Bayesian Semiparametric Longitudinal Inverse-Probit Mixed Models for Category Learning

Minerva Mukhopadhyay[1](minervam@iitk.ac.in)
Jacie R. McHaney[2] (j.mchaney@pitt.edu)
Bharath Chandrasekaran[2](b.chandra@pitt.edu)
Abhra Sarkar[3] (abhra.sarkar@utexas.edu)

[1]Department of of Mathematics and Statistics,
Indian Institute of Technology,
Kanpur, UP 208016, India

[2]Department of Communication Science and Disorders,
University of Pittsburgh,
4028 Forbes Tower, Pittsburgh, PA 15260, USA

[3]Department of Statistics and Data Sciences,
University of Texas at Austin,
105 East 24th Street D9800, Austin, TX 78712, USA

Abstract

Understanding how the adult human brain learns novel categories is an important problem in neuroscience. Drift-diffusion models are popular in such contexts for their ability to mimic the underlying neural mechanisms. One such model for gradual longitudinal learning was recently developed in Paulon *et al.* (2021). Fitting conventional drift-diffusion models, however, requires data on both category responses and associated response times. In practice, category response accuracies are often the only reliable measure recorded by behavioral scientists to describe human learning. To our knowledge, however, drift-diffusion models for such scenarios have never been considered in the literature before. To address this gap, in this article, we build carefully on Paulon *et al.* (2021), but now with latent response times integrated out, to derive a novel biologically interpretable class of 'inverse-probit' categorical probability models for observed categories alone. However, this new marginal model presents significant identifiability and inferential challenges not encountered originally for the joint model in Paulon *et al.* (2021). We address these new challenges using a novel projection-based approach with a symmetry-preserving identifiability constraint that allows us to work with conjugate priors in an unconstrained space. We adapt the model for group and individual-level inference in longitudinal settings. Building again on the model's latent variable representation, we design an efficient Markov chain Monte Carlo algorithm for posterior computation. We evaluate the empirical performance of the method through simulation experiments. The practical efficacy of the method is illustrated in applications to longitudinal tone learning studies.

Key Words: Category learning, B-splines, Drift-diffusion models, Functional models, Inverse Gaussian distributions, Longitudinal mixed models, Speech learning

Short/Running Title: Longitudinal Inverse-Probit Mixed Models

Corresponding Author: Abhra Sarkar (abhra.sarkar@utexas.edu)

1 Introduction

Scientific background.

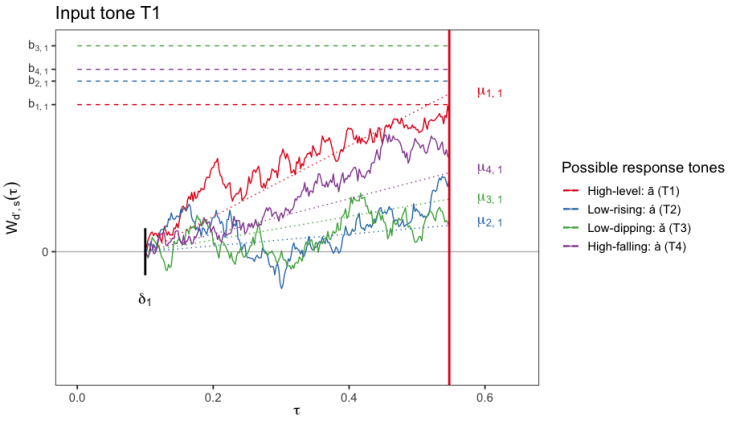
Scientific background. Categorization decisions are important in almost all aspects of our lives - whether it is a friend or a foe, edible or non-edible, the word /bat/ or /hat/, etc. The underlying cognitive dynamics are being actively studied through extensive ongoing research (Smith and Ratcliff, 2004; Heekeren *et al.*, 2004; Gold and Shadlen, 2007; Schall, 2001; Purcell, 2013; Glimcher and Fehr, 2013).

In typical multi-category decision tasks, the brain accumulates sensory evidence in order to make a categorical decision. This accumulation process is reflected in the increasing firing rates at local neural populations associated with different decisions. A decision is taken when neural activity in one of these populations reaches a particular threshold level. The decision category that is finally chosen is the one whose decision threshold is crossed first (Gold and Shadlen, 2007; Brody and Hanks, 2016). Changes in evidence accumulation rates and decision thresholds can be induced by differences in task difficulty and/or cognitive function (Cavanagh *et al.*, 2011; Ding and Gold, 2013). Decision-making is also regulated by demands on both the speed and accuracy of the task (Bogacz *et al.*, 2010; Milosavljevic *et al.*, 2010).

