Zoning Out? Inferring Latent Engagement States During Perceptual Decision-Making in Humans

Hoorde, Lorenzo Van

Psychologie
Faculteit der Sociale Wetenschappen

Zoning Out? Inferring Latent Engagement States During Perceptual Decision-Making in Humans

Lorenzo Van Hoorde

Research Master Thesis *Cognitive Neuroscience*
Date: 24/10/2025
Student number: 1984713
Supervisor: Philippa Johnson
Second reader: Michiel van Elk
Word count: 8991

Abstract

Introduction: Psychology experiments typically assume participants remain focused on the task, without explicitly accounting for periods of disengagement. However, attention naturally fluctuates, and standard experimental designs may not capture these dynamics. Here, we use a generalized linear model–hidden Markov model (GLM-HMM) to infer latent engagement states directly from human choice behaviour. The present study is part of a larger project on engagement dynamics and aimed to: 1) optimize the statistical pipeline for subsequent large-scale analyses, and 2) assess the robustness and reproducibility of the GLM-HMM framework by replicating findings from Ashwood et al. (2022).

Methods: We analysed choice behaviour from 27 participants performing a two-alternative motion discrimination task. A GLM-HMM was used to infer latent engagement states from trial-by-trial responses. Model parameters were optimized in a pilot dataset (n = 3) to evaluate convergence, prior settings, and model stability. The final analysis replicated Ashwood et al. (2022) using the refined pipeline. Model performance was assessed via cross-validation, and state-dependent effects on behavioural parameters, reaction times, and accuracy were evaluated using linear and generalized linear mixed-effects models

Results: Cross-validation indicated that a two-state GLM-HMM explained choice behaviour better than a single-state model (ΔLL = 0.012 bits/trial). Sensory evidence weights were higher in state 1 than state 2 ($t(25) = 2.63$, $p = .015$, $d = 0.52$), whereas use of the win–stay, lose–switch strategy was greater in state 2 ($t(25) = -4.73$, $p < .001$, $d = 0.93$). Reaction times were slightly longer in the less engaged state ($\beta = 0.015$, $p < .001$), and accuracy tended to be lower. The probability of engagement declined across trials ($\beta = -0.00060$, $p < .001$).

Discussion: The findings provide preliminary evidence that the GLM-HMM can identify latent fluctuations in task engagement from behavioural data. Two distinct states reflected varying reliance on task-relevant and task-irrelevant strategies. Our results closely replicate those of Ashwood et al. (2022), underscoring the robustness and reproducibility of the framework. By refining the statistical pipeline, this study establishes a foundation for large-scale analyses aimed at understanding how engagement fluctuates during perceptual decision-making.

Layman's abstract

Maintaining focus on tasks for a long time is notoriously difficult. Even when we try to concentrate, attention naturally fluctuates. Typically, psychology experiments implicitly assume that participants remain focused throughout a task, without accounting for periods of disengagement. In this study, we looked for periods of high and low engagement during a simple visual decision-making task. We asked 27 participants to judge the direction of moving dots on a screen. Instead of measuring attention directly, we analysed their choices using a statistical model called a generalized linear model–hidden Markov model (GLM-HMM). This model can detect hidden "states" in behaviour, such as when someone is more or less focused, by looking at patterns in their responses across trials.

First, we tested and refined the model using a small pilot dataset to make sure it worked reliably. Then, we applied it to the full group of participants. We checked whether the model's results were stable and whether it could explain behaviour better than a simpler model.

The analysis revealed two distinct states. In one state, participants relied more on the actual sensory information from the task, suggesting they were more engaged. In the other state, they used a simple "win–stay, lose–switch" strategy, indicating lower engagement. Reaction times were slower, and accuracy was marginally lower in the less engaged state. Across the session, the likelihood of being in the engaged state decreased gradually. These patterns closely matched results from a previous study, showing that our model produces reliable and reproducible results.

Overall, this work shows that hidden changes in engagement can be detected from normal choice behaviour. Even in tasks that do not explicitly induce inattention, people switch between more and less engaged states. Detecting these changes can help researchers understand how attention fluctuates and could improve the way experiments are designed or how data are interpreted. This study lays the groundwork for larger projects that will explore engagement in more complex tasks and across more people.

Introduction

How long are you able to stay focused on a task? Do you maintain your focus from start to end, or do you tend to periodically disengage from the task before completion? Chances are that your mind tends to wander regularly during effortful tasks, such as reading this research paper. Drawing from our experience, it seems evident that people frequently disengage from tasks before completion. Yet, fluctuating attention levels have often been ignored in psychological experiments. Current methodology to study (dis-)engagement during psychological experiments contain notable limitations or deliberately probe for disengagement, which raises the question of whether those findings generalise to more typical task settings.

The challenge of sustaining attention has long been recognized across research traditions, though fragmented terminology, ranging from absent-mindedness (e.g. Manly et al., 1999) and mind-wandering (e.g. Smallwood & Schooler, 2006), to attentional lapses (Unsworth et al., 2010), task-unrelated thought (Giambra, 1995) and task-engagement/disengagement (Smallwood et al., 2004), has obstructed conceptual progress (Cheyne et al., 2009). For clarity, we use the term engagement throughout to refer broadly to this phenomenon. Here, we briefly review key findings from research on sustained attention and mind-wandering to identify current methodological limitations and motivate the present work.

Mind-Wandering and Sustained Attention

Sustained attention refers to the ability to maintain engagement in a task over extended periods (Fortenbaugh et al., 2017). Failures to sustain attention seem prevalent based on our phenomenological experience, yet this ability is less studied than other aspects of attention like switching and selective attention (Esterman & Rothlein, 2019). Early theories attributed failures to sustain attention to resource depletion (overload) or understimulation (underload), while newer theories highlight other aspects such as cost-benefit trade-offs (see Fortenbaugh et al., 2017, for a review). Behaviourally, disengagement is consistently associated with increased error rates and more variable reaction times (RTs), while both slower and faster RTs are frequently reported (Unsworth, 2015; Isbell et al., 2018), likely reflecting different modes of disengagement. However, the mechanisms underlying these temporal fluctuations and how they dynamically unfold across trials remain poorly understood (Gaillard et al., 2021).

Relatedly, mind-wandering refers to the shift of attention away from the task at hand towards task-unrelated, self-generated thought (Smallwood & Schooler, 2006). According to the resource-control theory (Thomson et al., 2015), a prominent framework on sustained attention, this self-generated thought is the default state for individuals, such that there is a continuous bias for executive resources to be directed toward mind-wandering. Supporting this view, studies using

smartphone prompts in everyday life have estimated that people spend up to 50% of their waking time mind-wandering (Kane et al., 2007; Killingsworth & Gilbert, 2010).

Thought probes, which intermittently pause a task to ask participants whether their attention was on- or off-task, are widely used in mind-wandering research, yet they face several well-acknowledged limitations (Zhang & Kool, 2025). Critically, because probes are only presented at a few arbitrary time-points, they provide sparse sampling of internal states and consequently fail to capture the temporal dynamics of attentional fluctuations (Bastian & Sackur, 2013). Furthermore, thought probes can interrupt the natural flow of attentional states (Schubert et al., 2020) and depend on participants' honesty (Vinski & Watter, 2012) and meta-cognitive awareness (Schooler et al., 2011), introducing systematic biases that are difficult to quantify or control for. Taken together, these limitations highlight the need for methods that allow for continuous measurement of attentional fluctuations.

Several studies have sought to address this issue by leveraging intra-individual RT variability. RT variability is widely used to index fluctuations in attentional control and is thought to reflect how susceptible an individual is to frequently disengaging from task-relevant goals (Kane et al., 2016). RT stability is thus considered an indicator of an individual's ability to sustain focused performance over time (Unsworth et al., 2015). Broadly, two methodological approaches are commonly used to measure RT variability: distributional approaches and moving-window approaches, each carrying specific limitations.

Using an Ex-Gaussian model, which separates the Gaussian and exponential components of RTs to better handle their skewed distribution, and a Hidden Markov Model applied to fMRI data, Cai et al. (2020) identified latent brain states during a simple choice response task in children. They found that greater occupancy of a task-optimal state was associated with lower RT variability, faster evidence accumulation, and fewer inattention symptoms, whereas engagement of a second state showed the opposite pattern. However, like thought probes, distributional approaches cannot capture the temporal evolution of attentional fluctuations, as they aggregate responses across trials and thus only quantify overall behavioural stability.

An early investigation into attentional fluctuations by Esterman et al. (2012) introduced the *variance time course* (VTC), a measure designed to track trial-by-trial deviations from a participant's mean RT with high temporal precision. Using this approach in the gradual-onset Continuous Performance Task (gradCPT), the authors demonstrated that attention alternates between periods of low variability, high accuracy, and small error-related adjustments ("in the zone") and periods of higher variability, lower accuracy, and larger error-related adjustments ("out of the zone"). These dynamics occurred independently of general time-on-task effects, implying that the VTC successfully captured distinct attentional states in the experiment.

However, despite offering excellent temporal resolution, VTC and related moving-window approaches have limitations. They rely on a researcher-defined smoothing windows, which imposes a fixed timescale of fluctuations that may not reflect individual differences in state-switching dynamics (Rosenberg et al., 2013). In addition, the subsequent median-split classification of trials into “in-” and “out-of-the-zone” states enforces an equal distribution of attentional states across participants, disregarding natural variation in engagement frequency and duration (Unsworth et al., 2010). Yet, despite these limitations, the discussed findings emphasize the potential of continuous behavioural measures in characterising the latent structure of engagement dynamics, and suggest that subjects recruit distinct attentional states during psychological experiments.

Other studies that have aimed to characterise the temporal dynamics of fluctuating engagement during psychological experiments rely on M/EEG or fMRI. One such study, by Gaillard et al. (2021), investigated whether prolonged cognitive activity results in a monotonic decrease in the efficiency of recruited brain processes or whether the brain is able to sustain functions over timespans of one hour and more. Specifically, they trained humans and macaques on a vigilance task and simultaneously recorded multi-unit neuronal activity and local field potentials from multiple recording sites in the frontal eye field. The authors found consistent behavioural fluctuations between periods of high and low performance at a rhythm of 4 to 7 cycles per hour. They further showed that these oscillations coincide with 1) phase-locked rhythmic fluctuations in prefrontal encoding of visuospatial information 2) electrophysiological signatures of attention, and 3) rhythmic variations in pupil size in humans. Synthesizing these findings, the authors propose that attentional and perceptual encoding processes in the brain adaptively oscillate in their efficiency to sustain cognitive performance over prolonged periods of time.

Methodological Challenges and Implications

Taken together, converging evidence across behavioural, electrophysiological and neuroimaging studies portray rhythmic fluctuations of attentional and behavioural performance in humans while performing experimental tasks, suggesting that disengagement can reliably be captured in laboratory settings. Nevertheless, methods capable of inferring latent engagement states from purely behavioural data offer important complementary advantages to the field. Such behavioural methods capture the functional consequences of fluctuating attention more directly than neural correlates, and their ease of implementation provide practical benefits for researchers and clinicians over electrophysiological or neuroimaging approaches.

Additionally, the choice of experimental tasks used to study fluctuating engagement should be carefully considered. All empirical studies discussed so far rely on some form of continuous performance task (CPT), which are explicitly designed to induce disengagement through repetitive

response demands (Manly et al., 1999). This raises the question of whether naturally occurring periods of disengagement also arise in tasks not specifically designed to elicit them, especially given that unvarying, familiar, repetitive, and undemanding task environments have been shown to promote disengagement (Smallwood & Schooler, 2006; Cheyne et al., 2009). This question extends beyond sustained attention paradigms to experimental settings across other research domains.

