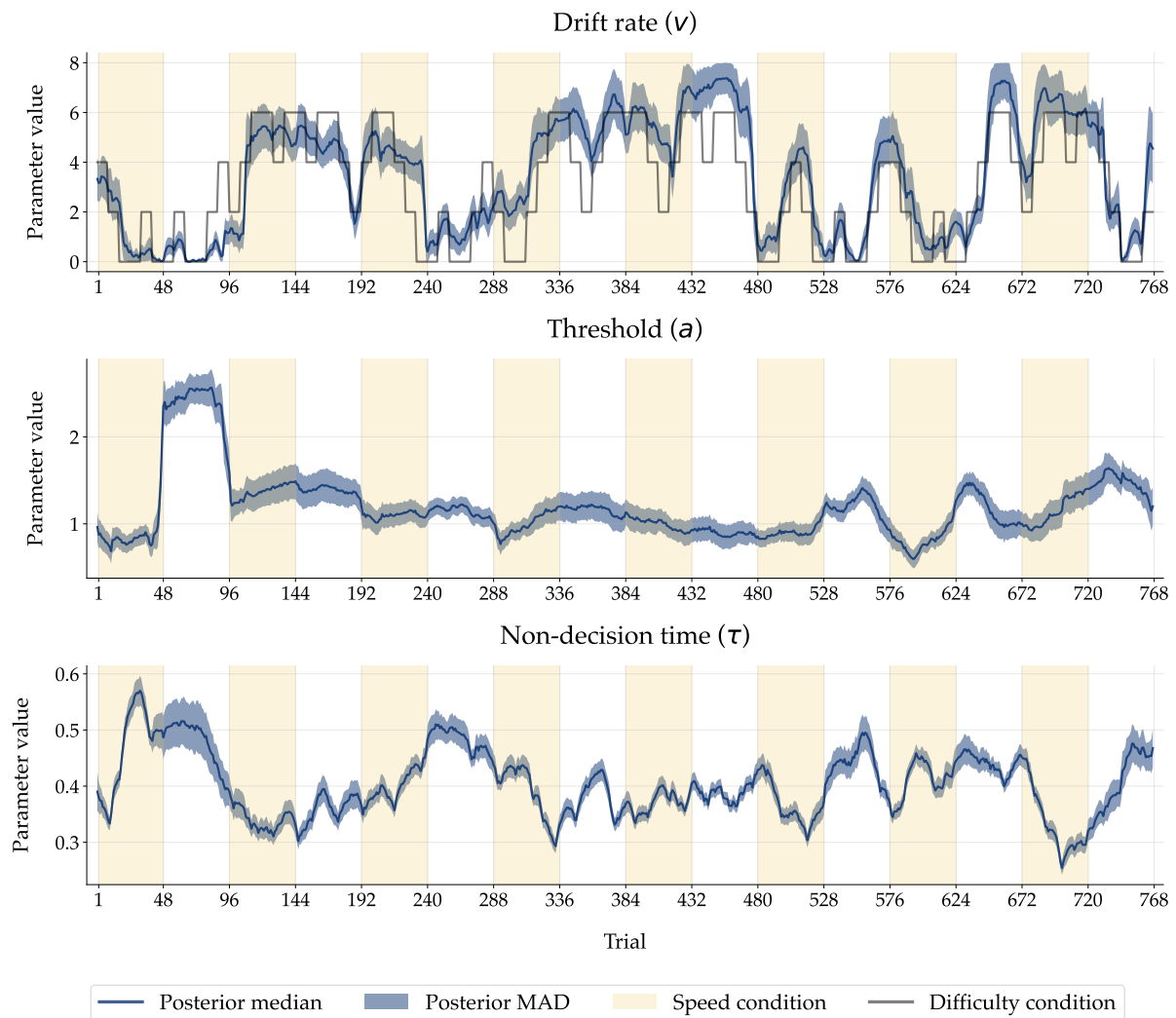


Figure 40: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 7 (a Lévy flight DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