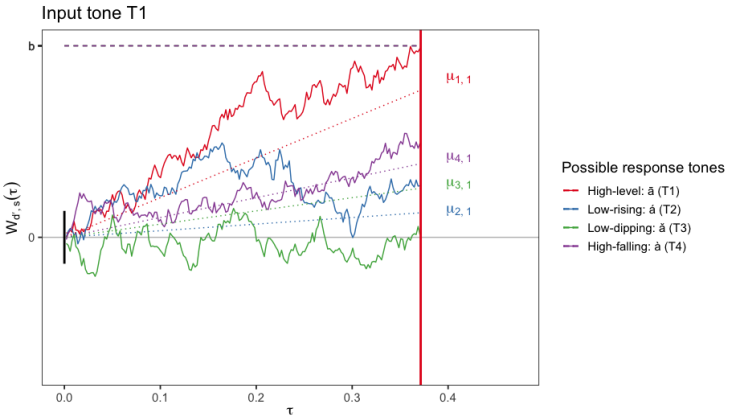
Understanding the brain activity patterns for different decision alternatives is a key scientific interest in modeling brain mechanisms underlying decision-making. Statistical approaches with biologically interpretable parameters that further allow probabilistic clustering of the parameters (Lau and Green, 2007; Wade, 2023) associated with different competing choices can facilitate such inference, the parameters clustering together indicating similar behavior and difficulty levels.

Drift-diffusion models. A biologically interpretable joint model for decision response accuracies and associated response times is obtained by imitating the underlying evidence accumulation mechanisms using latent drift-diffusion processes racing toward their respective boundaries, the process reaching its boundary first producing the final observed decision and the time taken to reach this boundary giving the associated response time (Figure 1, Panel (a)) (Usher and McClelland, 2001).

The literature on drift-diffusion processes for decision-making is rather vast but is mostly focused on simple binary decision scenarios with a single latent diffusion process with two boundaries, one for each of the two decision alternatives (Ratcliff, 1978; Ratcliff *et al.*, 2016; Smith and Vickers, 1988; Ratcliff and Rouder, 1998; Ratcliff and McKoon, 2008). Multi-category drift-diffusion models with multiple latent processes are mathematically more easily tractable (Usher and McClelland, 2001; Brown and Heathcote, 2008; Leite and Ratcliff, 2010; Dufau *et al.*, 2012; Kim *et al.*, 2017) but the literature is sparse and focused only on simple static designs.



(a) Drift-diffusion processes for tone learning when data on both response categories and response times are available.



(b) Drift-diffusion processes for tone learning proposed in this article – with additional identifiability restrictions imposed to accommodate scenarios when only data on response categories are available.

Figure 1: Drift-diffusion model for tone learning. The tones {T1, T2, T3, T4} represent the different categories; s denotes an input category, d' the different possible response categories, and d the final response category. Here we are illustrating a single trial with input tone T1 ($s = 1$) that was eventually correctly identified ($d = 1$). Panel (a) shows a process whose parameters can be inferred from data on both response categories and response times. Here, after an initial δ_s amount of time required to encode an input category s (here T1), the evidence in favor of different possible response categories d' accumulates according to latent Wiener diffusion processes $W_{d',s}(\tau)$ (red, blue, green, and purple) with drifts $\mu_{d',s}$. The decision d (here T1) is eventually taken if the underlying process (here the red one) is the first to reach its decision boundary $b_{d,s}$. Panel (b) shows a process with additional identifiability restrictions (for all d' and s, $\delta_s = 0$, $b_{d',s} = b$ fixed, and $\sum_{d'=1}^{d_0} \mu_{d',s} = d_0$) considered in this article which can be inferred from data on response categories alone.

Learning to make categorization decisions is, however, a dynamic process, driven by perceptual adjustments in our brain and behavior over time. Category learning

is thus often studied in longitudinal experiments. To address the need for sophisticated statistical methods for such settings, Paulon *et al.* (2021) developed an inverse Gaussian distribution-based multi-category longitudinal drift-diffusion mixed model.

Data requirements and related challenges. Crucially, measurements on both the final decision categories and the associated response times are needed to estimate the drift and the boundary parameters from conventional drift-diffusion models, including the work by Paulon *et al.* (2021). Unfortunately, however, researchers often only record the participants' decision responses as their go-to measure of categorization performance, ignoring the response times (Chandrasekaran *et al.*, 2014; Filoteo *et al.*, 2010). Additionally, eliciting accurate response times can be methodologically challenging, e.g., in the case of experiments conducted online, especially during the Covid-19 pandemic (Roark *et al.*, 2021), or when the response times from participants/patients are unreliable due to motor deficits (Ashby *et al.*, 2003). Participants may also be asked to delay the reporting of their decisions so that delayed physiological responses that relate to decision-making can be accurately measured (McHaney *et al.*, 2021). In such cases, the reported response times may not accurately relate to the actual decision times and hence cannot be used in the analysis. As a result, conventional drift-diffusion analysis that requires data on both response accuracies and response times, such as Paulon *et al.* (2021), cannot be used in such scenarios.

The research question. The main research question addressed in this article is to see if a new class of drift-diffusion models can be designed which will allow the biologically interpretable drift-diffusion process parameters to be meaningfully recovered from data on input-output category combinations alone.

The inverse-probit model. Categorical probability models that build on latent drift-diffusion processes can be useful in providing biologically interpretable inference in data sets comprising input-output categories but no response times. To our knowledge, however, the problem has never been considered in the literature before. We aim to address this remarkable gap in this article.

