RESEARCH

Validation and Comparison of Non-stationary Cognitive Models: A Diffusion Model Application

Lukas Schumacher[1] · Martin Schnuerch[2] · Andreas Voss[1] · Stefan T. Radev[3,4]

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.

Keywords Decision-making · Dynamics in cognition · Cognitive process models · Superstatistics · Amortized Bayesian inference

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 (Van Orden et al., 2003; Wagenmakers et al., 2004; Sebastian Castro-Alvarez & Tendeiro, 2023; Craigmile et al., 2010). 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 (Schurr et al., 2024).

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 (Schumacher et al., 2023; Beer, 2023; Evans and Brown, 2017; Li et al., 2023; Kucharský et al., 2021; Gunawan et al., 2022; Cochrane 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.

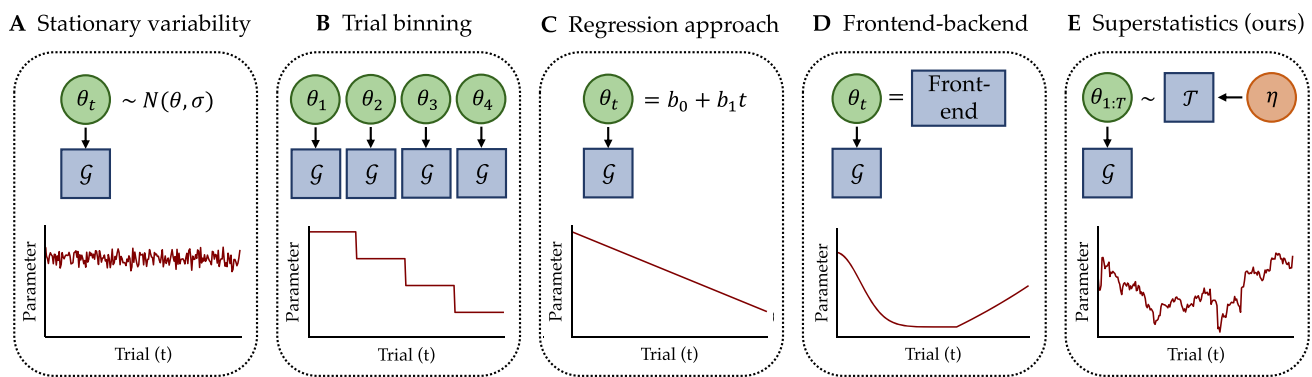Extended author information available on the last page of the article

A Stationary variability   **B** Trial binning   **C** Regression approach   **D** Frontend-backend   **E** Superstatistics (ours)

Fig. 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

The first approach assumes random fluctuations around a stable mean, referred to as stationary variability (see Fig. 1A). A prominent example of this approach is the "full" diffusion decision model (DDM), which allows for inter-trial variability of its core parameters (Ratcliff & Rouder, 1998; Ratcliff and Tuerlinckx, 2002). However, stationary inter-trial variability mainly improves in-sample model fit and cannot identify systematic changes or sudden shifts in 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 and 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 Fig. 1B). One can then examine variations in parameter estimates across these bins. The challenge in employing this approach is selecting 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 Fig. 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 Fig. 1D). Here, the backend model pertains to the cognitive model which formalizes how the behavioral data are 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 (Fontanesi et al., 2019; Osth et al., 2018; Schumacher & Voss, 2023; Brown et al., 2008). 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 (McDougle & Collins, 2021; Miletić et al., 2021; Fontanesi et al., 2019). 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 can be viewed from the lens of state space models and involves a superposition of multiple stochastic processes operating on distinct time scales (see Fig. 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 approach (cf. Fig. 1D), the transition model characterizes how the parameters change over time. However, the transition model in superstatistics is inherently a stochastic process, exemplified, for instance, by a Gaussian random walk, a regime-switching process, or a mixture between smooth and abrupt transitions.

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 artifact? To address this inquiry, we perform an experimental validation study in which 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 diffusion decision model (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 tasks, 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 (Voss et al., 2004; Lerche & Voss, 2019; Arnold et al., 2015). 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 task difficulty 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 parameters 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 in 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 from the behavioral data.

When dealing with various types of fluctuations, a crucial question arises: What kind of transition model is most suitable for capturing the expected dynamics? To address this, we implemented different NSDDMs that vary only in their transition model for the drift rate and threshold parameter. Specifically, we compare four distinct transition models: a Gaussian random walk, a mixture of a Gaussian random walk and uniformly distributed regime changes, a Lévy flight, and a regime-switching function, where parameters either remain constant from the previous time step or shift uniformly. These four transition models represent different prior assumptions about plausible parameter trajectories. They vary in complexity (i.e., the number of high-level parameters and functional expressiveness) 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, Talts et al., 2020; Säilynoja et al., 2022) and other validation methods (Schad et al., 2021; Gelman et al., 2020) 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 48th trial. 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:

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

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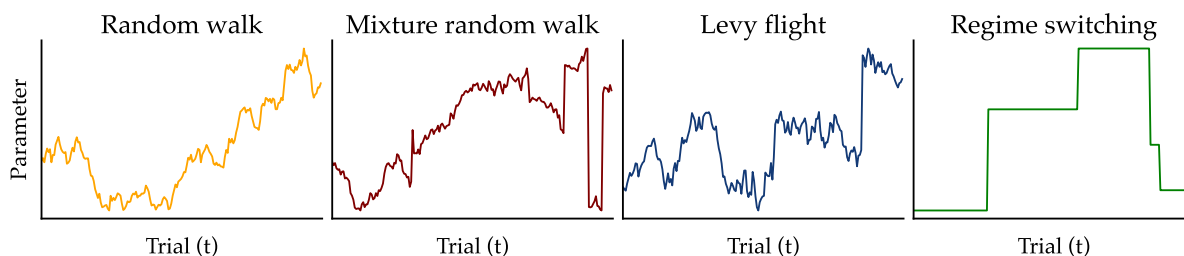
Fig. 2 An example illustration of the four high-level (transition) models considered in our study, governing the temporal variation of a hypothetical low-level model parameter

ing. 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

$$\mathbb{T}(\theta_t \mid \eta, \theta_{0:t-1}) = \int p(\theta_t, \xi \mid \eta, \theta_{0:t-1}) \, d\xi \qquad (2)$$

$$p(x_t \mid \theta_t, x_{1:t-1}) = \int p(x_t, z \mid \theta_t, x_{1:t-1}) \, dz, \qquad (3)$$

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 et al., 2020) as a principled approach to perform fully Bayesian inference by using only samples from the generative system defined by Eq. 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:

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

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:

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

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 $\mathbb{T}_1,..., \mathbb{T}_4$, which govern the trial-by-trial changes in local DDM parameters $\theta_{1:T}$. These transition models vary in flexibility in allowing changes to the low-level parameters and their underlying complexity, including the number of high-level parameters involved (see Fig. 2 for an exemplar trajectory generated by each transition model). To ensure that the low-level parameters remain within plausible ranges, we impose lower and upper bounds on their trajectories.[1] 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 simplicity, our transition models do not assume dependencies between the trajectories of the local DDM parameters a priori. We note that DDM parameters are typically found to be correlated (Boehm et al., 2018), and thus, priors with less entropy (e.g., correlated Gaussian random walk) are also plausible.

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:

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

where k denotes the individual model parameters. According to this transition model, the current value of each parameter

---

[1] To facilitate gradient-based training we transformed the parameters to an unbounded space via scaling.

$\theta_{k,t}$ 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:

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

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:

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

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:

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

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-switching transition function. Conversely, when $\rho = 1$, 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 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 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 Supplementary Material.

It is noteworthy that these transition models not only differ in their parameter counts but also in the degree to which they can generate diverse parameter trajectories. Thus, our Bayesian model comparison approach relies on a more general notion of model complexity as embodied by the *prior predictive distribution* (i.e., marginal likelihood). The next section discusses Bayesian model comparison from the lens of amortized Bayesian inference as an efficient approximation method.

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 sim-
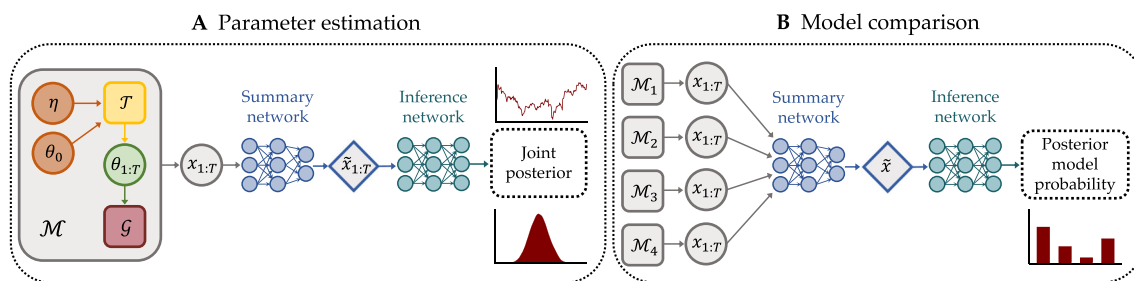
A Parameter estimation   **B** Model comparison

Fig. 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

ulations from the generative model and do not require an explicit likelihood or numerical integration. Thus, ABI recasts expensive Bayesian inference into a neural network prediction task, such that sampling from the target posterior and model refits happen almost instantaneously. In a previous study, we showed that ABI drastically outperforms traditional Bayesian methods for estimating time-varying parameters in terms of computation time (Schumacher et al., 2023).