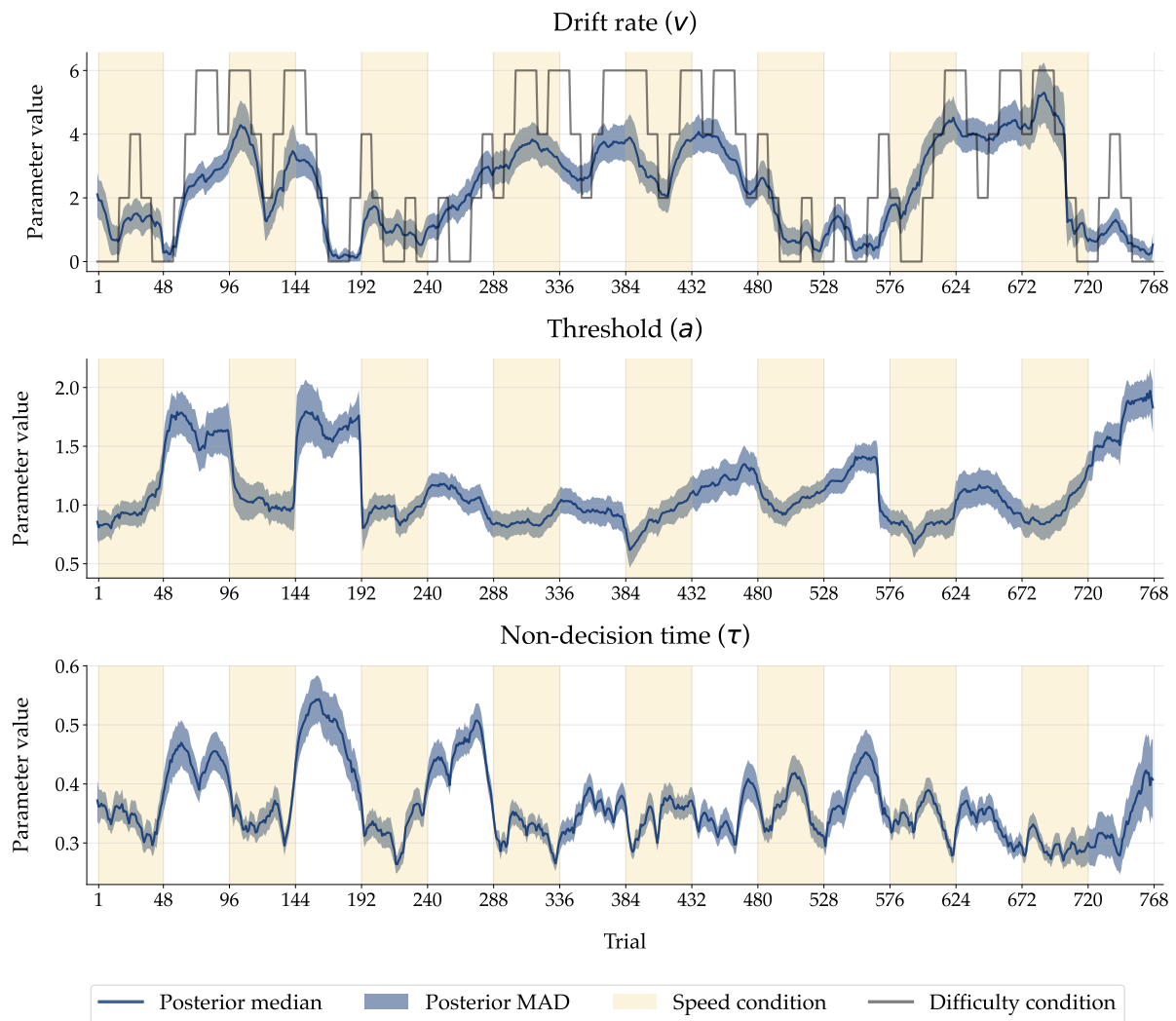


Figure 41: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 8 (a Lévy flight DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