By integrating out the latent response times from the joint inverse Gaussian drift-diffusion model for response categories and associated response times in Paulon *et al.* (2021), we can arrive at a natural albeit overparametrized model for the response categories. We refer to this as the 'inverse-probit' categorical probability model. This inverse-probit model serves as the starting point for the methodology presented in this article but, as we describe below, it also comes with significant and unique statistical challenges not encountered in the original drift-diffusion model.

Statistical challenges. While scientifically desirable, unfortunately, it is also mathematically impossible to infer both the drifts and the boundaries in the inverse-probit model from data only on the decision accuracies. We must thus have to keep the values of either the drifts or the boundaries fixed and focus on inferring the other.

However, even when we fix either the drift or the decision boundaries, the problem of overparametrization persists. In the absence of response times, only the information on relative frequencies, that is empirical probabilities of taking a decision is available. As the total probability of observing any of the competing decisions is one, the identifiability problem remains for the chosen main parameters of interest, and appropriate remedial constraints need to be imposed.

Setting an arbitrarily chosen category as the reference provides a simple solution widely adopted in categorical probability models but comes with serious limitations, including breaking the symmetry of the problem, potentially making posterior inference sensitive to the specific choice of the reference category (Burgette and Nordheim, 2012; Johndrow *et al.*, 2013).

By breaking the symmetry of the problem, a reference category also additionally makes it difficult to infer the potential clustering of the model parameters, especially across different panels. To see this, consider a problem with d_0 categories, with a logistic model for the probabilities $p_{s,d'} = \text{logistic}(\beta_{s,d'})$, $s, d' \in \{1 : d_0\}$, of choosing the d'^{th} output category for the s^{th} input category. For each input category s, by setting the s^{th} output category as a reference, e.g., by fixing $\beta_{s,s} = 0$, one can then cluster the probabilities of incorrect decision choices, $p_{s,d'}$, $d' \neq s$. However, it is not clear how to compare the probabilities across different input categories (i.e., across the four panels in Figure 2), e.g., how to test the equality of $p_{1,1}$ and $p_{2,2}$.

Finally, while coming up with solutions for the aforementioned issues, we must also take into consideration the complex longitudinal design of the experiments generating the data. Whatever strategy we devise, it should be amenable to a longitudinal mixed model analysis that ideally allows us to (a) estimate the smoothly varying longitudinal trajectories of the parameters as the participants learn over time, (b) accommodate participant heterogeneity, and (c) compare the estimates at different time points within and between different input categories.

Our proposed approach. As a first step toward addressing the identifiability issues and related modeling challenges, we keep the boundaries fixed but leave the drift parameters unconstrained. The decision to focus on the drifts is informed by the existing literature on such models cited above where the drifts have almost always been allowed more flexibility. The analysis of Paulon *et al.* (2021) also showed that it is primarily the variations in the drift trajectories that explain learning while the boundaries remain relatively stable over time.

As a next step toward establishing identifiability, we apply a 'sum to a constant'

condition on the drifts so that symmetry is maintained in the constrained model.

Implementation of this restriction brings in significant challenges. One possibility is to design a prior on the constraint space, a challenging task in itself. Additionally, posterior computation for such priors would also be extremely complicated in drift-diffusion models. Instead, we conduct inference with an unconstrained prior on the drift parameters and project the samples drawn from the corresponding posterior to the constrained space through a minimal distance mapping.

To adapt this categorical probability model to a longitudinal mixed model setting, we then assume that the drift parameters comprise input-response-category-specific fixed effects and subject-specific random effects, modeling them flexibly by mixtures of locally supported B-spline bases (de Boor, 1978; Eilers and Marx, 1996) spanning the length of the longitudinal experiment. These effects are thus allowed to evolve flexibly as smooth functions of time (Ramsay and Silverman, 2007; Morris, 2015; Wang *et al.*, 2016) as the participants get more experience and training in the decision tasks.

We take a Bayesian route to estimation and inference. Carefully exploiting conditional prior-posterior conjugacy as well as our latent variable construction, we design an efficient Markov chain Monte Carlo (MCMC) based algorithm for approximating the posterior, where sampling the latent response times for each observed response category greatly simplifies the computations.

We evaluate the numerical performance of the proposed approach in extensive simulation studies. We then apply our method to the PTC1 data set described below. These applications illustrate the utility of our method in providing insights into how the drift parameters characterize the rates of accumulation of evidence in the brain evolve over time, differ between input-output category combinations, as well as between individuals.

Differences from previous works. This article differs in many fundamental ways from all existing works on drift-diffusion models, including Paulon *et al.* (2021), where response categories and response times were *both* observed and therefore the drift and boundary parameters could be modeled jointly with no identifiability issues. In contrast, the current work is motivated by scenarios where data on *only* response categories are available, leading us to the inverse-probit categorical probability model which, with its complex identifiability issues, brings in new unique challenges to performing statistical inference, confining us only to infer the drift parameters on a relative scale, achieved via a novel projection-based approach. The introduction and analysis of the inverse-probit model, addressing the significant new statistical challenges posed by it, ranging across (a) identifiability issues, (b) assessment of intra and inter-panel similarities, (c) extension to complex longitudinal mixed effects settings to accommodate the motivating applications, (d) computational implementation of these new models, etc. are the novel contributions of this article.