Establishing whether disengagement is captured in such tasks is important for a number of reasons. First, disengagement is consistently associated with impaired task performance (Smallwoord & Schooler, 2006). Failing to detect periods of disengagement therefore risks misestimating participants' abilities, which is particularly consequential in high-stakes settings such as clinical assessment or cognitive screening batteries. Second, failing to account for disengagement risks confounding core psychological relationships under investigation. When examining the association between variables *X* and *Y* on a given task, any apparent relationship may simply reflect fluctuations in attention, rather than meaningful interactions between the constructs of interest. Moreover, in psychophysical and decision-making paradigms, undetected attentional lapses can lead to serious misestimation of psychometric parameters (Carandini & Churchland, 2013; Prins, 2012). Finally, uncaptured periods of disengagement violate key assumptions of widely used statistical models such as regression and ANOVA (Gunawan et al., 2021), which assume that repeated observations are independent and identically distributed. If those observations are more likely to be generated by multiple latent states, these assumptions are clearly untenable, thereby undermining the validity of the resulting findings.

These methodological constraints have motivated the development of alternative modelling frameworks that can infer latent engagement states more flexibly. One particularly promising tool has recently surfaced in perceptual decision-making.

Latent State Modelling during Perceptual Decision-Making

In recent years, computational modelling has emerged as a powerful approach for uncovering the cognitive mechanisms underlying behavioural responses across a wide range of research domains. Computational models can break down overt behaviour into latent subprocesses which can then be tested by fitting the models to experimental data. In the context of fluctuating task-engagement, Generalised Linear Model-Hidden Markov Models (GLM-HMMs) have proven particularly promising for inferring latent states underlying decision-making behaviour. Conceptually, the GLM-HMM assumes that behaviour arises from a small number of hidden cognitive states, each reflecting a distinct decision-making strategy, between which participants transition over time. By linking behavioural choices to external task variables within each state, and modelling transition probabilities between them, the model captures both *how* and *when*

engagement fluctuates. Recent investigations using this approach provide compelling evidence that fluctuations in task-engagement extend beyond sustained attention paradigms.

A landmark study by Ashwood et al. (2022), applied a GLM-HMM to mice performing perceptual decision-making tasks. Their analysis demonstrated that mice rely on multiple discrete decision-making strategies, with varying degrees of task-engagement, that persist over tens to hundreds of trials. Specifically, the mice alternated between an ‘engaged strategy’, in which choices were strongly driven by sensory evidence, and other less-engaged states characterised by biased or weakly stimulus-dependent responding. Since then, similar latent state structures have been reported across diverse methodologies.

Hulsey et al. (2024) further advanced this line of work by extending the GLM-HMM to incorporate non-responses, and linking transitions between decision-making strategies to changes in arousal. Applying this extended framework to auditory and visual discrimination tasks in mice, they recovered a similar optimal, stimulus-driven state and two biased response states. Interestingly, the extended model also revealed a *fully disengaged state* characterised by the complete absence of responses. Additionally, state occupancy was systematically related to physiological arousal: the likelihood of optimal state occupancy followed an inverted-U relationship with pupil diameter and uninstructed movement, whereas disengagement exhibited the opposite pattern. Moreover, reductions in arousal variability reliably preceded transitions into the optimal state, such that they could be used to predict the onset and offset of latent decision-making states.

Collectively, these studies demonstrate a) the potential of latent-state modelling to infer internal engagement dynamics directly from behavioural data, and b) that fluctuations in task-engagement extend beyond restrictive sustained attention paradigms, offering a powerful framework for identifying and characterising periods of (dis)engagement during experimental tasks. However, these findings are exclusively based on rodent behaviour. Whether they extend to humans remains uncertain, especially given evidence that human decision-making is considerably more stable than that of rodents (Roy et al., 2021; Chakravarty et al., 2024).

Human Findings

Critically, Ashwood et al. (2022) also applied the GLM-HMM to the choice behaviour of humans performing a visual discrimination task and found far less conclusive results. Human decision-making behaviour was characterized by two strategies, both of which relied heavily on the sensory stimulus and only differed in their bias toward one of two decision alternatives. Notably, no clear 'engaged' or 'disengaged' states were identified, reaffirming the need to investigate engagement dynamics in humans during non-CPTs.

Since the start of the present work, one study has applied a GLM-HMM to human choice behaviour during perceptual decision-making (Zhang & Kool, 2025). In this experiment, participants performed a random dot motion task in which 90% of trials featured the same ('dominant') motion direction, to mimic the repetitive response requirement of sustained attention paradigms, with thought probes interrupting the task every 30 trials. The authors first characterised behavioural profiles using a series of regression-based models and thought probes and found that participants rely more on the task's bias when off-task, leading to faster and more accurate responses on dominant trials but increased errors on rare ones. They then fit individual GLM-HMMs to choice behaviour, using the previously defined behavioural profiles to label the two resulting states as on- and off-task. Finally, the authors performed additional analyses to validate the model-predicted states and determine whether they truly capture attentional fluctuations. Using linear mixed models, the authors established that a) model-predicted states aligned with self-reported focus levels, b) the average probability of the on-task state decreased throughout the experiment and c) off-task periods were marked by faster reaction times on dominant trials, a behavioural signature of reduced attentional control.

While their findings reinforce the use of GLM-HMMs to infer latent engagement states during perceptual decision-making tasks, even in human subjects, a few problems appear in the context of our study. First, because the task mimics the repetitive demands of sustained attention paradigms, it remains unclear whether these findings generalise to less monotonous tasks. Second, the authors use thought probes to asses self-reported focus. Although they were only used to validate the novel modelling approach in this particular study, their occurrence could have disrupted the natural flow of attentional states (Schubert et al., 2020). Finally, we believe that the authors' analytical approach introduces an inherent circularity, as self-reported focus was used to both define the behavioural profiles that guided the labelling of model-predicted states and subsequently to validate those same states.

Here, we address these issues by fitting a GLM-HMM to human perceptual decision-making data collected under more natural response conditions and without the use of thought probes.

Present Study

The present study is part of a larger project that aims to establish whether humans exhibit multiple (dis)engagement states in perceptual decision-making, as inferred using a GLM-HMM. The research will be conducted on a large body of data, comprising several datasets and perceptual tasks, allowing for a rigorous examination of this research question and an assessment of whether such states generalise across tasks or depend on the specific experimental context.

The present study serves as preparatory and exploratory investigation within this ongoing project. The central research question is: "*Does the GLM-HMM provide a robust and reproducible*

framework for identifying and characterising latent engagement states in human perceptual decision-making?" Specifically, it aims to i) identify suitable datasets from the Confidence Database, ii) explore modelling options and finalize the analysis pipeline on pilot data and iii) replicate the findings on human data reported by Ashwood et al., (2022).

The results obtained from the pilot data will only be discussed descriptively, as no formal hypotheses are formulated for this exploratory phase. For the replication analysis of Ashwood et al. (2022), we expect to reproduce the same pattern of results as reported by the authors. Specifically, we hypothesize that: i) a two-state model will provide the best explanation of the data and ii) no apparent differences in task-engagement will emerge between states. These hypotheses are based on previous evidence of distinct engagement states observed in humans during sustained attention paradigms (Esterman et al., 2012; Cai, et al., 2020; Gaillard et al., 2021), perceptual decision-making in rodents (Ashwood et al., 2022; Hulsey et al., 2024) and humans (Zhang & Kool, 2025), as well as the specific results of the study we are replicating. Successfully replicating these patterns using our analysis pipeline would demonstrate the robustness and validity of our framework.

Beyond replication, the present study extends the original analysis by examining subjects' state-specific GLM weights to characterise individual engagement profiles, and by conducting additional validation analyses following the approach of Zhang and Kool (2025).
These additions allow for a more comprehensive assessment of fluctuating engagement and behavioural performance during human perceptual decision-making, laying the groundwork for future research to apply this framework across a wider range of datasets and experimental contexts, as planned in subsequent stages of the research project.

The current work represents an important step towards building a robust and generalizable framework for modelling engagement dynamics in perceptual decision-making. Understanding how engagement fluctuates during perceptual decision-making is of central importance to psychological and neuroscientific research. Many widely used paradigms implicitly assume stable cognitive engagement across trials, yet unaccounted attentional fluctuations risk misestimating participants' abilities, confounding experimental relationships, and violating key assumptions of statistical models (Gunawan et al., 2021). Demonstrating that engagement fluctuates, and that these fluctuations can be identified directly from behavioural data, has broad implications. Beyond advancing theoretical models of decision-making to account for temporal structure in sequences of decisions (Urai, 2025), this framework improves accessibility of attentional dynamics for researchers and clinicians. This enables large-scale scientific investigations, re-analysis of a large body of existing behavioural datasets, and potential applications in cognitive assessment and interventions for populations with attentional difficulties

Methods

Design

The present study used a cross-sectional observational design based on secondary analysis of two previously collected datasets. No experimental manipulation or random assignment was conducted. Instead, latent cognitive states underlying performance in the behavioural tasks were inferred using a GLM-HMM, and the resulting parameters served as the basis for subsequent analyses. The work forms part of a larger research project on engagement dynamics in humans during perceptual decision-making. No new data were collected for the current study, but existing datasets were selected from the Confidence Database (Rahnev et al., 2020) to serve as research sample for future stages of the project.

Participants

The final sample consisted of 30 subjects, drawn from two separate datasets. Twenty-seven subjects (10 male, aged 23±5.2 years) were taken from data by Urai et al. (2017), described in Ashwood et al. (2022). The remaining three participants (2 male; age range: 25-30 yrs) were taken from Zylberberg et al. (2016) and were used exclusively as pilot data. No exclusion criteria were reported for both studies.

A sensitivity power analysis (α = .05, *power* = .80, two-tailed) indicated that the available sample of 27 participants allowed for detection of a minimal effect size of Cohen's dz = 0.56, meaning the replication analysis was sufficiently powered to detect medium within-person effects.

Ethical Approval

All procedures in the datasets used in this study were approved by the relevant institutional review boards. The procedures from Zylberberg et al. (2016) were approved by the Institutional Review Board of Columbia University, and procedures from Urai et al. (2017) by the Ethics Committee at the University of Amsterdam. All participants provided informed consent prior to participation. Further details are available in the original publications (Urai et al., 2017; Zylberberg et al., 2016).

Procedures

The study comprised three components; i) selecting suitable datasets for the future research sample, ii) exploring modelling options on pilot data and iii) replicating the analysis on human data reported in Ashwood et al. (2022).

First, we examined and selected datasets from the Confidence Database (Rahnev et al., 2020), a collection of datasets that examine the relation between perceptual decisions, confidence