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 (Radev et al., 2020; Gonçalves et al., 2020) to non-stationary Bayesian models. Accordingly, our goal is not only to infer the trajectories 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 $p(\theta_{1:T}, \eta \mid x_{1:T})$ from the observed time series $\{x_t\}_{t=1}^{T}$:

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

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

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

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 Fig. 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)

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

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 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 training the neural approximators.

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 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 Supplementary Material 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). These measures can be classified as *prior predictive*, since they depend on the *marginal likelihood* (see below) as a proxy for a model's generative diversity, penalizing prior complexity (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 (van Ravenzwaaij & Wagenmakers, 2022; Tendeiro & Kiers, 2019). 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 analyze both quantities in our study.

Following the common Bayesian terminology (MacKay, 2003), we can refer to the four competing models through an index set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$. Prior predictive Bayesian model comparison aims to find the simplest most plausible model within $\mathcal{M}$. To this end, we can compute PMPs for each of the competing models

$$p(\mathcal{M}_j \mid x_{1:T}) = \frac{p(x_{1:T} \mid \mathcal{M}_j)\, p(\mathcal{M}_j)}{\mathbb{E}_{p(\mathcal{M})}\left[p(x_{1:T} \mid \mathcal{M})\right]}, \tag{13}$$

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

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

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,

$$\mathrm{BF}_{ji} = \frac{p(x_{1:T} \mid \mathcal{M}_j)}{p(x_{1:T} \mid \mathcal{M}_i)}. \tag{15}$$

This ratio is referred to as the Bayes factor (BF). Consequently, a $\mathrm{BF}_{ji} > 1$ 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 Radev et al. (2020) and Elsemüller et al. (2023) which enables amortized Bayesian model comparison for arbitrary computational models (see Fig. 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, $q_\phi(\mathcal{M} \mid x_{1:T})$ 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,

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

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.

A key concern in amortized Bayesian model comparison is whether the network outputs truly reflect the underlying probabilities (Guo et al., 2017). Ideally, a posterior probability estimate of 0.9 for a given model suggests that a decision in favor of this model should be correct in 90% of cases. However, if it is only correct in 80% of cases, the estimate is *overconfident*. This discrepancy is quantified using the Expected Calibration Error (ECE; Naeini et al., 2015), which ranges from 0 (best) to 1 (worst). In practice, we estimate the ECE by averaging the deviations between predicted and true probabilities, calculated as relative frequencies within each probability bin.

---

[2] Note that Schumacher et al. (2023) focused exclusively on the filtering distribution in their benchmarking experiments.
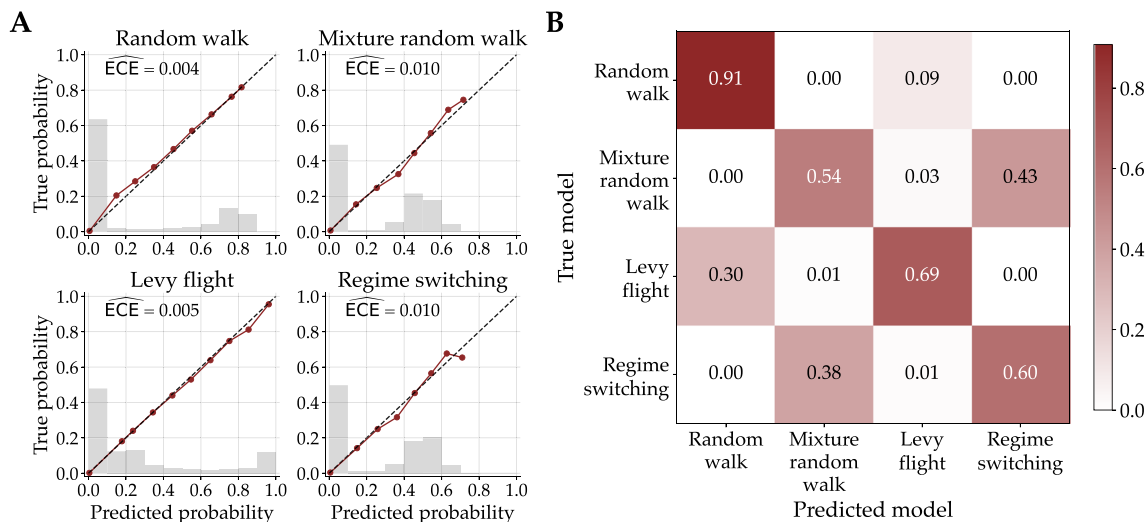
Fig. 4 In silico model comparison and sensitivity results. **A** Calibration curves of all four NSDDMs aggregated across the neural approximator ensemble. Additionally, the expected calibration error ($\widehat{ECE}$) is annotated within each subfigure. The gray histograms depict the relative frequencies 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 approximators within the ensemble

More recently, Elsemüller 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 Supplementary Material.