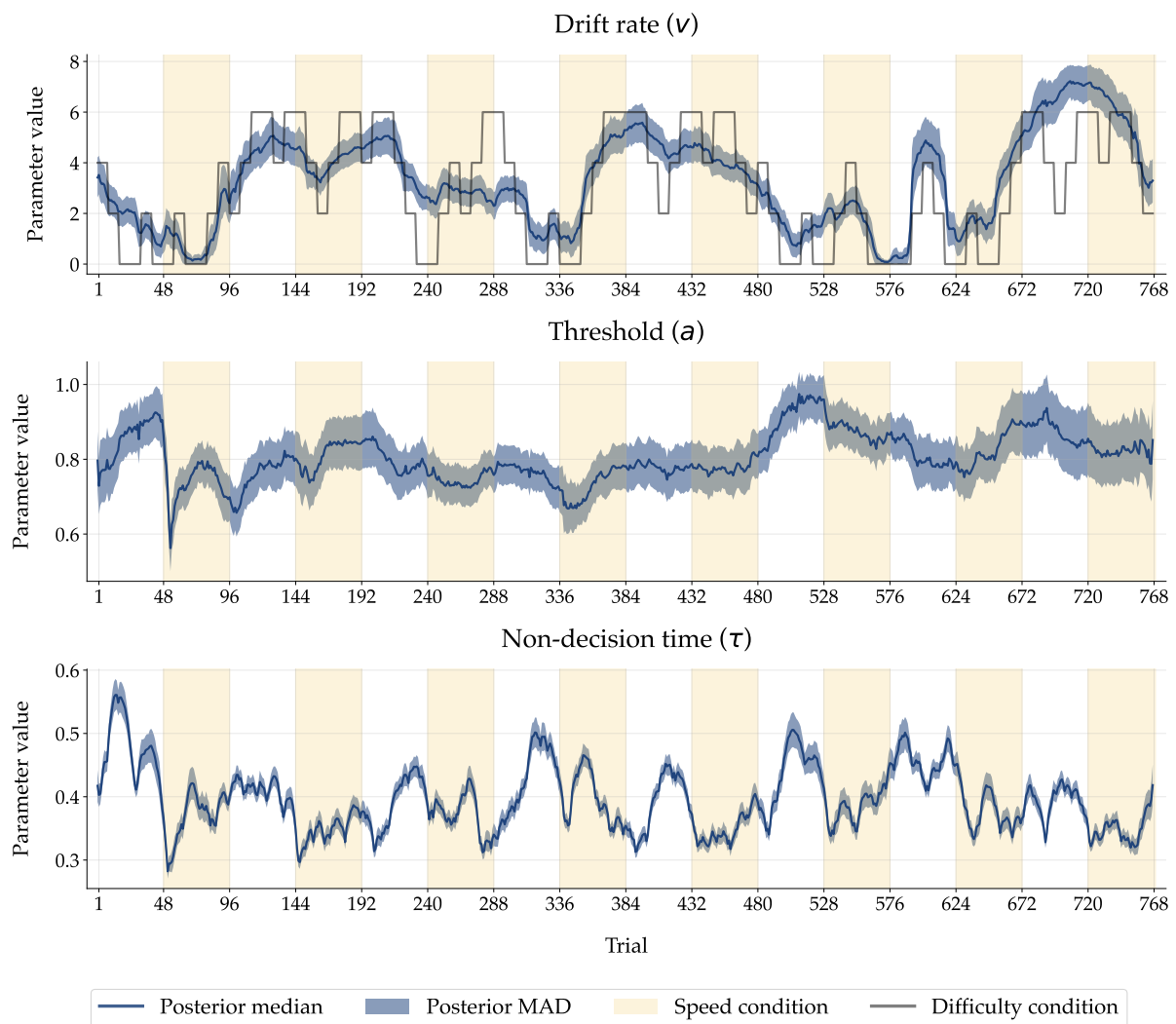


Figure 42: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 9 (a Lévy flight DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