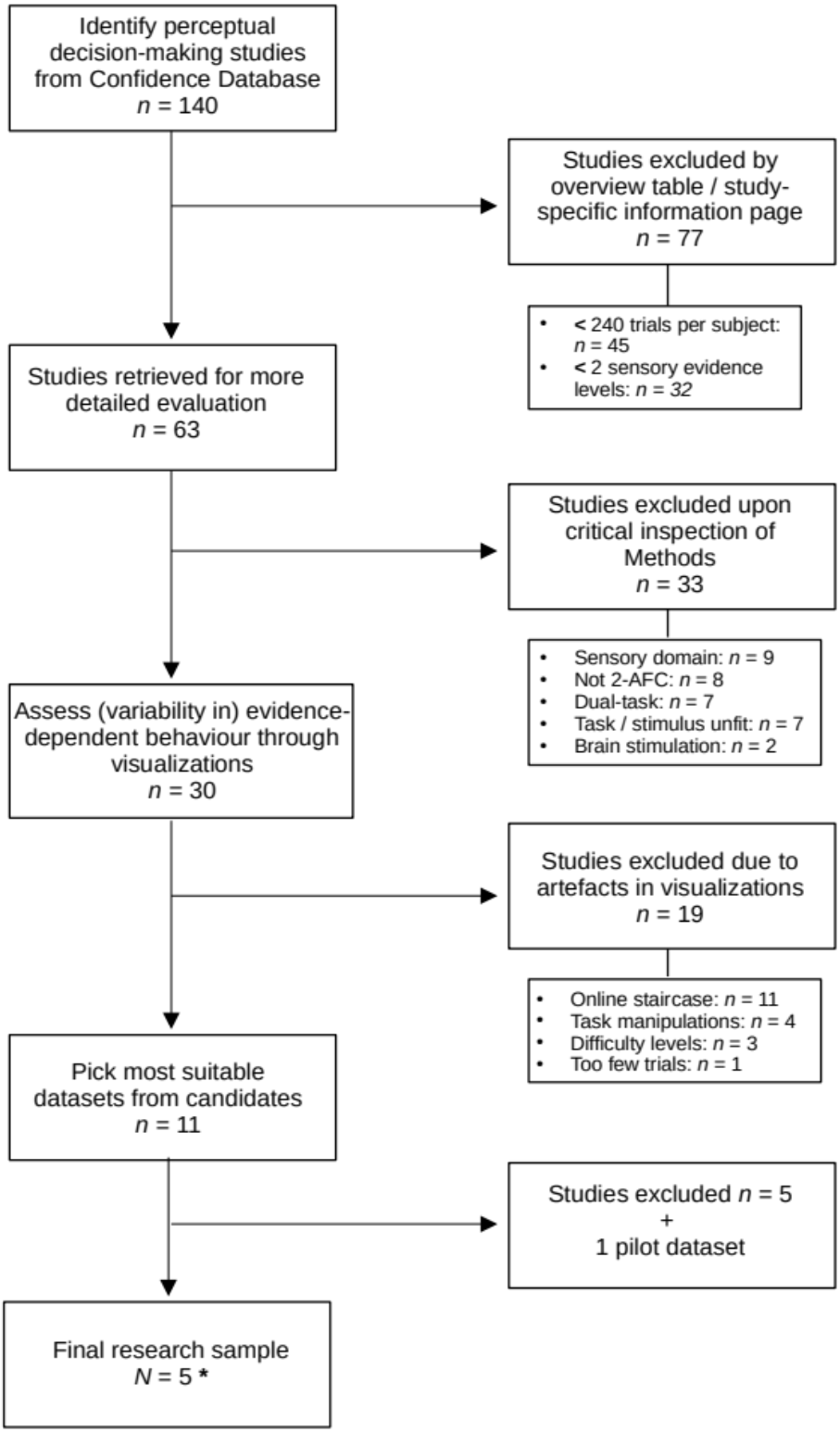
ratings and underlying cognitive processes. Confidence ratings are not relevant to this research, so they were excluded from analyses. To ensure a broad representation of perceptual tasks, we aimed to include at least three datasets per task. Datasets were further selected to meet the assumptions and requirements of the GLM-HMM and to ensure robust estimation of model parameters. Specifically, selected datasets (a) employed a two-alternative forced-choice (2AFC) design, (b) included multiple discrete levels of sensory evidence, and (c) contained a sufficient number of trials per subject for model fitting, ideally exceeding 400. In a subsequent selection step we assessed evidence-dependent variability in choices and RTs using psychometric and chronometric plots, respectively, alongside session-wise performance trajectories. After careful inspection, many datasets were found unsuitable for our research design, so one additional dataset collected in our lab at Leiden University (Enwereuzor et al., in prep.) was added to the sample. A full overview of the dataset selection procedure is provided in Figure 1.

Six datasets, comprising a total of 162 participants were selected, along with one pilot dataset. A sensitivity power analysis ($\alpha = .05$, power = .80, two-tailed) indicated that this sample size allows for the detection of a minimal effect size of Cohen's $dz = 0.22$, meaning the future analyses will be sufficiently powered to detect small-to-medium within-person effects. Table 1 summarizes the final research sample and key characteristics relevant to the ongoing research program.

Second, we prepared the pilot data for model exploration. This involved retrieving the datasets from the Confidence Database, performing standard pre-processing (e.g., computing covariates, aligning stimulus and response codes), and formatting the data for input to the GLM-HMM. These steps allowed us to explore modelling options and examine the effect of model priors on independent data, before applying the analysis pipeline to the unseen research sample.

Third, we replicated the analysis of Ashwood et al. (2022) using the finalized pipeline. The data was pre-processed in the same manner as the pilot data, and GLM-HMMs were fitted to each subject's data to infer latent engagement states.

Figure 1
Overview of the dataset selection procedure



Note. *Five datasets were selected from the Confidence Database, one additional study collected in our lab at Leiden University was added to form the final sample of six datasets

Table 1
Characteristics of the selected research sample

Dataset	Task	Stimuli	*n* subjects	*n* trials	Feedback	Blocks	Diff. levels
Desender et al. (2021)	Dot-discrimination	RDK	32	540	After block (ACC & RT)	9 blocks, 60 trials	4
Hellman et al. (2023)	Dot-discrimination	RDK	42	640	Trial-by-trial error feedback	8 blocks, 80 trials	5
Law (in prep.)	Dot-discrimination	RDK	16	960	After block (ACC & cum.)	16 blocks*, 60 trials	5
Rausch (2016)	Orientation discrimination	Low-contrast grating	20	378	Trial-by-trial error feedback + after block (% errors)	9 blocks, 42 trials	6
Rahnev et al. (2013)	Orientation discrimination	Low-contrast grating	12	700	After block (ACC)	5 blocks, 140 trials	4
Enwereuzor et al. (in prep.)	Contrast discrimination	Low-contrast gratings	40	600	Trial-by-trial error feedback	1 block	5

Note. The number of trials refers to the number of *usable* trials included in the present analyses. Some studies originally contained additional experimental conditions that were excluded due to task-irrelevant manipulations. An asterisk (*) indicates that participants took brief breaks between blocks.

Operationalisation

The central research question of this project is whether humans alternate between latent engagement states during perceptual decision-making. Engagement refers to the subjects' attention levels and associated task performance. When engaged, participants attend closely to stimuli, follow task instructions and respond with high accuracy.

In this study, engagement is operationalised through state-specific GLM weights. Following Ashwood et al. (2022), four covariates were included in the model: sensory evidence, bias, previous choice and win-stay/lose-switch (WSLS). Sensory evidence reflects the task-relevant decision strategy, as the correct response on each trial depends solely on the presented evidence. The remaining covariates capture task-irrelevant strategies: bias reflects a consistent preference for one response alternative independent of evidence; previous choice represents the simple repetition of the prior response; and WSLS combines choice and outcome history by repeating a previous correct choice or switching following an incorrect one.

Thus, engagement was operationalised as the weighting of task-relevant (sensory evidence) versus task-irrelevant (bias, previous choice, WSLS) covariates within each state. While task-irrelevant strategies may occasionally lead to correct responses, they do not reflect task-relevant

decision-making and are therefore considered ‘disengaged’. Each covariate represents a distinct decision strategy, allowing the model to distinguish between engaged and disengaged patterns of behaviour throughout the experiment. A detailed description of the GLM-HMM specification and fitting procedure is provided in the *Measurements* section.

To assess the adequacy and reliability of this operationalisation, model performance was evaluated using cross-validated log-likelihood, a measure of how well the model explains the observed data. For each participant, the performance of *K*-state GLM-HMMs was compared against a one-state model (i.e. ordinary logistic regression). A multi-state model was only retained for further analysis when it improved explanatory power compared to one state. See *Fitting Procedure* for further details.

Measurements

Motion Discrimination Task

Both studies administered a two-alternative forced choice (2-AFC) motion discrimination task, using a Random Dot Kinematogram (RDK), in which a cloud of randomly moving dots are presented, with a subset of dots moving coherently in the same direction. Subjects are asked to decide in which of two directions these target dots are moving. Trial difficulty is controlled by manipulating the proportion of coherently moving dots, i.e. motion coherence, which is picked randomly on each trial out of a few discrete levels. Stimuli are repeated for a large number of trials to probe perceptual decision-making processes.

In Zylberberg et al. (2016), the researchers manipulated the within-trial consistency of motion coherence: in low-volatility trials, the probability of coherent motion remained constant, whereas in high-volatility trials, this probability varied over time. Low- and high-volatility trials were uncued and randomly interleaved. Mean levels of motion coherence used were 0%, 3.2%, 6.4%, 12.8%, 25.6% and 51.2%. Participants indicated the perceived motion direction (left or right) and their confidence simultaneously by selecting one of two crescent-shaped targets presented on either side of fixation. The side indicated perceived direction, and the vertical position within the crescent reflected confidence, ranging from guessing at the bottom to full certainty at the top. Participants completed between 1,536 and 2,107 trials, divided into blocks of 96 trials. Each subject performed four to six blocks per day over four days to reach the total number of trials. The authors’ goal was to test whether selectively increasing evidence volatility, without changing mean evidence, would produce changes in choice accuracy, RT and confidence, as expected under a bounded accumulation model. While the volatility manipulation and small sample size make this dataset

unsuitable for inclusion in the research sample, the small sample size offered practical advantages for model fitting, so it was used as pilot data.

In Urai et al. (2017), subjects judged the difference in motion coherence between two successively presented RDKs: a constant reference stimulus (70% motion coherence) and a test stimulus with varying motion coherence levels. The difference between motion coherence of test and reference was taken from three sets: easy (2.5%, 5%, 10%, 20%, 30%), medium (1.25%, 2.5%, 5%, 10%, 30%) and hard (0.625%, 1.25%, 2.5%, 5%, 20%). All subjects began with the easy set. They switched to the medium or hard set if their psychophysical threshold fell below 15% or 10%, respectively, in a given session. Motion coherence differences were randomly shuffled within each block. Participants completed one practice session and five main experimental sessions, each comprising 500 trials organized in blocks of 50. Subjects were asked to indicate whether the test stimulus contained higher or lower motion coherence than the reference stimulus. The goal of the study was to test whether pupil-linked arousal signals reflect decision uncertainty, as predicted by computational models of dynamic decision-making, and whether these signals predicted changes in subsequent decision behaviour, including serial choice bias.

Generalised Linear Model – Hidden Markov Model (GLM-HMM)

To identify latent engagement states during decision-making, we fitted a GLM-HMM (Ashwood et al., 2022) to each subject's choice behaviour. The GLM-HMM consists of two components: a hidden Markov model (HMM) that governs the transitions between latent states and a set of state-specific generalized linear models (GLMs), that define the decision-making strategy employed in each state.

For K latent states, the HMM includes 1) a $K*K$ transition matrix, which specifies the probability of transitioning from one state to another, and 2) a distribution over initial states, represented by a K-element vector, whose elements sum to 1. The 'Markov' property of the HMM reflects the assumption that the state on any trial depends only on the state from the previous trial, and the 'hidden' property refers to the fact that states are latent or hidden from external observers. The model further assumes that the probability of transitioning between states is constant for all time-points within a session.

To describe the state-dependent mapping from inputs (e.g., stimulus, trial history) to decisions, the GLM-HMM contains K independent Bernoulli GLMs, each defined by a weight vector that determines how inputs are integrated within a specific state. Each latent state corresponds to a distinct decision-making strategy that the subject may switch between over time. These states capture fluctuations in engagement or cognitive strategies that influence choices but

cannot be directly measured. The number of states that best describes the data will be determined through cross-validation, as outlined further below.

Model parameters, including the transition probabilities, initial state distribution, and state-specific GLM weights, were estimated using the expectation-maximisation (EM) algorithm. The EM algorithm is a commonly used iterative method to estimate values of latent variables. It alternates between two steps:

1. *Expectation step* (*E-step*)*:* estimates the probability that the subject was in each of the possible states at every trial, given the current model, and
2. *Maximization step* (*M-step):* updates the model parameters to best fit the data (i.e. maximize the log-likelihood), given the estimated state probabilities from the E-step.

The steps are repeated until the parameters converge to a stable solution. The EM algorithm does not guarantee reaching the global optimum; so to reduce the risk of the EM algorithm getting stuck in poor local optima, the fitting procedure was repeated with twenty different random initializations.

The GLM-HMM was fit using the Dynamax package (Lindeman et al., 2025) in Python, as opposed to SSM (Linderman et al., 2020) that was used in the analysis by Ashwood et al. (2022), because SSM is not compatible with the current Python version and is no longer actively maintained, unlike Dynamax. Prior to model fitting, the datasets were pre-processed to ensure consistency and compatibility with the GLM-HMM framework. Stimulus and response variables were recoded to binary format, variable names were standardised across datasets, and sensory evidence was normalised to a common scale between 0 and 1. Trial-history covariates (i.e. previous choice and WSLS) were computed based on individual trial sequences Next, the GLM was fitted using the multi-stage fitting procedure outlined below.