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 calibration (SBC; Talts et al., 2020; Säilynoja et al., 2022) based on 10,000 synthetic data sets each consisting of 800 trials 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 Fig. 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 our 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

Fig. 5 Empirical model comparison results. **A** Distribution of posterior model probabilities (PMP) from all ensembles and 14 individual participants. **B** Heatmap of average $\log_{10}$ Bayes factors (BF). Both metrics agree on favoring the Lévy flight over the other transition models
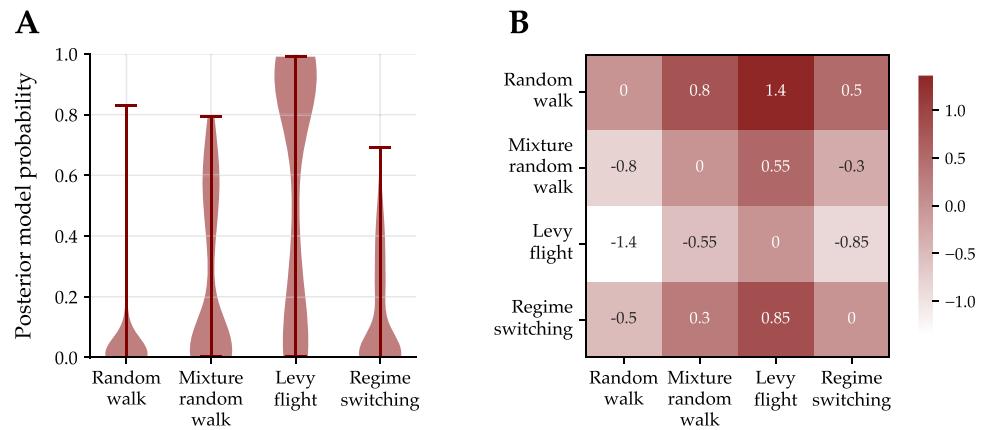
often yield data patterns that could have just as plausibly originated from the simpler random walk DDM.

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 a large overlap in synthetic data sets. Interestingly, the mixture random walk and the Lévy flight DDM are rarely 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.

To better understand the similarities between the transition models, we conducted a model misspecification analysis focusing on the mixture random walk and regime-switching models, as these two models exhibited the highest model mimicry. We cross-fitted the models to 100 synthetic data sets, each consisting of 800 trials. We then evaluated parameter recovery performance by computing the normalized root mean squared error (NRMSE) between true and estimated parameters for both the well-specified and misspecified scenarios. We found no notable difference in parameter recovery between the two scenarios (see Fig. 9 in the Supplementary Material).

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 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 displayed the distribution of the PMPs over all ensembles and individuals (Fig. 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 & Raftery (1995), an absolute value of $\log_{10}(\text{BF}) > 2$ indicates *decisive* evidence, absolute values between 1 and 2 signify *strong*, and between 0.5 and 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 Fig. 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
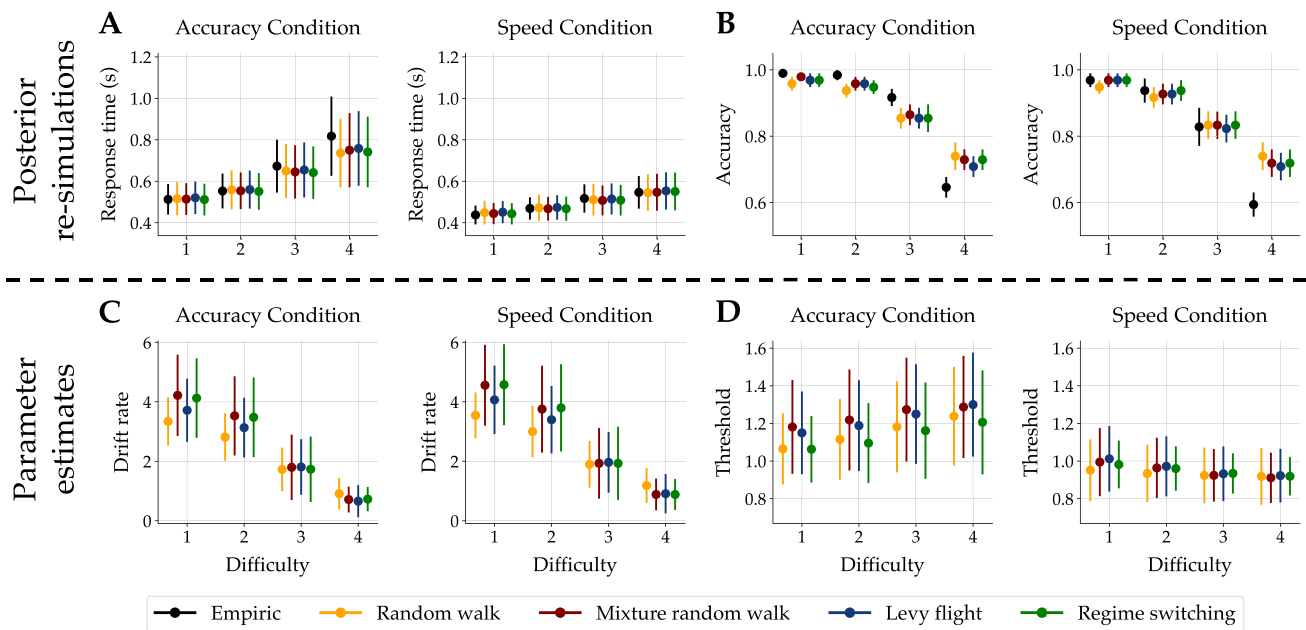
Fig. 6 Aggregated results from all models fitted to the empirical data. The top row illustrates posterior re-simulations as a measure of the model's generative performance and absolute goodness-of-fit to the data. The bottom row depicts parameter estimates of the drift rate and the threshold parameter 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 cor-rect 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 sep-arately. Points indicate medians and the error bars represent the median absolute deviations (MAD) across individuals and re-simulations