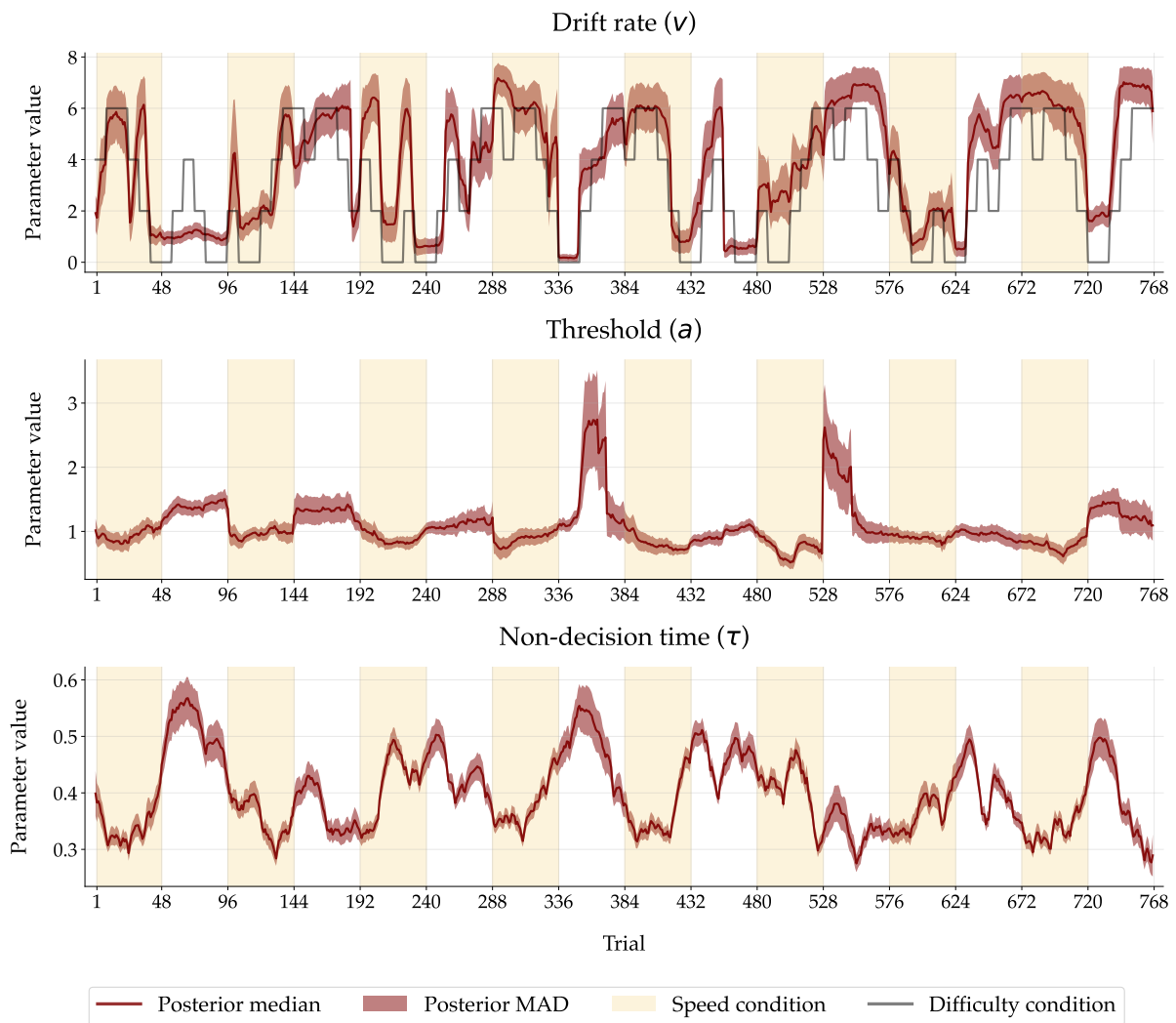


Figure 43: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 10 (a mixture random walk DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