Assessing assumptions in GLM-HMMs is challenging, as many of these are conceptual rather than directly testables. The core premise, however, is that the input data is generated by multiple latent states with distinct properties. To determine whether our data justified fitting a GLM-HMM, we compared the performance of *K*-state GLM-HMMs to that of a one-state model, using cross-validation. Improved held-out log-likelihood for models with additional states indicates that the latent-state structure is warranted. If no improvement is observed over the single-state model, this suggests the data are better described by a unitary strategy, so no GLM-HMM would be applied to that data.

Priors and Hyperparameters

To further avoid overfitting, prior distributions are imposed on model parameters. A few deviations from Ashwood et al. (2022) should be noted here. First of all, the original analysis imposes a Gaussian prior over GLM weights (zero mean, $\sigma^2 = 0.2$) directly regularising the

objective function during training. We were unable to replicate this component, since Dynamax does not allow imposing prior distributions over emission weights. We instead approximated the effect of a weak prior, by adding small, structured noise (zero mean, $\sigma^2 = 0.2$) to the emission parameter initialisations. This only affects starting values rather than the objective function, loosely mimicking prior regularization and reducing overfitting. We acknowledge this methodological divergence and discuss potential implications in the *Discussion* section.

In addition, Dirichlet priors were placed on each row of the transition matrix and the initial state distribution. The concentration parameter controls how evenly probability is spread across possible state transitions and initial states, such that higher values encourage more uniform transitions and prevent overfitting to sharp, unlikely switches. Dynamax supports Dirichlet priors over initial state distributions, so this aspect of the original pipeline was retained. However, Ashwood et al., (2022) used a concentration parameter of $\alpha = 1$ in the global fit to explore the full parameter space, and $\alpha = 2$ in the subject fit to provide mild regularization of state transitions. We retain the default value of $\alpha = 1.1$, because in our analysis the stickiness parameter is tuned separately to bias self-transitions, making additional regularization via α unnecessary.

Finally, the 'stickiness' parameter governs the tendency of the model to remain in the same state across trials. Low stickiness values make the model non-sticky, meaning it does not favour staying in the same state over switching to others, whereas high values bias the model toward staying in the same state. In Ashwood et al. (2022), stickiness was set to zero, meaning no bias toward remaining in the same state was imposed. In our exploratory fits on the pilot data, however, we consistently observed excessive switching between states for stickiness = 0. Therefore, we conducted a grid search for stickiness $\in \{0, 1, 2, 5, 10\}$, and evaluated performance using cross-validation, see *Results*. For the replication analysis, we present the results for both the original stickiness value from Ashwood et al. (2022), and the stickiness value that produced best model performance on the pilot data.

Fitting Procedure

We employed a multi-stage fitting procedure to ensure robust parameter estimation (see *Algorithm 1*). First, the data of all subjects was concatenated, and a GLM (one-state GLM-HMM) is fit to that data using maximum likelihood estimation. The resulting weights are then used to initialize the GLM weights of a *K*-state GLM-HMM, which is also fit to the concatenated data (i.e. global fit). This model is initialised 20 times, with the addition of Gaussian noise, to ensure that initialised states are distinct and vary across initialisations. The model with the largest log-likelihood is selected as global fit estimate and is subsequently used to initialise 20 subject-specific GLM-HMMs, in two further fitting rounds.

First, we conducted a cross-validation procedure to determine the number of states that best describe the data. Each subject's data was split into folds based on measurement sessions, as both datasets included multiple sessions per subject. This procedure resulted in four folds for the Zylberberg et al. (2016) dataset and five folds for the Urai et al. (2017) dataset. Then, in each iteration, all but one fold were used to train successive models with K states, and the held-out fold (i.e. validation set) is used to evaluate each model's performance. This process is repeated until every fold has been used as validation set once. The test log-likelihoods were then aggregated across folds to compare model performance and determine the number of latent states that best describe the data.

Finally, the best performing model with K states, is refit to the complete data of every subject to obtain reliable parameter estimates. Unlike earlier implementations, we did not order states on any external variable. Instead, the multi-stage fitting procedure differentiated states in a fully data-driven manner, without requiring additional assumptions about their characteristics or relationships.

Algorithm 1 – Multi-stage GLM-HMM fitting procedure

---

1. Fit GLM (one-state GLM-HMM) to concatenated data from all subjects
2. Fit global GLM-HMM to concatenated data:
3. **for** $K \in \{2, \ldots 4\}$ **do**
4. **for** init. $\in$\{1, ... 20\} **do**
5. Initialize K-state GLM-HMM using noisy GLM weights
6. Run EM algorithm until convergence
7. **end for**
8. **end for**
9. Select global model with highest log-likelihood for each K
10. Fit separate GLM-HMM to each subject by initializing with global fit:
11. **for** each individual subject **do**
12. **for** $K \in$\{2, ... 4\} **do**
13. **for** init. $\in$\{1, ... 20\} **do**
14. Initialize K-state GLM-HMM using best global
15. GLM-HMM parameters for this K
16. Run EM algorithm within cross-validation folds
17. Aggregate CV log-likelihoods to evaluate K
18. **end for**
19. **end for**
20. Refit best-performing K-state model to each subject's full data

---

Statistical Analysis

Assumptions of the GLM-HMM are discussed in the *Measurements* section. To assess whether latent states reflected distinct behavioural strategies, paired-samples t-tests were conducted on the state-specific GLM weights across subjects, focusing on the evidence (stimulus) weight as an index of engagement. For the bias covariate, absolute weight differences between states were used, since negative bias values merely reflect a preference for the opposite response alternative rather than reduced bias. To control the family-wise error rate across multiple comparisons, p-values were adjusted using the Holm–Bonferroni correction. One participant was excluded from this analysis due to unstable weight estimates, resulting from spending 99.8% of trials in a single state.

To validate model-predicted states and further characterise behavioural differences, linear mixed-effects models were fitted to RTs, accuracy rates and engagement probability. RTs and accuracy were separately predicted from inferred states, including random intercepts for subjects and random slopes for sessions, to account for baseline differences and within-subject variability. The probability of being in the engaged state was modelled as a binary outcome across trials, with trial number as a predictor and random intercepts for each subject and session, allowing us to capture gradual changes in engagement over time.

Results

To evaluate and refine the statistical pipeline on the pilot dataset, we first examined how the choice of model priors affected parameter estimation and model convergence. Subsequent sections report the final results on the pilot data, followed by the replication analysis of Ashwood et al. (2020), conducted using both the original prior specifications and the set of best-performing prior values identified in the pilot data.

Exploratory Model Fits and Effects of Priors

In initial model fits, we employed Dynamax's default prior settings and initialized per-subject GLM weights using logistic regression. This approach revealed two key issues. First, fitted GLM weights failed to converge consistently across subject-specific initialisations (see Fig. 2a). Non-convergence suggests that the optimization algorithm is not locating a stable global optimum, but instead settling in multiple local optima. Furthermore, this procedure frequently produced implausibly large weight estimates, reinforcing the conclusion that the model was not reaching a well-defined optimum.

Second, differences in GLM weights between states were minimal. Across subjects, variability in weights was greater *between* individuals than between states *within* individuals, suggesting poor identifiability of the state-dependent structure in the model (see Fig. 2b).

Figure 2

Fitted GLM weights using logistic regression initialisation

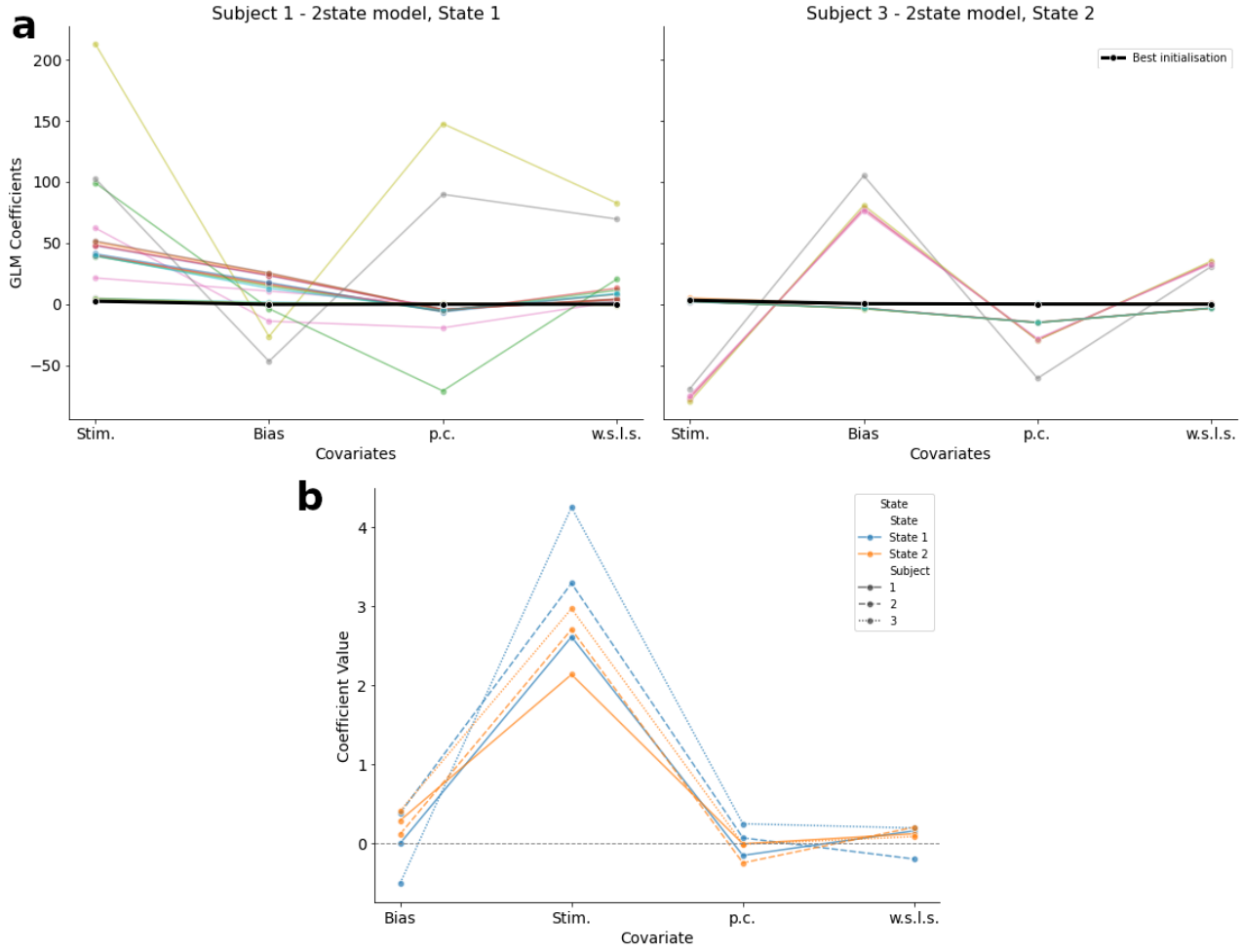
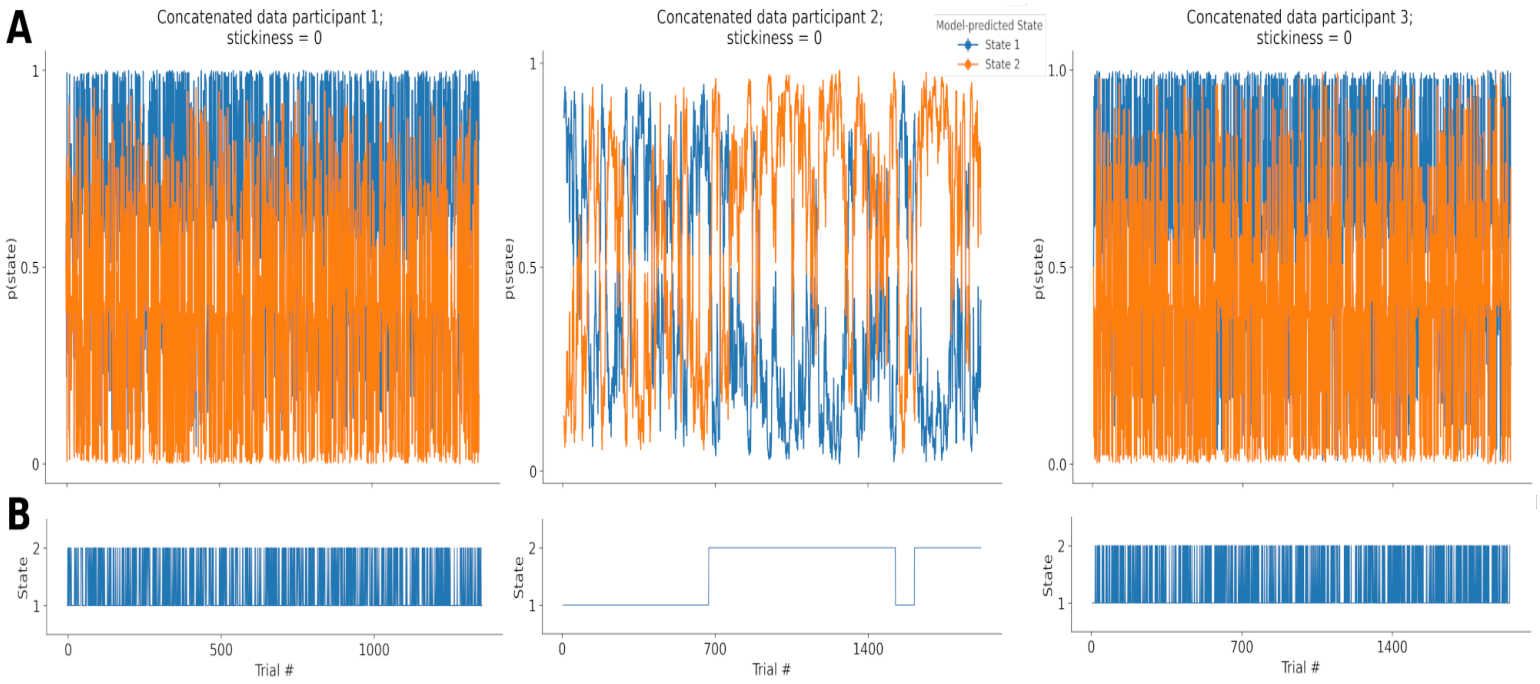