plausible than their simpler counterparts despite the implicit penalty for prior complexity imposed by Bayesian model comparison.

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 Fig. 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 Fig. 6B). Again, the empirical data align with 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 RT 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 behav-ioral 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 con-clude that all NSDDM implementations successfully capture the general patterns in the empirical RT data. Individual par-ticipant analyses, detailed in the Supplementary Material, affirm the same conclusions.
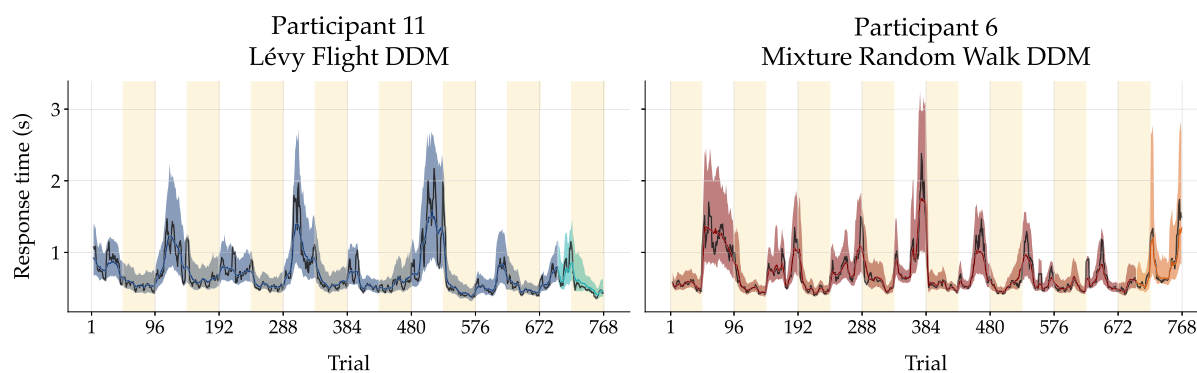
Participant 11
Lévy Flight DDM

Participant 6
Mixture Random Walk DDM

Fig. 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 emphasized over accuracy, while blank white areas denote instances where the opposite emphasis was applied

In addition to analyzing the absolute model fit at the aggregate level, we evaluated the fit across the RT time series. For each participant, we generated 250 posterior re-simulations for the first 700 trials using the corresponding best-fitting NSDDM. The remaining 68 data points were reserved for predictive analysis. For this analysis, we employed a *one-step-ahead* prediction approach, where we iteratively forecasted the subsequent data point starting at time step 700, followed by re-fitting the model for each of the remaining steps.

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 Supplementary Material. The colored lines depict the median, and the shaded bands represent 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 emphasized 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 Fig. 6 illustrates the posterior medians and MADs collapsed onto the different experimental contexts for the drift rate (Fig. 6C) and threshold parameter (Fig. 6D).