Outline of the article. Section 2 describes our motivating tone learning study. Sections 3 and 4 develop our longitudinal inverse-probit mixed model. Section 5 outlines our computational strategies. Section 6 presents the results of simulation experiments. Section 7 presents the results of the proposed method applied to our motivating PTC1 study. Section 8 concludes the main article with a discussion. Additional details, including Markov chain Monte Carlo (MCMC) based posterior inference algorithms, are deferred to the supplementary material.

2 The PTC1 Data Set

The PTC1 (pupillometry tone categorization experiment 1) data set is obtained from a Mandarin tone learning study conducted at the Department of Communication Science and Disorders, University of Pittsburgh (McHaney *et al.*, 2021). Mandarin Chinese is a tonal language, which means that pitch patterns at the syllable level differentiate word meanings. There are four linguistically relevant pitch patterns in Mandarin that make up the four Mandarin tones: high-flat (Tone 1), low-rising (Tone 2), low-dipping (Tone 3), and high-falling (Tone 4). For example, the syllable /ma/ can be pronounced using the four different pitch patterns of the four tones, which would result in four different word meanings. Adult native English speakers typically experience difficulty differentiating between the four Mandarin tones because pitch contrasts at the syllable level are not linguistically relevant to word meanings in English (Wang *et al.*, 1999, 2003). Thus, Mandarin tones are valid stimuli to examine how non-native speech sounds are acquired, which has implications for second language learning in adulthood. In PTC1, a group of native English-speaking younger adults learned to categorize monosyllabic Mandarin tones in a training task. During a single trial of training, an input tone was presented over headphones, and the participants were instructed to categorize the tone into one of the four tone categories via a button press on a keyboard. Corrective feedback in the form of "Correct" or "Wrong" was then provided on screen. A total of $n = 28$ participants completed the training task across $T = 6$ blocks of training, each block comprising $L = 40$ trials. Figure 2 shows the middle 30% quantiles of the proportion of times the response to an input tone was classified into different tone categories over blocks across different subjects, each for the four input tones.

Pupillometry measurements were also taken during each trial. It is commonly used as a metric of cognitive effort during listening because increases in pupil diameter are associated with greater usage of cognitive resources (Zekveld *et al.*, 2011; Peelle, 2018; Winn *et al.*, 2018; Robison and Unsworth, 2019; Parthasarathy *et al.*, 2020). One issue with pupillary responses however is that they unfold slowly over time. In view of that, unlike standard Mandarin tone training tasks, where the participants hear the input tone, press the keyboard response, and are provided feedback all within a few seconds

(Chandrasekaran *et al.*, 2016; Reetzke *et al.*, 2018; Llanos *et al.*, 2020; Smayda *et al.*, 2015), in the PTC1 experiment, there was an intentional four-second delay from the start of the input tone to the response prompt screen where participants made their category decision via button press. This four-second delay allows the pupil to dilate in response to hearing the tone and begin to return to baseline before the participant makes a motor response to the button press. During this four-second period, participants have likely already made conscious category decisions. As such, the response times that are recorded in the end are not meaningful measures of their actual decision times.

This presents a critical limitation for using these response times for further analysis. Conventional drift-diffusion analysis that requires data on response times, such as the one presented in Paulon *et al.* (2021), can no longer be directly applied here. The focus of this article is to see if the drift-diffusion parameters can still be meaningfully recovered from input-output tone categories alone in the PTC1 data.

We found drift-diffusion analysis in the absence of reliable data on response times challenging enough to merit its separate treatment presented here. Relating drift-diffusion parameters to measures of cognitive effort such as pupillometry is another challenging problem that we are pursuing separately elsewhere.