Figure 3

Inferred State Dynamics With Stickiness Set to 0



Note. Results are based on model fits to concatenated data from all subjects across all sessions. (**a**) Trial-by-trial posterior state probabilities. (**b**) Inferred state sequences.

To address this, we examined the influence of the stickiness prior on state transitions. Stickiness penalizes rapid switching by increasing the self-transition probabilities of each state. We conducted a grid search for stickiness $\in \{0, 1, 2, 5, 10\}$, and evaluated performance using cross-validation. Figure 4 displays the improvement in log-likelihood relative to a one-state GLM on held-out validation data, for each stickiness level.

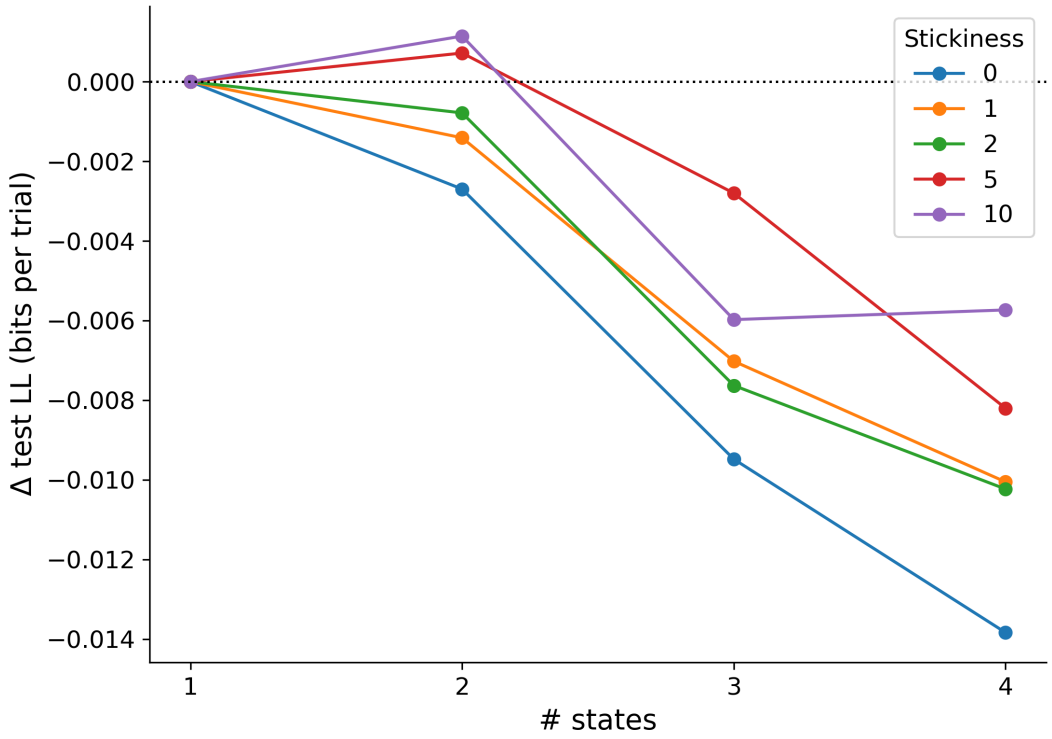
Model performance on the validation set was highest when stickiness was set to 10, yielding an improvement of 0.00114 bits per trial over logistic regression. The next best value was 5, whereas all other stickiness values resulted in substantially poorer performance relative to the baseline. Although the improvement in log-likelihood may appear modest, an increase of 0.00114 bits per trial, renders a dataset of 480 trials 1.46 times more likely to have been generated by a two-state GLM-HMM than by a single-state GLM.

Regarding noise added to each initialisation, SD values larger than 0.2 prevented convergence. Conversely, lower noise values produced GLM estimates that remained contiguous to the global fit estimate, suggesting insufficient flexibility to capture individual differences across subjects.

Based on these exploratory fits, we selected a value of 10 for the stickiness parameter and 0.2 for added noise. The concentration parameter was retained at the default value of 1.1

Figure 4

Cross-Validated Improvement in Log-Likelihood Relative to One-State Baseline



Note. Log-likelihood was computed as the average across subjects and compared against an ordinary logistic regression model. The difference was divided by the number of trials per subject to yield improvement in log-likelihood per trial, consistent with Ashwood et al. (2022).

Surprisingly, increasing stickiness did not always lead to increased self-transitions. In some cases, higher stickiness values produced more frequent switching between states in the inferred sequences. For example, with stickiness set to 5, one of the three subjects occupied a single state for nearly the entire session, whereas with stickiness set to 10, state occupancy was more evenly distributed (see Supplementary Figure 1).

This counter-intuitive effect highlights a central property of the optimization landscape underlying the EM algorithm. Changes to model priors alter the posterior landscape and can shift the optimizer towards different basins of attraction. As a result, higher stickiness values may shift the model towards qualitatively different solutions in which the emission parameters and transition matrix jointly explain the data with increased state-switching.

Final Results Pilot Data

As noted before, the pilot dataset (n = 3) was used exclusively to optimise the statistical pipeline rather than test hypotheses. Given the small sample size and exploratory purpose, no inferential statistics were conducted. Instead, the results are described descriptively to illustrate the behaviour of the model and inform parameter selection, see Figure 5.

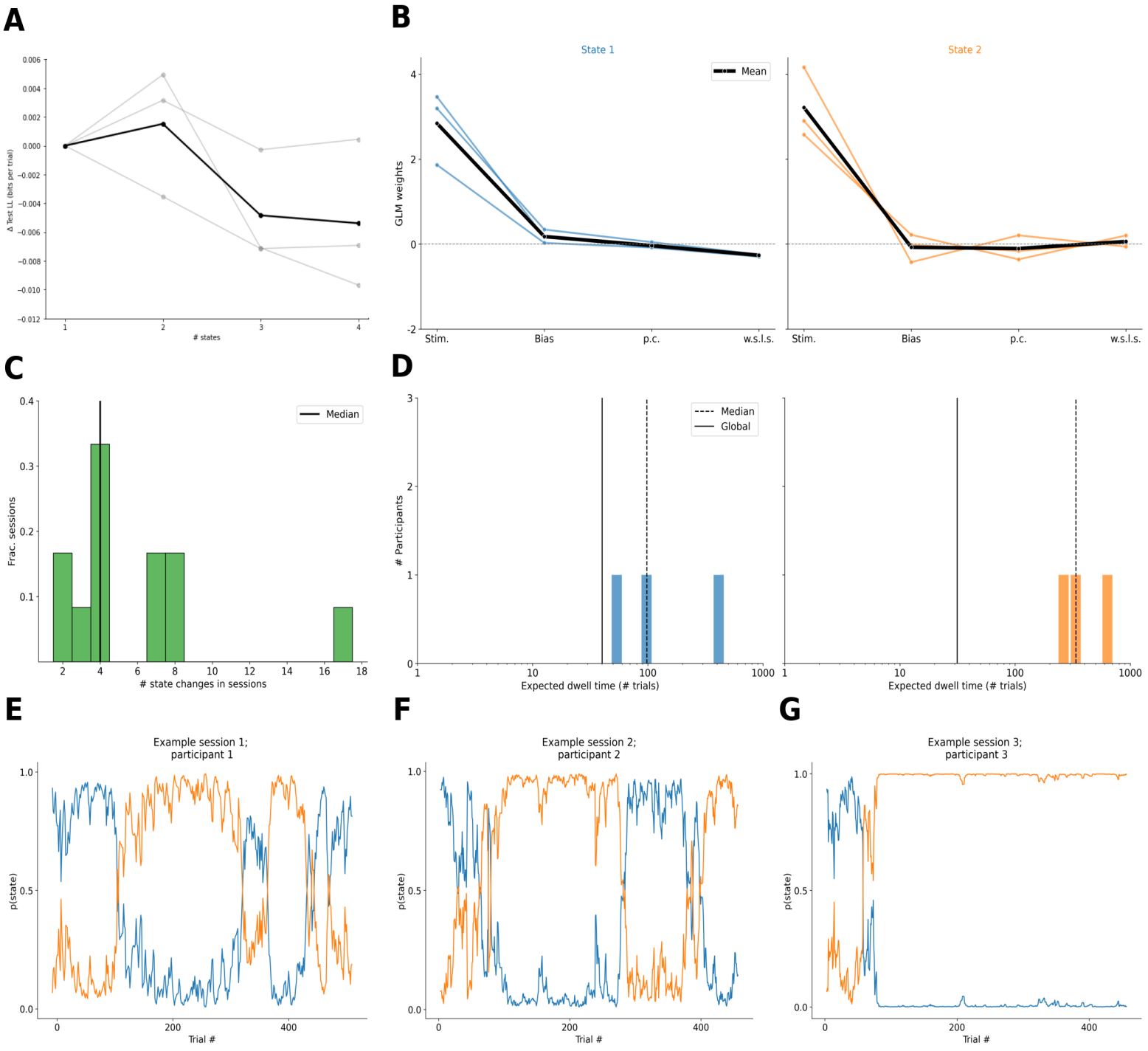
Cross-validation showed that the choice data of two out of three subjects was better explained by a two-state GLM-HMM than one-state GLM (Fig. 5a). The mean improvement across subjects was 0.00114 bits per trial, rendering a test dataset of 480 trials 1.46 times more likely to have been generated by a two-state GLM-HMM than by a single-state GLM.

Visual inspection of the fitted weights (Fig. 5b), appear to indicate some differences in engagement between states. The sensory evidence weight tended to be larger in state 2, reflecting increased engagement, whereas bias and WSLS weights were slightly more pronounced in state 1, suggesting larger disengagement.

Inferred state sequences were highly stable. Subjects exhibited few state transitions within sessions (Fig. 5c) and expected dwelling times were long (Fig. 5d), suggesting that once a state was entered, it tended to persist for many trials.