Analyzing the aggregated drift rate estimates reveals 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 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 Fig. 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
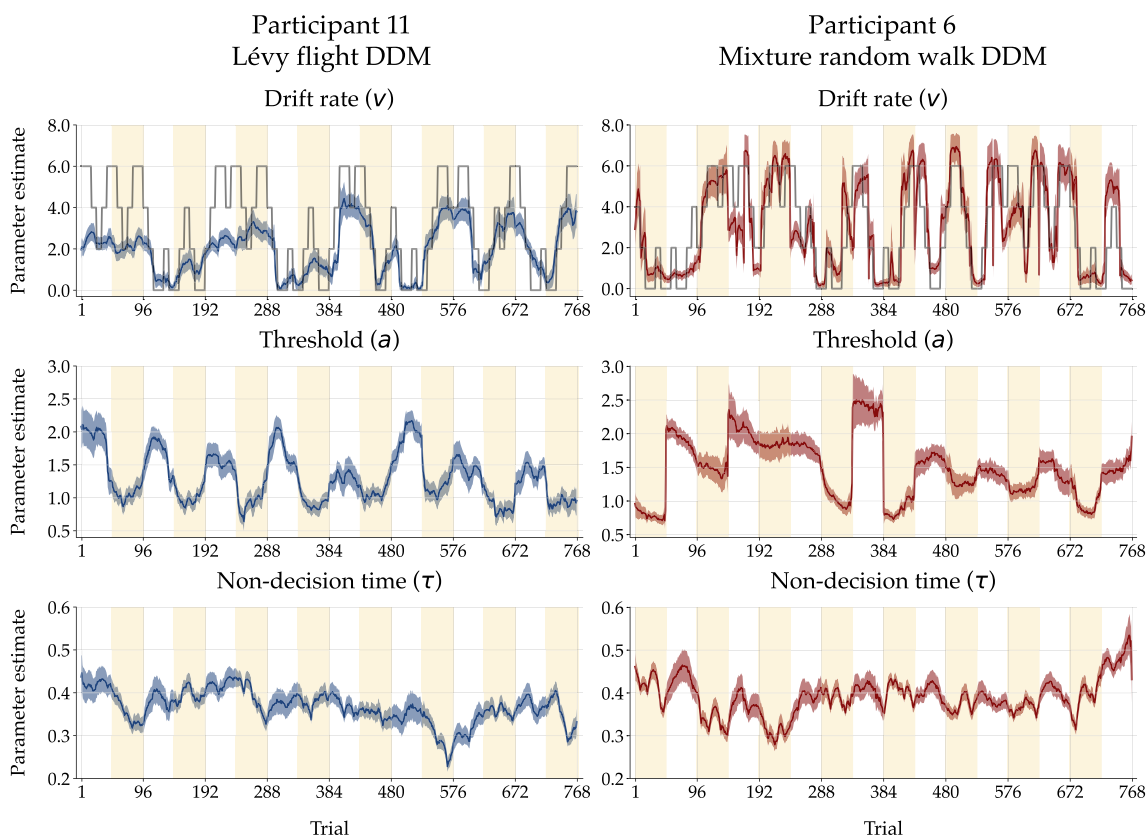
Fig. 8 Estimated parameter trajectories of two exemplar individuals corresponding to the respective best-fitting non-stationary diffusion decision model (NSDDM). In this instance, the left column showcases results from a Lévy flight DDM, while the right column displays parameter trajectories from a mixture 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 emphasized 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

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 Fig. 8. The Supplementary Material contains the inferred parameter trajectories of the remaining 12 participants. Each 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 a 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 Fig. 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 Fig. 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 (Voss et al., 2004; Arnold et al., 2015). 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 Supplementary Material, 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 can 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 the following core question: Can non-stationary DDMs (NSDDM) effectively detect experimentally induced changes in cognitive constructs from behavioral data alone? If so, our findings can provide the first substantial evidence for the validity of the superstatis-

tics 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 a good fit to the general data pattern, both on an aggregate level and on the level of the raw time series. This performance stands as compelling evidence supporting the validity of NSDDMs.

Nevertheless, we observed some misfits in accuracy in the most difficult condition regardless of the emphasis on speed or accuracy. Identifying the exact reasons for these discrepancies is challenging, but several factors may have contributed to the difficulty in achieving an accurate fit. First and foremost, our models did not receive information about the condition of the current trial. We aimed to validate whether the inferred parameter trajectories could capture changes in conditions, but this made it significantly harder to fit specific data patterns accurately. Additionally, some prior assumptions might have made it unlikely to fit the behavior in the most difficult condition accurately. Furthermore, we did not perform any pretreatment of the data, such as excluding trials with very short or long response times. The goal of this study was to experimentally validate the parameter trajectories, not to achieve perfect data fitting. Therefore, we did not conduct a detailed analysis to improve our modeling decisions to address these misfits.

Despite these challenges, the inferred parameter trajectories provided valuable insights into how the model responded to varying task conditions. 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 (Healey & Kahana, 2014, 2016; Forstmann et al., 2008; Wagenmakers et al., 2008; Wynton & Anglim, 2017). In fact, practice effects have been studied with various dynamic cognitive modeling approaches (Kahana et al., 2018; Gunawan et al., 2022; Evans & Hawkins, 2019; Evans 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 capture parameter changes 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 analyzing practice effects. However, we suggest that our flexible framework could be a promising alternative for investigating practice effects. Unlike pure regime-switching models, it can 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 of further informing 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 these 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 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.

Comparison to Time-Variant Models

Throughout this article, we have focused exclusively on the DDM as a specific example within the broader class of evidence accumulation models. The DDM assumes that certain parameters, such as the threshold parameter, remain constant during a single trial, but we allowed them to vary across different trials.

The literature has introduced models like the collapsing bounds DDM (Bowman et al., 2012; Shadlen and Kiani, 2013) and the urgency-gating model (Ditterich, 2006), which allow for time-variance within a trial. The collapsing bounds DDM posits that the threshold decreases throughout the decision-making process, while the urgency-gating model incorporates both the leaky integration of evidence samples and an "urgency signal" that prevents excessive delays in decision-making.