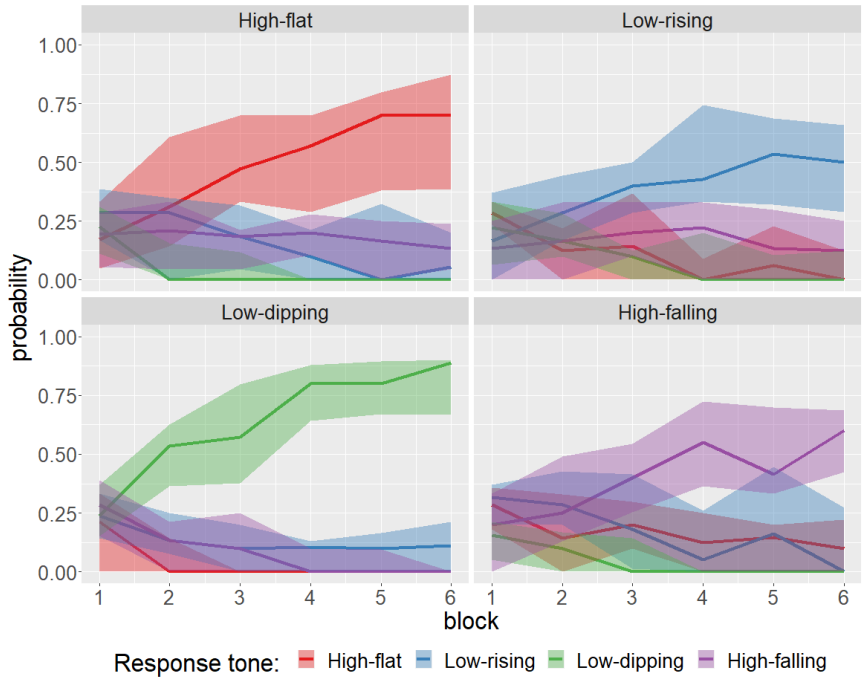


Figure 2: Description of PTC1 data: The proportion of times the response to an input tone was classified into different tone categories over blocks across different subjects, each for the four input tones (indicated in the panel headers). The thick line represents the median performance and the shaded region indicates the corresponding middle 30% quantiles across subjects.

3 Inverse-Probit Model

The starting point for the proposed inverse-probit categorical probability model follows straightforwardly by integrating out the (unobserved) response times from the joint model for response categories and associated response times developed in Paulon *et al.* (2021). The derivation of this original joint model illustrates its latent drift-diffusion process-based underpinnings (Figure 1, panel a). Later such construction will also be crucial in understanding the diffusion process-based foundations of the marginal categorical probability model modified with identifiability constraints proposed in this article (Figure 1, panel b). We therefore present the derivation from Paulon *et al.* (2021) ditto here which also keeps the main paper self-contained.

To begin with, a Wiener diffusion process $W(\tau)$ over domain $\tau \in (0, \infty)$ can be specified as $W(\tau) = \mu\tau + \sigma B(\tau)$, where $B(\tau)$ is the standard Brownian motion, μ is the drift rate, and σ is the diffusion coefficient (Cox and Miller, 1965; Ross *et al.*, 1996). The process has independent normally distributed increments, i.e., $\Delta W(\tau) = \{W(\tau + \Delta\tau) - W(\tau)\} \sim \text{Normal}(\mu\Delta\tau, \sigma^2\Delta\tau)$, independently from $W(\tau)$. The first passage time of crossing a threshold b, $\tau = \inf\{\tau' : W(0) = 0, W(\tau') \geq b\}$, is then distributed according to an inverse Gaussian distribution (Whitmore and Seshadri, 1987; Chhikara, 1988; Lu, 1995) with mean b/μ and variance $b\sigma^2/\mu^3$.

Given a perceptual stimulus s and a set of decision choices $d' \in \{1 : d_0\}$, the neurons in the brain accumulate evidence in favor of the different alternatives. Modeling this behavior using latent Wiener processes $W_{d',s}(\tau)$ with unit variances, assuming that a decision d is made when the decision threshold $b_{d,s}$ for the d^{th} option is crossed first, as illustrated in Figure 1, Panel (a), a probability model for the time τ_d to reach decision d is obtained as

$$f(\tau_d \mid \delta_s, \mu_{d,s}, b_{d,s}) = \frac{b_{d,s}}{\sqrt{2\pi}}(\tau_d - \delta_s)^{-3/2} \exp\left[-\frac{\{b_{d,s} - \mu_{d,s}(\tau_d - \delta_s)\}^2}{2(\tau_d - \delta_s)}\right], \tag{1}$$

where $\mu_{d,s}$ denotes the rate of accumulation of evidence, $b_{d,s}$ the decision boundaries, and δ_s an offset representing time not directly related to the underlying evidence accumulation processes (e.g., the time required to encode the s^{th} signal before evidence accumulation begins, etc.). We let $\boldsymbol{\theta}_{d',s} = (\delta_s, \mu_{d',s}, b_{d',s})^{\mathrm{T}}$.

Joint model for (d, τ): Since a decision d is reached at response time τ if the corresponding threshold is crossed first, that is when $\{\tau = \tau_d\} \cap_{d' \neq d} \{\tau_{d'} > \tau_d\}$, we have $d = \arg\min_{d' \in \{1:d_0\}} \tau_{d'}$. Assuming simultaneous accumulation of evidence for all decision categories, modeled by independent Wiener processes, and termination when the threshold for the observed decision category d is reached, the joint distribution of (d, τ) is thus given by

$$f(d, \tau \mid s, \boldsymbol{\theta}) = g(\tau \mid \boldsymbol{\theta}_{d,s}) \prod_{d' \neq d} \{1 - G(\tau \mid \boldsymbol{\theta}_{d',s})\}, \tag{2}$$

where, to distinguish from the generic notation f, we now use $g(\cdot \mid \boldsymbol{\theta})$ and $G(\cdot \mid \boldsymbol{\theta})$ to denote, respectively, the probability density function (pdf) and the cumulative distribution function (cdf) of an inverse Gaussian distribution, as defined in (1).