Finally, Figure 5e displays posterior state probabilities for three example sessions. The model assigned high confidence to the inferred state on most trials, suggesting a good identifiability of retrieved states. Furthermore, the stability of state sequences is also reflected in the long continuous epochs dominated by a single state.

Figure 5
GLM-HMM Results to Pilot Data

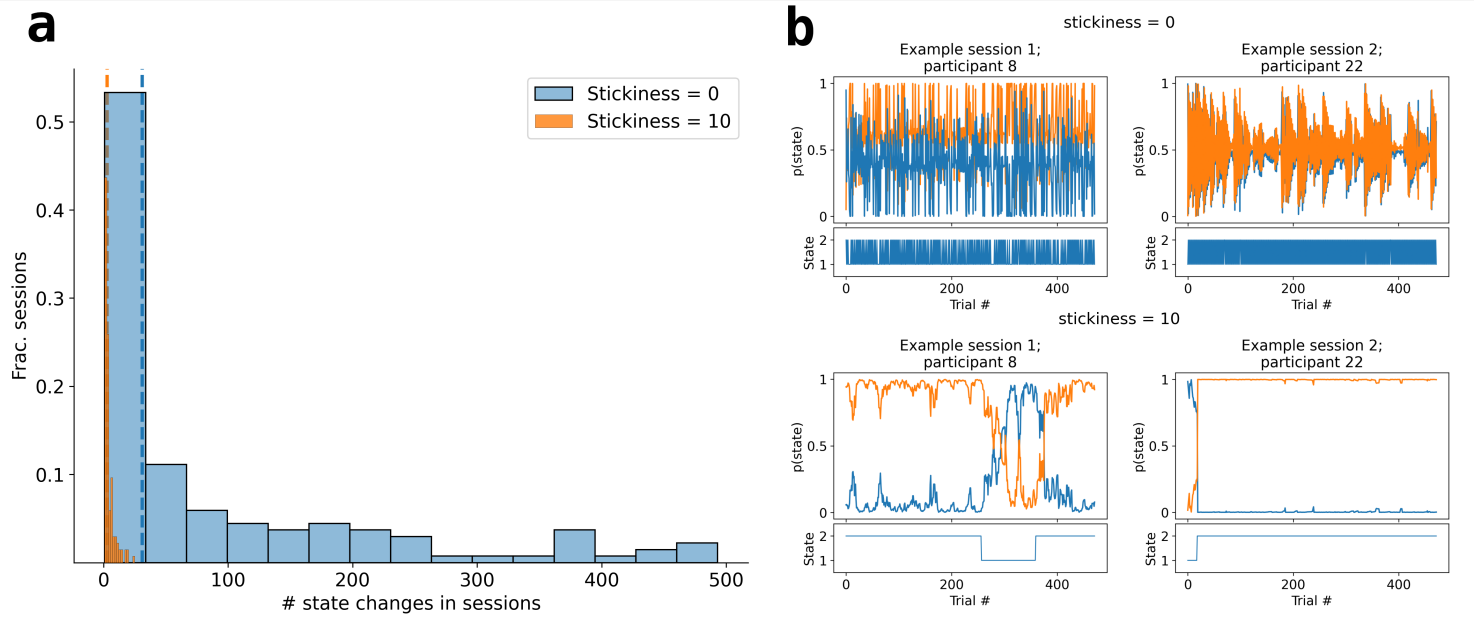


Note. Results from the GLM-HMM applied to the full participant dataset are shown across multiple measures. (**a**) Change in test set log-likelihood relative to a GLM for each of the 27 participants; black indicates the mean across participants. (**b**) Retrieved weights for a two-state GLM-HMM for each participant; black indicates the mean across participants. (**c**) Number of state changes per session obtained from posterior state probabilities; median session length is 500 trials, and black indicates the median across sessions. (**d**) Expected dwell times for each participant in each state, obtained from the inferred transition matrices; black dashed line indicates the median dwell time across participants, and black solid line indicates the global fit. (**e**) Posterior state probabilities for three example sessions corresponding to three different participants.

Ashwood Replication

Consistent with the pilot analysis, where higher stickiness values supported stable dynamics, applying a stickiness value of 0 produced largely unstable state dynamics in our replication analysis. Using this prior, 100 out of 135 sessions showed rapid switching between states and minimal dwelling times, indicating that the inferred states did not reflect meaningful temporal structure By contrast, reintroducing a stickiness value of 10 restored stable temporal dynamics, with extended dwelling times and markedly fewer state transitions. These dynamics are more consistent with those reported by Ashwood et al. (2022), so we picked this prior setting for the final analysis. Figure 6 illustrates the impact of the prior settings on the inferred state dynamics.

Figure 6
Effect of Stickiness Prior on Inferred State Dynamics



Note. (**a**) Number of state switches per session for different stickiness values. (b) Posterior state probabilities and inferred state sequences for the Urai (2017) dataset.

In our final analysis, cross-validation revealed that the choice data of 18 out 27 subjects was best explained by a two-state GLM-HMM. The mean improvement of the two-state GLM-HMM compared to one state was 0.012 bits per trial, making a test dataset of 500 trials 67 times more likely to have been generated by a two-state GLM-HMM (Fig. 7a).

We observed slight differences in engagement and behavioural strategies used in each state. Subjects relied more on sensory evidence in state 1 compared to state 2 ($t(25) = 2.63$, $p_{unc} = .015$, $p_{Holm} = .045$, Cohen's $d = 0.52$, 95% CI [0.09, 0.95]), reflecting a higher degree of engagement. Conversely, subjects used the WSLS strategy more strongly in state 2 compared to state 1 ($t(25) = -4.73$, $p_{unc} < .001$, $p_{Holm} < .001$, Cohen's $d = -0.93$, 95% CI [-1.38, -0.47]), potentially

reflecting higher disengagement. While there was a numerical trend toward larger *absolute* biases in state 2 relative to state 1 ($t(25) = -2.05$, p_{unc} =.051, p_{Holm} =. 102, Cohen's d = -0.402, 95% CI [-0.82, 0.02], visual inspection of the fitted GLM weights suggests that the states differ primarily in their directional bias towards one alternative rather than magnitude (Fig. 7b), consistent with Ashwood et al. (2022). Finally, subjects' choices were independent of their response on the previous trial, with no differences observed between states ($t(25) = 0.998$, p_{unc} = .328, p_{Holm}= .328), and fitted GLM weights close to zero in both states.

Inferred state dynamics were highly stable across subjects. Subjects typically remained in the same state for extended periods (Fig. 7d), but switched between states several times within a session (Fig. 7c). Posterior state probabilities further confirmed this stability, with high confidence assigned to the inferred state on most trials (Fig. 7e). Overall, these results indicate that the two-state GLM-HMM captured fluctuations in engagement throughout the experiment, and closely mirror the results reported by Ashwood et al. (2022).

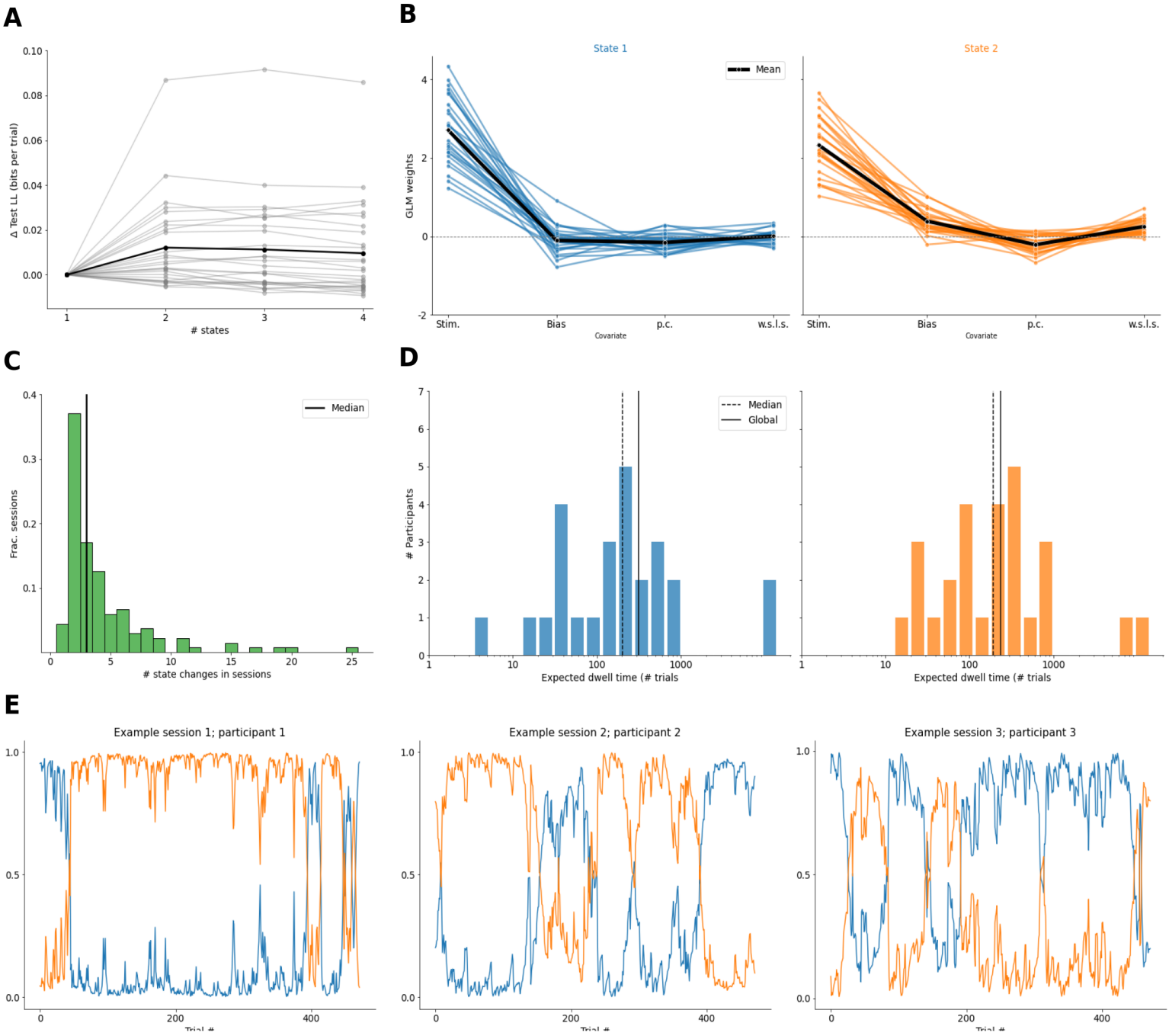
Linear mixed-effects analyses further revealed differences in RTs, accuracy rates and the probability of being engaged throughout the experiment.

A linear mixed-effects model predicting RTs from inferred states showed that responses were faster in state 1 compared to state 2, even though the effect was small ($\beta = 0.015$, $SE = 0.002$, $z = 6.35$, $p < .001$, 95% CI [0.011, 0.020], Cohen's $d = 0.06$). Figure 8 illustrates the mean RTs for all levels of sensory evidence in each state.

Accuracy rates did not differ significantly between states ($\beta = -0.033$, $SE = 0.020$, $z = -1.61$, $p = 0.108$, 95% CI [-0.073, 0.007], Cohen's $d = 0.02$), but tended to be higher in state 1 compared to state 2 (76.0 vs 73.7%).

Finally, a generalized linear mixed-effects model predicting the probability of being in the most engaged state showed a gradual decline over trials ($\beta = -0.00060$, $SE = 0.00009$, $z = -6.40$, $p < .001$, 95% CI [-0.00042, -0.00079]). While the per-trial decrease of 0.02% appears minimal, this accumulates to a total decrease of 7.46% over 500 trials, corresponding to a cumulative effect size of Cohen's $d = 0.17$.