These models have been developed primarily to explain specific empirical data patterns, such as slow errors, without relying on the assumption of random trial-to-trial variability in the core parameters of the DDM (but see Hawkins et al., 2015). This raises the question of whether these models could offer a simpler explanation than our NSDDMs.

While the collapsing bounds DDM and the urgency-gating model can account for certain data patterns without assuming trial-to-trial variability, they cannot identify or explain systematic changes in the underlying constructs caused by factors such as learning, fatigue, motivation, or sudden insights. Uncovering such changes is precisely the goal of NSDDMs.

Nonetheless, the collapsing bounds DDM and the urgency-gating model could be valuable to explore within a superstatistical framework. It has been suggested that these models may provide more plausible explanations for tasks in which stimuli or conditions are dynamic within a single trial (Palestro et al., 2018; Evans et al., 2017). With the superstatistical framework, it would be straightforward to implement the collapsing bounds DDM and the urgency-gating model, resulting in models that allow for within-trial dynamics but

also aim to uncover systematic changes over the course of an experiment.

Limitations

Psychological research is usually interested in 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 were 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, our superstatistics framework offers numerous opportunities for future research. It could become a powerful tool in the methodological toolkit of researchers interested in temporal changes in cognitive constructs. As a general framework, it provides significant flexibility to uncover potential parameter dynamics in a data-driven manner. While we have focused on a specific model describing the evidence accumulation process during speeded binary decision-making, many other cognitive process models stand to benefit from such an approach. For instance, parameters in reinforcement learning models, such as the learning rate or the softmax temperature parameter, are likely to change over time (Li et al., 2023; Ger et al., 2024).

Furthermore, even when the temporal evolution of cognitive parameters is not the central research question, employing non-stationary models can offer advantages over stationary models (Schumacher et al., 2023). Our analysis of estimated trajectories demonstrates clear changes in parameters, highlighting how assuming stationarity could result in misleading conclusions. Exploring these dynamics more

deeply could provide valuable insights and drive advancements in various areas of cognitive science. For instance, researchers could attempt to link non-stationary parameter trajectories to additional neurophysiological time series data, such as EEG or eye movements. This approach could be beneficial in two ways: incorporating additional data as a constraint might improve the precision of cognitive parameter trajectory estimates, and it could enhance our understanding of the relationship between brain activity and specific cognitive constructs.

With great flexibility comes a 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 et al. (2023) advocate 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.

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).

Declarations

Ethics Approval This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of the Heidelberg University.

Consent to Participate Informed consent was obtained from all individuals who participated in the study.

Conflict of Interest The authors declare no competing interests.

References

Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E.W., Klessen, R.S.. Köthe, U. (2018). Analyzing inverse problems with invertible neural networks. arXiv preprint arXiv:1808.04730 https://doi.org/10.48550/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., 79*(5), 882–898. https://doi.org/10.1007/s00426-014-0608-y

Beck, C. (2004). Superstatistics: Theory and applications. *Continuum Mechanics and Thermodynamics, 16*, 293–304.

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

Beer, R.D. (2023). On the proper treatment of dynamics in cognitive science. *Topics in cognitive science*

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

Bowman, N. E., Kording, K. P., & Gottfried, J. A. (2012). Temporal integration of olfactory perceptual evidence in human orbitofrontal cortex. *Neuron, 75*(5), 916–927. https://doi.org/10.1016/j.neuron.2012.06.035

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

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. https://doi.org/10.1037/0033-295X.115.2.396

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

Craigmile, P. F., Peruggia, M., & Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika, 75*(4), 613–632. https://doi.org/10.1007/s11336-010-9172-6

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

Ditterich, J. (2006). Evidence for time-variant decision making. *European Journal of Neuroscience, 24*(12), 3628–3641. https://doi.org/10.1111/j.1460-9568.2006.05221.x

Elsemüller, L., Schnuerch, M., Bürkner, P.C., Radev, S.T. (2023). A deep learning method for comparing Bayesian hierarchical models.

Elsemüller, L., Olischläger, H., Schmitt, M., Bürkner, P.C., Köthe, U., Radev, S.T. (2023). *Sensitivity-aware amortized Bayesian inference* https://doi.org/10.48550/arXiv.2310.11122

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. https://doi.org/10.3758/s13423-016-1135-1

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

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

Evans, N.J., Hawkins, G.E., Boehm, U., Wagenmakers, E.J., & Brown, S. D. (2017). The computations that support simple decision-making: A comparison between the diffusion and urgency-gating models. *Scientific Reports, 7*(1), 16433. https://doi.org/10.1038/s41598-017-16694-7

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

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. https://doi.org/10.1073/pnas.0805903105

Galadí, J. A., Silva Pereira, S., Sanz Perl, Y., Kringelbach, M. L., Gayte, I., Laufs, H., & Deco, G. (2021). Capturing the non-stationarity of whole-brain dynamics underlying human brain states. *NeuroImage, 244*, 118551. https://doi.org/10.1016/j.neuroimage.2021.118551

Gelman, A., Vehtari, A., Simpson, D., Margossian, C.C., Carpenter, B., Yao, Y.. Modrák, M. (2020). Bayesian workflow. arXiv preprint arXiv:2011.01808

Ger, Y., Nachmani, E., Wolf, L., & Shahar, N. (2024). Harnessing the flexibility of neural networks to predict dynamic theoretical parameters underlying human choice behavior. *PLoS Computational Biology, 20*(1), e1011678. https://doi.org/10.1371/journal.pcbi.1011678

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., 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* (pp. 2404–2414).