Marginal model for d: When the response times τ are unobserved, the probability of taking decision d given the stimulus s is thus obtained from (2) by integrating out the τ as

$$P(d \mid s, \boldsymbol{\theta}) = \int_{\delta_s}^{\infty} g(\tau \mid \boldsymbol{\theta}_{d,s}) \prod_{d' \neq d} \{1 - G(\tau \mid \boldsymbol{\theta}_{d',s})\} \, d\tau. \tag{3}$$

The construction of model (3) is similar to traditional multinomial probit/logit regression models except that the latent variables are now inverse Gaussian distributed as opposed to being normal or extreme-value distributed, and the observed category is associated with the minimum of the latent variables in contrast to being identified with the maximum of the latent variables. We thus refer to this model as a 'multinomial inverse-probit model'.

With data on both response categories d and response times τ available, the joint model (2) was used to construct the likelihood function in Paulon *et al.* (2021). In the absence of data on the response times τ, however, the inverse-probit model in (3) provides the basic building block for constructing the likelihood function for the observed response categories. As mentioned in the Abstract, discussed in the Introduction, and detailed in Section 3.1 below, the marginal inverse-probit model (3) for observed categories brings in many new identifiability issues and inference challenges not originally encountered for the joint model (2) developed in Paulon *et al.* (2021). Solving these new challenges for the marginal model (3) to infer the underlying drift-diffusion parameters $\boldsymbol{\theta}_{d',s}$, for all d', is the focus of this current article.

3.1 Identifiability Issues and Related Modeling Challenges

To begin with, we note that model (3) in itself cannot be identified from data on only the response categories. The offset parameters can easily be seen to not be identifiable since $P(\tau_d \leq \wedge_{d'} \tau_{d'}) = P\{(\tau_d - \delta) \leq \wedge_{d'} (\tau_{d'} - \delta)\}$ for any δ, where $\wedge_{d'} \tau_{d'}$ denotes the minimum of $\tau_{d'}, d' \in \{1 : d_0\}$. As is also well known in the literature, in categorical probability models, the location and scale of the latent continuous variables are not also separately identifiable. The following lemma establishes these points for the inverse-probit model.

Lemma 1. *The offset parameters δ_s are not identifiable in model (3). The drift and the boundary parameters, respectively $\mu_{d',s}$ and $b_{d',s}$, are also not separately identifiable in model (3).*

In the proof of Lemma 1 given in Appendix A, we have specifically shown that $P(d \mid s, \boldsymbol{\theta}) = P(d \mid s, \boldsymbol{\theta}^{\star})$, where the drift and boundary parameters in $\boldsymbol{\theta} =$

$\{(\mu_{d',s}, b_{d',s})\,; d' = 1, \ldots, d_0\}$ and $\boldsymbol{\theta}^\star = \left\{\left(\mu^\star_{d',s}, b^\star_{d',s}\right); d' = 1, \ldots, d_0\right\}$ satisfy $\mu^\star_{d',s} = c\mu_{d',s}$ and $b^\star_{d',s} = c^{-1}b_{d',s}$ for some constant $c > 0$. The result follows by noting that the transformation $\tau^\star_{d',s} = c^{-2}\tau_{d',s}$ does not change the ordering between the $\tau_{d',s}$'s and hence the probabilities of the resulting decisions $d = \arg\min_{d' \in \{1:d_0\}} \tau_{d'} = \arg\min_{d' \in \{1:d_0\}} \tau^\star_{d'}$ also remain the same. This has the simple implication that if the rate of accumulation of evidence is faster, then the same decision distribution is obtained if the corresponding boundaries are accordingly closer and conversely.

In fact, given the information on input and output categories alone, if d_0 denotes the number of possible decision categories, at most $d_0 - 1$ parameters are estimable. To see this, consider the probabilities $P(d' \mid s, \boldsymbol{\theta})$, $d' = 1, \ldots, d_0$, where $\boldsymbol{\theta}$ is the m-dimensional vector of parameters, possibly containing drift parameters and decision boundaries. Given the perceptual stimulus s as input, the probabilities satisfy $\sum_{d'=1}^{d_0} P(d' \mid s, \boldsymbol{\theta}) = 1$. Thus, the function $P_s(\boldsymbol{\theta}) = \{P(1 \mid s, \boldsymbol{\theta}), \cdots, P(d_0 \mid s, \boldsymbol{\theta})\}^{\mathrm{T}}$ lie on a $d_0 - 1$-dimensional simplex, $P_s(\boldsymbol{\theta}) : \boldsymbol{\theta} \to \Delta^{d_0 - 1}$, and by the model in (3) the mapping is continuous. Thus, it can be shown by the *Invariance of Domain* theorem (see, e.g., Deo, 2018) that if P_s is injective and continuous, then the domain of P_s must belong to $\mathbb{R}^m$, where $m \leq d_0 - 1$. Thus in order to ensure identifiablilty of $\{P_s(\boldsymbol{\theta}); \boldsymbol{\theta}\}$, we must parametrize the probability vector with at most $d_0 - 1$ parameters.