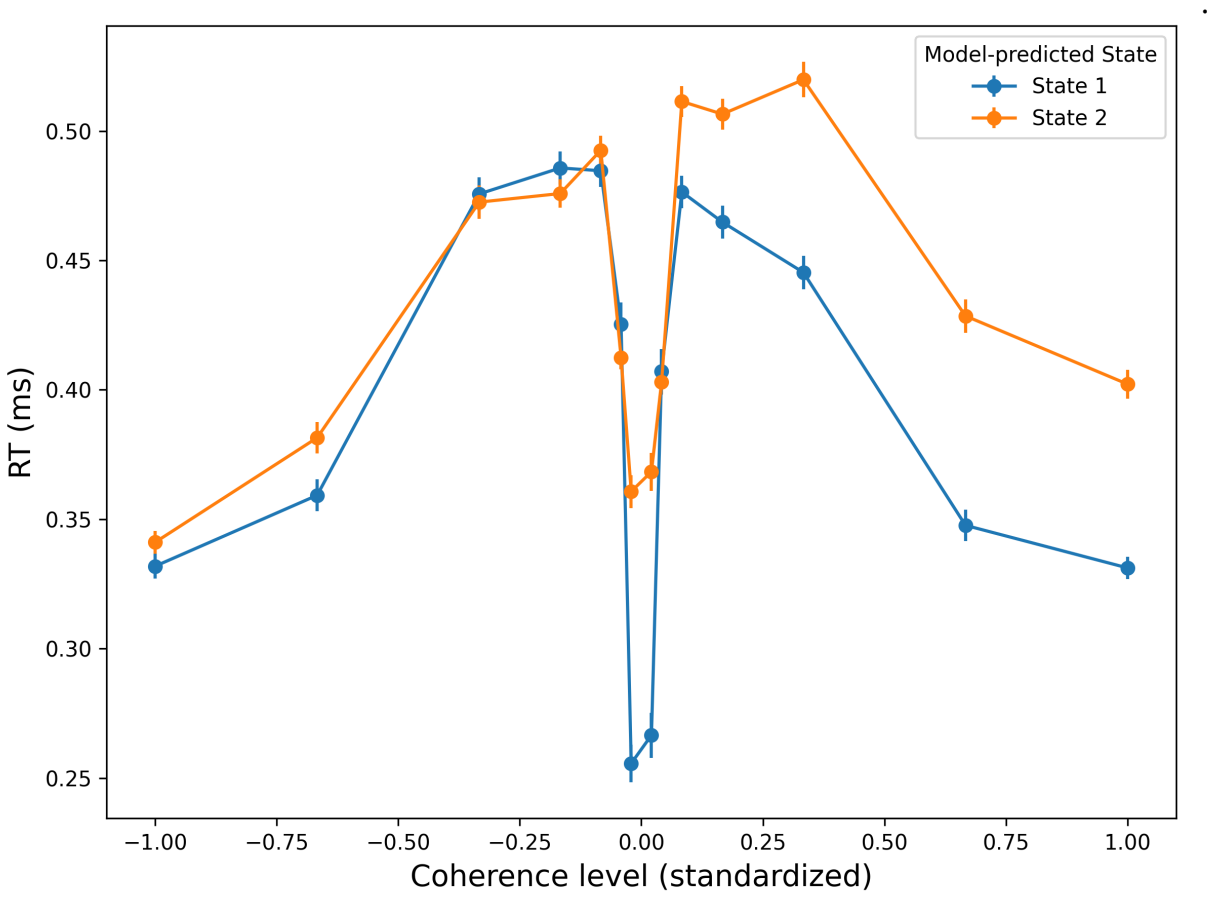
Figure 7
Replication Results Ashwood et al. (2022)



Note. Results from the GLM-HMM applied to the full Urai dataset are shown across multiple measures. (a) Change in test set log-likelihood relative to a GLM for each of the 27 participants; black indicates the mean across participants. (b) Retrieved weights for a two-state GLM-HMM for each participant; black indicates the mean across participants. (c) Number of state changes per session obtained from posterior state probabilities (such as those shown in e); median session length is 500 trials, and black indicates the median across sessions. (d) Expected dwell times for each participant in each state, obtained from the inferred transition matrices; black dashed line indicates the median across participants, and black solid line indicates the global fit. (e) Posterior state probabilities for three example sessions corresponding to three different participants.

Figure 8

Mean Reaction Time by Normalized Coherence Level Across States



*Note***.** Mean reaction times are shown for each normalized coherence level averaged across subjects in both states. Error bars indicate standard deviation across participants.

Discussion

The present study served as an exploratory step in a broader research program aimed at establishing whether fluctuations in task engagement can be inferred from human choice behaviour using a GLM-HMM. The goals of this research were to 1) optimize the statistical pipeline for subsequent large-scale analysis, and 2) asses the robustness of the GLM-HMM framework by replicating the results of Ashwood et al. (2022). Overall, our findings are consistent with the original analysis, showing similar profiles in terms of latent state structure, behavioural strategies used in each state and transition dynamics between them.

Cross-validated model comparisons revealed highly similar results: in both analyses, a two-state model outperformed alternative explanations, with an improvement of 0.013 bits per trial in the original analysis and 0.012 bits in the present study. These results indicate strong identifiability of the latent state structure and support the robustness of the GLM-HMM framework.

In terms of engagement strategies, the most notable finding was that subjects relied more on sensory evidence in state 1 compared to state 2. This suggests that state 1 reflects a more engaged mode of task performance, whereas state 2 represents a slightly less engaged strategy. Importantly, however, the difference between states was modest. State 2 cannot be described as fully disengaged, but rather reflects a weaker reliance on task-relevant information. The same argumentation applies to the use of the WSLS strategy, which was slightly more pronounced in state 2. While the original analysis does not report these effects, the overall pattern of state-specific GLM weights per subject looked highly similar. Ashwood et al. (2022) did not formally test for state-specific weight differences, so these subtle effects may have existed in their data but went unreported.

With respect to response bias, we replicated the qualitative pattern observed by Ashwood et al. (2022), showing a difference in the direction of bias between states rather than its magnitude. At the level of state-switching dynamics, our findings again closely mirrored the original findings. The number of state switches per session, dwell times, and posterior state probabilities were highly similar across studies, providing strong evidence that the GLM-HMM framework recovers reproducible state dynamics, even when different computational implementations are used.

Taken together, the results reported in Ashwood et al. (2022) were successfully replicated in this study, with only minor differences between analyses. These small discrepancies are not unexpected, given the differences in the analysis pipeline between the studies, particularly with respect to model priors. Crucially, the fact that nearly identical results emerged despite these methodological differences further strengthens confidence in the GLM framework as reliable tool to infer latent states from behavioural data, as it highlights the robustness of the retrieved states and associated parameters.

Engagement States in Perceptual Decision-Making

The central question motivating this study was whether periods of disengagement can be detected in psychological experiments, particularly in paradigms that are not explicitly designed to elicit them. Together, our findings provide preliminary support that human choice behaviour reflects multiple engagement states during perceptual decision-making. Specifically, we identified two states that differed in both reliance on sensory evidence and use of WSLS heuristics.
This converges with a broader literature documenting multiple attentional states underlying performance in sustained attention tasks (Estermam et al., 2012; Cai et al., 2021; Gaillard et al., 2021) and perceptual decision-making in mice (Ahswood et al., 2022; Hulsey et al., 2024, Tlaie et al., 2025). Our findings suggest that disengagement is not restricted to CPTs but also arises in experimental contexts more typical of mainstream cognitive psychology.

A critical question for any latent-state model is whether the inferred states reflect meaningful aspects of behaviour or merely serve as computational abstractions. In line with Zhang and Kool (2025), we conducted validation analyses to assess the behavioural relevance of the recovered states. The results provide tentative support for the former interpretation. Reaction times were, on average, 15 ms longer in state 2 than in state 1, consistent with the interpretation of state 2 as less engaged. While Zhang and Kool reported faster responses in their disengaged state, such differences are likely attributable to task-specific demands (i.e., the task being set up to induce repetitive responding); indeed, across the literature, both faster and slower reaction times have been associated with disengagement (Unsworth et al., 2015; Isbell et al., 2018). Accuracy was numerically lower in state 2, although the difference did not reach significance, which nevertheless offers a useful sanity check. Finally, the probability of being in the engaged state decreased significantly across the course of the experiment, a robust finding that aligns with Zhang and Kool and a broad literature demonstrating declining engagement with time-on-task (Zanesco et al., 2024; Zhang & Kool, 2025). Together, these validation analyses confirm that the inferred states capture psychologically meaningful dimensions of behaviour rather than arbitrary statistical groupings.

Strengths and Limitations

A key strength of the present study is the use of an independent pilot dataset to explore and optimize the GLM-HMM pipeline. This approach ensured model convergence, stability, and well-tuned prior settings without influencing the main analysis. By successfully replicating the findings of Ashwood et al. (2022), the study demonstrates the robustness and reproducibility of the framework. Importantly, we extended the original analysis by examining state-specific GLM weights and conducting additional validation analyses, providing a more detailed characterization of

individual engagement profiles. This work establishes a solid foundation for applying the GLM-HMM to a broader range of data and research contexts in future research.

Nevertheless, the present findings should be interpreted with caution. The observed effects were modest, and our exploratory approach to setting model priors, while conducted on independent data, was unstructured. A more systematic procedure, such as cross-validation on held-out data (see Tlaie et al., 2025), would further reduce researcher degrees of freedom, ensure optimal model fit that balances the bias–variance trade-off, and facilitate comparability across datasets. Moreover, the study was underpowered: a sensitivity power analysis indicated a minimal detectable effect size of $d = 0.56$, meaning only medium-to-large effects could be reliably detected. Consequently, these results should be regarded as preliminary evidence rather than definitive proof of multiple engagement states in human perceptual decision-making.

Another methodological consideration further qualifies our conclusions. While we aimed to closely replicate the analysis pipeline of Ashwood et al. (2022), some deviations were unavoidable. Most notably, we employed the Dynamax package rather than the original SSM toolkit, which did not support all functions that were used in the original analysis. Nevertheless, reproducing a similar pattern of results despite these differences highlights the robustness of the GLM-HMM framework. Finally, the pilot dataset may not fully represent the broader data pool. The uncued and randomly interleaved manipulation of stimulus volatility likely introduced additional external variability that was not explicitly modelled.

Implications and Future Directions

The present study provides preliminary evidence that human engagement fluctuates during perceptual decision-making, and that these fluctuations can be inferred directly from choice behaviour using the GLM-HMM framework. The implications of these results are substantial as unaccounted fluctuations in attentional state can: a) confound experimental relationships, b) bias parameter estimates and c) violate key assumptions of widely-used statistical models. Theoretically, our results highlight the need to account for temporal variability in cognitive engagement (Urai, 2025), challenging assumptions of stable performance in many standard paradigms. Methodologically, the GLM-HMM framework addresses the limitations of current methodology and provides an elegant solution by identifying latent engagement states directly from behavioural data, thereby allowing researchers to control for these dynamics in their analyses. Practically, this framework improves accessibility of attentional dynamics for researchers and clinicians, thereby enabling large-scale scientific investigations, re-analysis of a large body of existing behavioural datasets, and potential applications in cognitive assessment and interventions for populations with attentional difficulties.

Nevertheless, future studies are needed to confirm the preliminary results found here. These studies should use larger research samples and a more systematic approach to prior specification, to enable a more rigorous assessment of the GLM-HMM's capacity to capture fluctuations in engagement. Future work within this research program will directly address these limitations: a research sample comprising 162 participants, drawn from six experiments spanning multiple perceptual tasks, has already been selected in the current work for this next phase.

If future research confirms these preliminary findings, subsequent work should also relate the inferred engagement states to other physiological and behavioural indices of engagement, such as arousal, as demonstrated in mice by Hulsey et al. (2024). Moreover, methodological extensions of the GLM-HMM should be explorxed. For instance, following Mohammadi et al. (2025), incorporating time-varying transition probabilities could capture dynamic fluctuations in engagement more accurately, addressing the limitation of fixed transition probabilities in standard GLM-HMM implementations, which is likely invalid (Zanesco et al.., 2024).