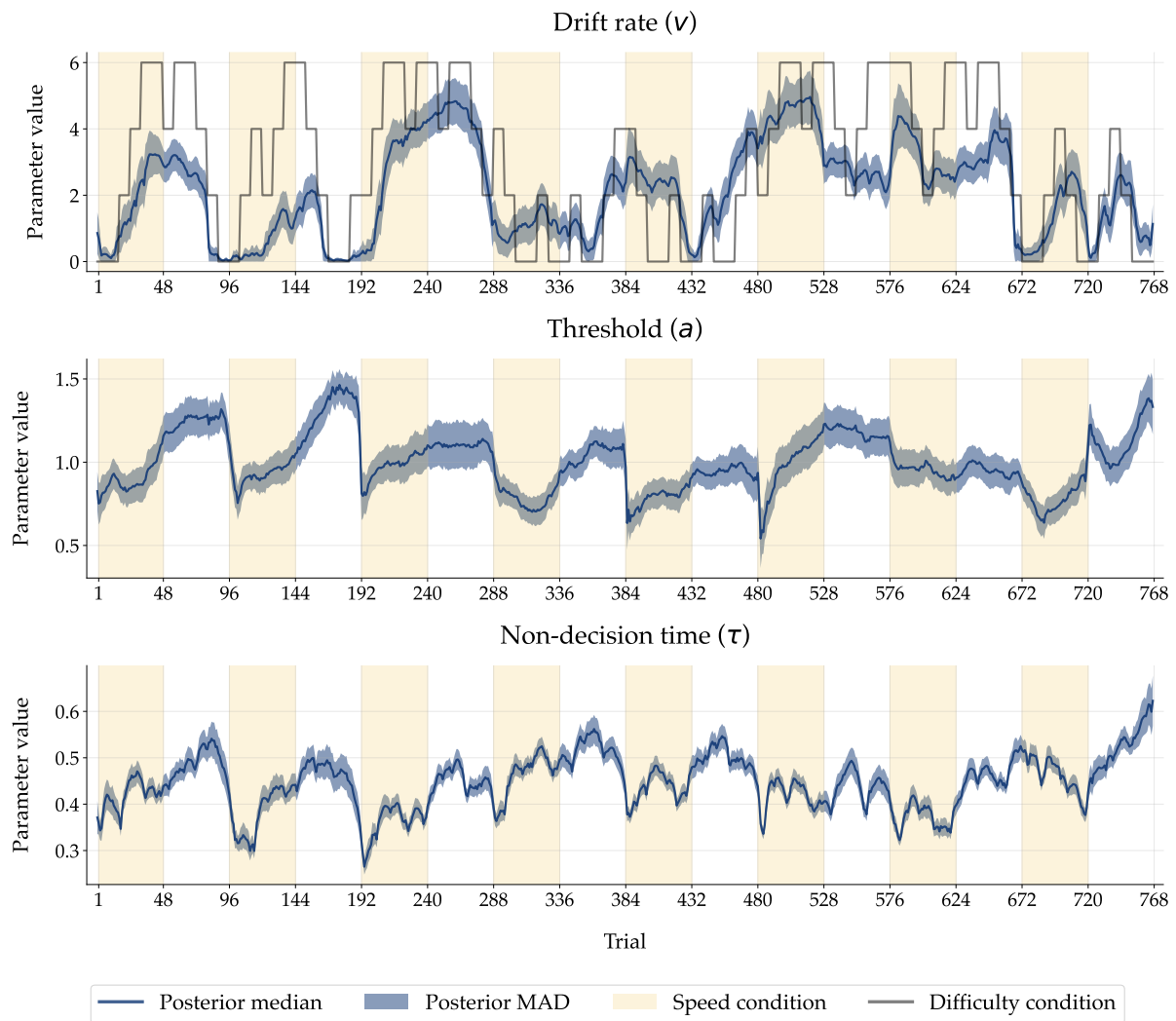


Figure 44: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 12 (a Lévy flight DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