The existing literature on drift-diffusion models discussed in the Introduction has traditionally put more emphasis on modeling the drifts (as their reference in the literature as 'drift'-diffusion models suggests). Previous research on joint models for response tones and associated response times in Paulon *et al.* (2021) also suggest that the boundaries remain stable around a value of 2 and it is primarily the changes in the drift rates that explain longitudinal learning. In view of this, we keep the boundaries fixed at the constant $b = 2$ and treat the drifts to be the free parameters instead. In our simulations and real data applications, it is observed that the estimates of the drift parameters and the associated cluster configurations are not very sensitive to small-to-moderate deviations of b around 2. In our codes implementing our method, available as part of the supplementary materials, we allow the practitioner to choose a value of b as they see fit for their specific application. The latent drift-diffusion process based with these constraints, namely $\delta_s = 0$ and $b_{d',s} = b$ for all d', is shown in Figure 1, Panel (b).

While fixing $\delta_s = 0$ and $b_{d',s}$ to some known constant b reduces the size of the parameter space to d_0, to ensure identifiability, we still need at least one more constraint on the drift parameters $\boldsymbol{\mu}_{1:d_0,s}$. In categorical probability models, the identifiabily problem of the location parameter is usually addressed by setting one category as a reference and modeling the probabilities of the others (Albert and Chib, 1993; Chib and Greenberg, 1998; Borooah, 2002; Agresti, 2018). However, posterior predictions from Bayesian categorical probability models with asymmetric constraints may be sensitive to the choice of reference category (see Burgette and Nordheim,

2012; Johndrow *et al.*, 2013). Further, as also discussed in the Introduction, the goal of clustering the drift parameters $\boldsymbol{\mu}_{1:d_0,s}$ across s can not be accomplished by this apparently simple solution.

The problem can be addressed by imposing a symmetric constraint on $\boldsymbol{\mu}_{1:d_0,s}$ instead. A symmetric identifiability constraint has been previously proposed by Burgette *et al.* (2021) in the context of multinomial probit models, where they considered a sum-to-zero constraint on the latent utilities. To implement the constraint, they introduced a *faux base category* indicator parameter, which is assigned a discrete uniform prior and then learned via MCMC. Given this faux base category indicator, the other parameters are adjusted so that the sum-to-zero restriction is satisfied. However, the introduction of a base category, even if adaptively chosen, does not facilitate the clustering of $\boldsymbol{\mu}_{1:d_0,s}$ within and across the different input categories s.

3.2 Our Proposed Approach

In coming up with solutions for these challenges, we take into consideration the complex design of our motivating tone learning experiments, so that our approach is easily extendable to longitudinal mixed model settings, allowing us to (a) estimate the smoothly varying trajectories of the parameters as the participants learn over time, (b) accommodate the heterogeneity between the participants, and (c) compare between the estimates not just within but also crucially between the different panels.

Similar to the *sum to zero* constraint in the multinomial probit model of Burgette *et al.* (2021), we impose a symmetric *sum to a constant* constraint on the drift parameters $\boldsymbol{\mu}_{1:d_0,s}$ to identify our new class of inverse-probit models, although our implementation is quite different from theirs. To conduct inference, we start with an unconstrained prior, then sample from the corresponding unconstrained posterior, and finally project these samples to the constrained space through a minimal distance mapping. Similar ideas have previously been applied to satisfy natural constraints in other contexts. See, e.g., Dunson and Neelon (2003); Gunn and Dunson (2005).

This approach is significantly advantageous both from a modeling and a computational perspective. On one hand, the basic building blocks are relatively easily extended to complex longitudinal mixed model settings, on the other, posterior computation is facilitated as this allows the use of conjugate priors for the unconstrained parameters. Projection of the drift parameters onto the same space further makes them directly comparable, allowing clustering within and across the panels. The projected drifts can now be interpreted only on a relative scale but such compromises are not avoidable given the challenges we face.

3.2.1 Minimal Distance Mapping

As the drift parameters are positive, the *sum to a constant k* constraint leads to the constrained space $\mathcal{S}_k = \{\boldsymbol{\mu} : \mathbf{1}^T\boldsymbol{\mu} = k,\ \mu_j > 0,\ j = 1, \ldots, d_0\}$ on which $\boldsymbol{\mu}_{1:d_0,s}$ should be projected. The space $\mathcal{S}_k$ is semi-closed, and therefore, the projection of any point $\boldsymbol{\mu}$ onto $\mathcal{S}_k$ may not exist. As a simple one dimensional example, let $x = -1$ and $\mathcal{S} = (0, 1]$, then $\arg\min_{y\in\mathcal{S}} |y - x| = 0 \notin \mathcal{S}$. Further, from a practical perspective, a drift parameter infinitesimally close to zero makes the distribution of the associated response times very flat which is typically not observed in real data. Therefore, we choose a small $\varepsilon > 0$ and project $\boldsymbol{\mu}$ onto $\mathcal{S}_{\varepsilon,k} = \{\boldsymbol{\mu} : \mathbf{1}^T\boldsymbol{\mu} = k,\ \mu_j \geq \varepsilon,\ j = 1, \ldots, d_0\}$. We then define the projection of a point $\boldsymbol{\mu}$ onto $\mathcal{S}_{\varepsilon,k}$ through minimal distance mapping as