Conclusion

The present study provides preliminary evidence that the GLM-HMM framework can capture latent fluctuations in task engagement during perceptual decision-making. We replicated key findings from Ashwood et al. (2022), extended their work by identifying subtle state-level differences in reliance on stimulus weight and WSLS strategies, and validated the behavioural relevance of the recovered states. However, given the modest effect sizes and exploratory analytic choices, these results should be regarded as proof of concept rather than definitive evidence. Despite this, the close correspondence with previous findings underscores the robustness and promise of the GLM-HMM approach. By establishing and testing the analysis pipeline, this study lays the groundwork for the next phase of the research program, which will apply the refined framework to a larger and more diverse dataset to confirm and extend the present findings.

References

Ashwood, Z. C., Roy, N. A., Stone, I. R., International Brain Laboratory, Urai, A. E., Churchland, A. K., … Pillow, J. W. (2022). Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience, 25*(2), 201–212. https://doi.org/10.1038/s41593-021-01007-z

Bastian, M., & Sackur, J. (2013). Mind wandering at the fingertips: Automatic parsing of subjective states based on response time variability. *Frontiers in Psychology, 4*, 573. https://doi.org/10.3389/fpsyg.2013.00573

Cai, W., Warren, S. L., Duberg, K., Pennington, B., Hinshaw, S. P., & Menon, V. (2021). Latent brain state dynamics distinguish behavioral variability, impaired decision-making, and inattention. *Molecular Psychiatry, 26*(9), 4944–4957. https://doi.org/10.1038/s41380-021-01022-3

Carandini, M., & Churchland, A. K. (2013). Probing perceptual decisions in rodents. *Nature Neuroscience, 16*(7), 824-831. https://doi.org/10.1038/nn.3410

Chakravarty, S., Delgado-Sallent, C., Kane, G. A., Xia, H., Do, Q. H., Senne, R. A., & Scott, B. B. (2024). A cross-species framework for investigating perceptual evidence accumulation (Preprint). *bioRxiv*. https://doi.org/10.1101/2024.04.17.589945

Cheyne, J. A., Solman, G. J., Carriere, J. S., & Smilek, D. (2009). Anatomy of an error: A bidirectional state model of task engagement/disengagement and attention-related errors. *Cognition, 111*(1), 98–113. https://doi.org/10.1016/j.cognition.2008.12.009

Desender, K., Donner, T. H., & Verguts, T. (2020). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition, 207*, Article 104522. https://doi.org/10.1016/j.cognition.2020.104522

Esterman, M., Noonan, S., & Rosenberg, M. (2012). In the zone or zoning out? Behavioral and neural evidence for distinct attentional states. *Journal of Vision, 12*(9), 383–383. https://doi.org/10.1167/12.9.383

Esterman, M., & Rothlein, D. (2019). Models of sustained attention. *Current Opinion in Psychology, 29*, 174–180. https://doi.org/10.1016/j.copsyc.2019.03.005

Fortenbaugh, F. C., DeGutis, J., & Esterman, M. (2017). Recent theoretical, neural, and clinical advances in sustained attention research. *Annals of the New York Academy of Sciences, 1396*(1), 70–91. https://doi.org/10.1111/nyas.13318

Gaillard, C., Sousa, C. D., Amengual, J., Loriette, C., Ziane, C., Hassen, S. B. H., … Hamed, S. B. (2021). Attentional brain rhythms during prolonged cognitive activity (Preprint). *bioRxiv*. https://doi.org/10.1101/2021.05.26.445730

Giambra, L. M. (1995). A laboratory method for investigating influences on switching attention to task-unrelated imagery and thought. *Consciousness and Cognition, 4*, 1–21. https://doi.org/10.1006/ccog.1995.1001

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), 438–468. https://doi.org/10.48550/arXiv.1906.10838

Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review, 130*(6), 1521–1543. https://doi.org/10.1037/rev0000411

Hulsey, D., Zumwalt, K., Mazzucato, L., McCormick, D. A., & Jaramillo, S. (2024). Decision-making dynamics are predicted by arousal and uninstructed movements. *Cell Reports, 43*(2), 113709. https://doi.org/10.1016/j.celrep.2024.113709

Isbell, E., Calkins, S. D., Swingler, M. M., & Leerkes, E. M. (2018). Attentional fluctuations in preschoolers: Direct and indirect relations with task accuracy, academic readiness, and school performance. *Journal of Experimental Child Psychology, 167*, 388–403. https://doi.org/10.1016/j.jecp.2017.11.013

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science, 18*(7), 614–621. https://doi.org/10.1111/j.1467-9280.2007.01948.x

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General, 145*(8), 1017–1048. https://doi.org/10.1037/xge0000248

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science, 330*(6006), 932–932. https://doi.org/10.1126/science.1192439

Linderman, S., Antin, B., Zoltowski, D., & Glaser, J. (2020, October 15). *SSM: Bayesian learning and inference for state space models* (Version 0.0.1) [Computer software]. GitHub. https://github.com/lindermanlab/ssm

Linderman, S. W., Chang, P., Harper-Donnelly, G., Kara, A., Li, X., Duran-Martin, G., & Murphy, K. (2025). Dynamax: A Python package for probabilistic state space modeling with JAX. *Journal of Open Source Software, 10*(108), 7069. https://doi.org/10.21105/joss.07069

Manly, T., Robertson, I. H., Galloway, M., & Hawkins, K. (1999). The absent mind: Further investigations of sustained attention to response. *Neuropsychologia, 37*(6), 661–670. https://doi.org/10.1016/S0028-3932(98)00127-4

Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision, 12*(6), 25-25. https://doi.org/10.1167/12.6.25

Rahnev, D., Kok, P., Munneke, M., Bahdo, L., de Lange, F. P., & Lau, H. (2013). Continuous theta burst transcranial magnetic stimulation reduces resting state connectivity between visual areas. *Journal of Neurophysiology, 110*(8), 1811–1821. https://doi.org/10.1152/jn.00209.2013

Rahnev, D., Desender, K., Lee, A. L., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., … Zylberberg, A. (2020). The confidence database. *Nature Human Behaviour, 4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rausch, M., & Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology, 7*, 591. https://doi.org/10.3389/fpsyg.2016.00591

Rosenberg, M. D., Noonan, S., DeGutis, J., & Esterman, M. (2013). Sustaining visual attention in the face of distraction: A novel gradual-onset continuous performance task. *Attention, Perception, & Psychophysics, 75*, 426–439. https://doi.org/10.3758/s13414-012-0413-x

Roy, N. A., Bak, J. H., Akrami, A., Brody, C. D., & Pillow, J. W. (2021). Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron, 109*(4), 597–610. https://doi.org/10.1016/j.neuron.2020.12.004

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences, 15*(7), 319–326. https://doi.org/10.1016/j.tics.2011.05.006

Schubert, A. L., Frischkorn, G. T., & Rummel, J. (2020). The validity of the online thought-probing procedure of mind wandering is not threatened by variations of probe rate and probe framing. *Psychological Research, 84*(7), 1846–1856. https://doi.org/10.1007/s00426-019-01194-2

Smallwood, J. M., Davies, J. B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R., et al (2004). Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and Cognition, 13*, 657–690. https://doi.org/10.1016/j.concog.2004.06.003

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin, 132*(6), 946–958. https://doi.org/10.1037/0033-2909.132.6.946

Thomson, D. R., Besner, D., & Smilek, D. (2015). A resource-control account of sustained attention: Evidence from mind-wandering and vigilance paradigms. *Perspectives on Psychological Science, 10*(1), 82–96. https://doi.org/10.1177/1745691614556681

Tlaie, A., Abd El Hay, M. Y., Mert, B., Taylor, R., Ferracci, P. A., Shapcott, K., … Schölvinck, M. L. (2025). Inferring internal states across mice and monkeys using facial features. *Nature Communications, 16*(1), 5168. https://doi.org/10.1038/s41467-025-60296-1

Unsworth, N., Redick, T. S., Lakey, C. E., & Young, D. L. (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: An individual differences investigation. *Intelligence, 38*(1), 111–122. https://doi.org/10.1016/j.intell.2009.08.002

Unsworth, N. (2015). Consistency of attentional control as an important cognitive trait: A latent variable analysis. *Intelligence, 49*, 110–128. https://doi.org/10.1016/j.intell.2015.01.005

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications, 8*(1), 14637. https://doi.org/10.1038/ncomms14637

Urai, A. E. (2025). Structure uncovered: Understanding temporal variability in perceptual decision-making. *Trends in Cognitive Sciences*. Advance online publication. *https://doi.org/10.1016/j.tics.2025.06.003*

Vinski, M. T., & Watter, S. (2012). Priming honesty reduces subjective bias in self-report measures of mind wandering. *Consciousness and Cognition, 21*(1), 451–455. https://doi.org/10.1016/j.concog.2011.11.001

Zanesco, A. P., Denkova, E., & Jha, A. P. (2024). Mind-wandering increases in frequency over time during task performance: An individual-participant meta-analytic review. *Psychological*

Bulletin, 151(2), 217–239. https://doi.org/10.1037/bul0000424

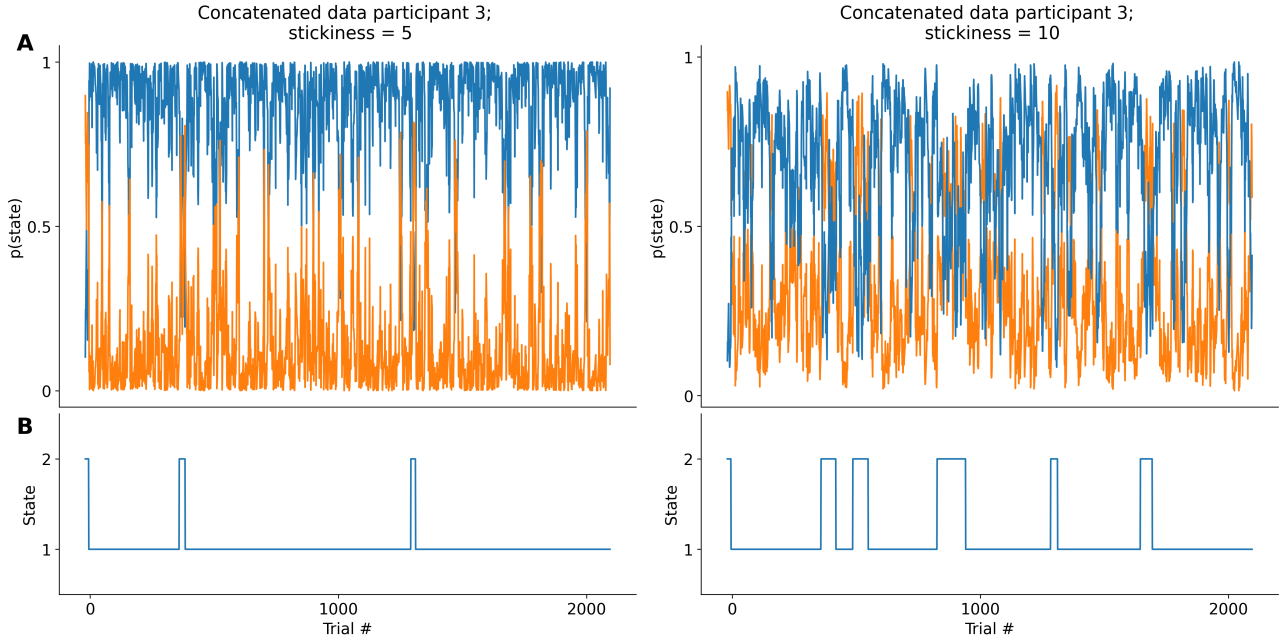
Zhang, C. K., & Kool, W. (2025). Inferring mind wandering from perceptual decision making (Preprint). *PsyArXiv*. https://doi.org/10.31234/osf.io/mxtbh

Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *Elife, 5*, e17688. https://doi.org/10.7554/eLife.17688

Appendix

Supplementary Figure 1

Comparison of inferred state dynamics between stickiness values of 5 and 10



Note. State sequences were inferred from model fits on concatenated data from subject 3. Stickiness values of 5 and 10 are compared to illustrate the effect of parameter choice. **(a)** P*osterior state probabilities (**b**) inferred state sequences (b).*