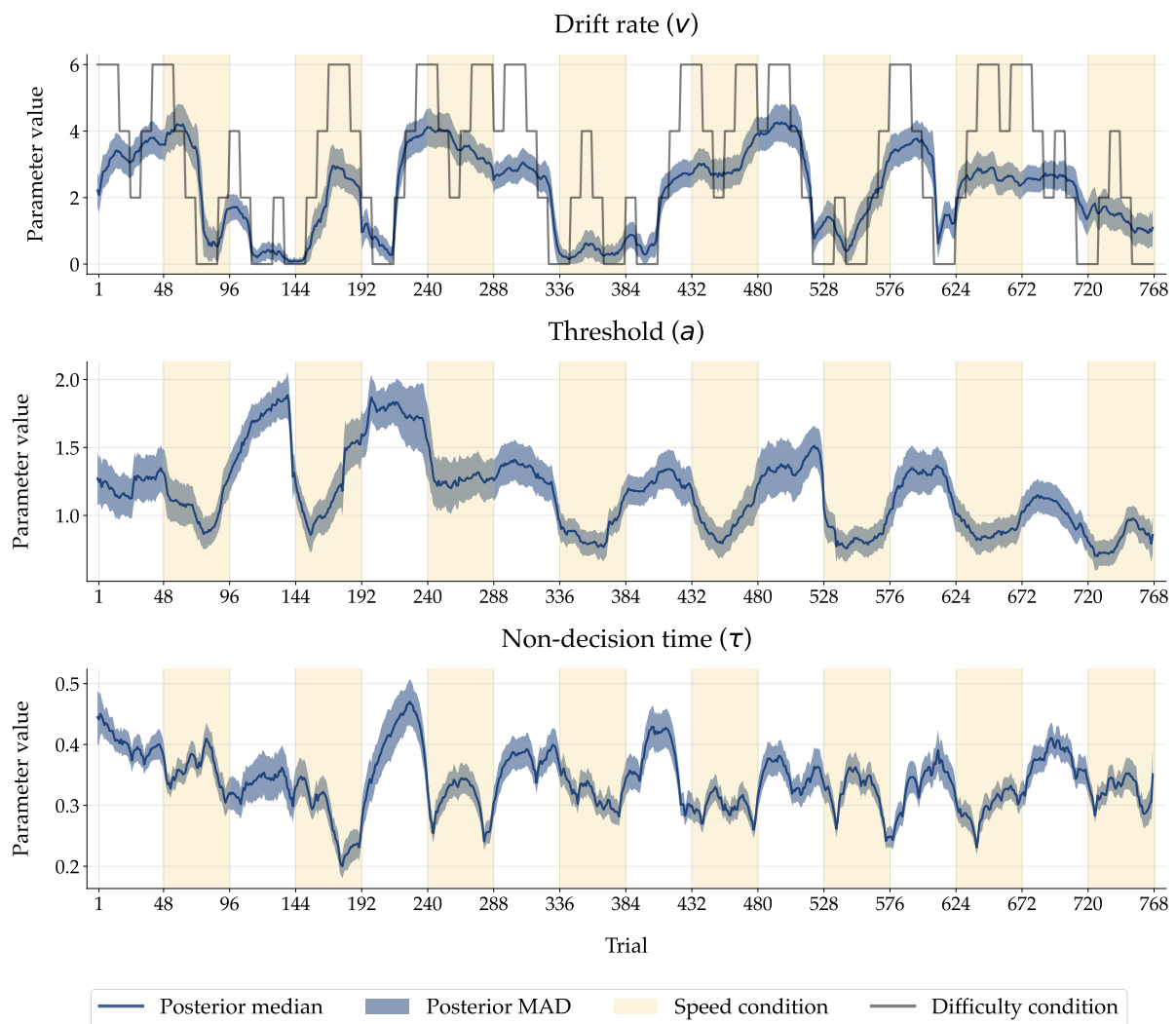


Figure 45: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 13 (a Lévy flight DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.