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., & 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.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q. (2017). On calibration of modern neural networks. *International conference on machine learning* (pp. 1321–1330).

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

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. https://doi.org/10.1037/a0033715

Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review, 123*(1), 23–69. https://doi.org/10.1037/rev0000015

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P. C., Derks, K., Dienes, Z., & Hoijtink, H. (2023). A review of applications of the Bayes factor in psychological research. *Psychological Methods, 28*(3), 558–579. https://doi.org/10.1037/met0000454

Kahana, M.J., Aggarwal, E.V., Phan, T.D. (2018). The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* textit44(12), 1857–1863 https://doi.org/10.1037/xlm0000553

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

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. https://doi.org/10.1007/s42113-021-00115-0

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. https://doi.org/10.3758/s13428-018-1034-7

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

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*. https://doi.org/10.1101/2023.06.19.545524

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. https://doi.org/10.1038/s41467-018-04241-5

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

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), https://doi.org/10.1523/ENEURO.0355-16.2017

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*, e63055 https://doi.org/10.7554/eLife.63055

Naeini, M.P., Cooper, G., Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the aaai conference on artificial intelligence* (Vol. 29).

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., Weichart, E., Sederberg, P. B., & Turner, B. M. (2018). Some task demands induce collapsing bounds: Evidence from a behavioral analysis. *Psychonomic Bulletin & Review, 25*(4), 1225–1248. https://doi.org/10.3758/s13423-018-1479-9

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. https://doi.org/10.1109/TNNLS.2021.3124052

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., Elsemüller, L., Pratz, V., Schälte, Y.. Bürkner, P.C. (2023). BayesFlow: Amortized Bayesian workflows with neural networks. *Journal of Open Source Software, 8*(89), 5702 https://doi.org/10.21105/joss.05702

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., & 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., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Neural Computation, 9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

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

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. https://doi.org/10.1073/pnas.1100483108

Särkkä, S. (2013). *Bayesian filtering and smoothing*. Cambridge: 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. https://doi.org/10.1038/s41598-023-40278-3

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

Schurr, R., Reznik, D., Hillman, H., Bhui, R., & Gershman, S. J. (2024). Dynamic computational phenotyping of human cognition. *Nature Human Behaviour, 1–15*,. https://doi.org/10.1038/s41562-024-01814-x

Sebastian Castro-Alvarez, R. R. M., Bringmann, Laura F., & Tendeiro, J. N. (2023). A time-varying dynamic partial credit model to analyze polytomous and multivariate time series data. *Multivariate Behavioral Research,* 1–20. https://doi.org/10.1080/00273171.2023.2214787

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron, 80*(3), 791–806. https://doi.org/10.1016/j.neuron.2013.10.047

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. https://doi.org/10.1007/s11222-022-10090-6

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., Gelman, A. (2020). Validating Bayesian inference algorithms with simulation-based calibration. https://doi.org/10.48550/arXiv.1804.06788

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods, 24*(6), 774–795. https://doi.org/10.1037/met0000221

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. https://doi.org/10.1037/met0000415

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. https://doi.org/10.1037/0096-3445.132.3.331

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. https://doi.org/10.1037/rev0000258

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. https://doi.org/10.3758/BF03196893

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. https://doi.org/10.3758/BF03196615

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

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. https://doi.org/10.3758/s13423-017-1243-6

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 (arXiv:2202.07125)

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. https://doi.org/10.1037/xlm0000404

Authors and Affiliations

Lukas Schumacher[1] · **Martin Schnuerch**[2] · **Andreas Voss**[1] · **Stefan T. Radev**[3,4]

✉ Lukas Schumacher
schuma.luk@gmail.com

Martin Schnuerch
martin.schnuerch@uni-mannheim.de

Andreas Voss
andreas.voss@psychologie.uni-heidelberg.de

Stefan T. Radev
stefan.radev93@gmail.com

[1] Institute of Psychology, Heidelberg University, Heidelberg, Germany

[2] Institute of Psychology, University of Mannheim, Mannheim, Germany

[3] Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY, USA

[4] Center for Modeling, Simulation and Imaging for Medicine, Rensselaer Polytechnic Institute, Troy, NY, USA