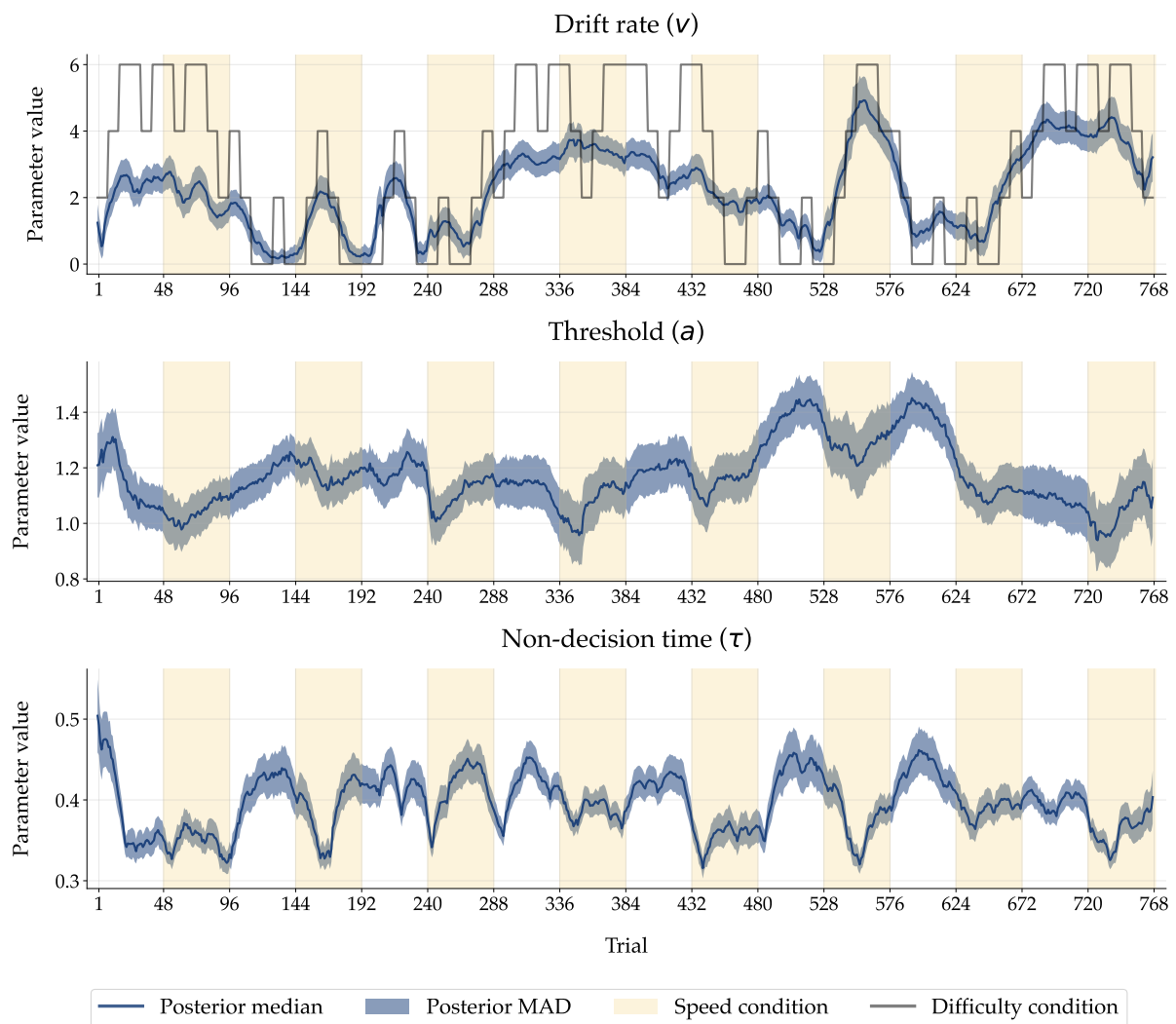


Figure 46: Posterior parameter trajectory inferred with the best fitting NSDDM of participant 14 (a Lévy flight DDM in this case) for all three DDM parameters (drift rate, threshold, and non-decision time) separately. The yellow shaded areas indicate trials where speed was emphasised over accuracy and blank white area indicated where the opposite was asked for. In the top panel, the task difficulty levels sequence is depicted in black lines.