$$\boldsymbol{\mu}^\star = \text{Proj}_{\mathcal{S}_{\varepsilon,k}}(\boldsymbol{\mu}) := \{\text{argmin}_{\boldsymbol{\nu}} \|\boldsymbol{\mu} - \boldsymbol{\nu}\| : \boldsymbol{\nu} \in \mathcal{S}_{\varepsilon,k}\},$$

where $\|\cdot\|$ is the Euclidean norm. Note that for appropriate choices of (k, ε), $\mathcal{S}_{\varepsilon,k}$ is non-empty, closed and convex. Therefore, $\boldsymbol{\mu}^\star$ exists and is unique by the Hilbert projection theorem (Rudin, 1991). The solution to this projection problem comes from the following result from Beck (2017).

Lemma 2. *Let $\mathcal{S}_{\varepsilon,k}$ be as defined above, and $\mathcal{S}_\varepsilon = \{\boldsymbol{\mu} : \mu_j \geq \varepsilon,\ j = 1, \ldots, d_0\}$. Then, $\text{Proj}_{\mathcal{S}_{\varepsilon,k}}(\boldsymbol{\mu}) = \text{Proj}_{\mathcal{S}_\varepsilon}(\boldsymbol{\mu} - u^\star\mathbf{1})$, where $u^\star$ is a solution to the equation $\mathbf{1}^T\text{Proj}_{\mathcal{S}_\varepsilon}(\boldsymbol{\mu} - u^\star\mathbf{1}) = k$.*

Although the analytical form of the solution is not available, as is evident from the above result, the solution mainly relies on finding a root $u^\star$ of the non-increasing function $\phi(u^\star) = \mathbf{1}^T\text{Proj}_{\mathcal{S}_\varepsilon}(\boldsymbol{\mu} - u^\star\mathbf{1}) - k$. We apply an algorithm based on Duchi *et al.* (2008) to reach the solution. The algorithm is described in Appendix C.

3.2.2 Identifiability Restrictions

The projection approach solves the problem of identifiability and maps the probability vector corresponding to an input tone s to the constraint space of $\boldsymbol{\mu}_{1:d_0,s}$, $\mathcal{S}_{\varepsilon,k}$. The following theorem shows that the mapping from the constrained space of $\boldsymbol{\mu}_{1:d_0,s}$ to the probability vector $\mathbf{P}(\boldsymbol{\mu}_{1:d_0,s}) = \{p_1(\boldsymbol{\mu}_{1:d_0,s}), \ldots, p_{d_0}(\boldsymbol{\mu}_{1:d_0,s})\}^{\mathrm{T}}$ is injective. To keep the ideas simple, we consider the domain of the function to be $\mathcal{S}_{0,k}$ (i.e., $\varepsilon = 0$) instead of $\mathcal{S}_{\varepsilon,k}$ although a very similar proof would follow if $\mathcal{S}_{\varepsilon,k}$ were considered.

Theorem 1. *Let $p_d(\boldsymbol{\mu}_{1:d_0,s})$ be the probability of observing the output tone d given the input tone s and the drift parameters $\boldsymbol{\mu}_{1:d_0,s}$, as given in (3), for each $d = 1 : d_0$. Suppose $\boldsymbol{\mu}_{1:d_0,s}$ lies on the space $\mathcal{S}_{0,k}$. Then, the function from $\mathcal{S}_{0,k}$ to the space of probabilities $\{p_d(\boldsymbol{\mu}_{1:d_0,s}); d = 1 : d_0\}$ is injective.*

A proof is presented in Appendix B.

3.2.3 Conjugate Priors for the Unconstrained Drifts

From (3), given $\tau_1, \ldots, \tau_{d_0}$, such that $\tau_d \leq \min\{\tau_1, \ldots, \tau_{d_0}\}$, the posterior full conditional of $\boldsymbol{\mu}_{1:d_0,s}$ is proportional to $\pi^{(n)}_{\boldsymbol{\mu}} \propto \pi(\boldsymbol{\mu}_{1:d_0,s}) \times \prod_{d'=1}^{d_0} g(\tau_{d'} \mid \mu_{d',s})$, where $\pi(\cdot)$ is the prior of $\boldsymbol{\mu}_{1:d_0,s}$. Observe that $\prod_{d'=1}^{d_0} g(\tau_{d'} \mid \mu_{d',s})$ is Gaussian in $\boldsymbol{\mu}_{1:d_0,s}$. A Gaussian prior on $\boldsymbol{\mu}_{1:d_0,s}$ thus induces a conditional posterior for $\boldsymbol{\mu}_{1:d_0,s}$ that is also Gaussian and hence very easy to sample from. Importantly, these benefits also extend naturally to multivariate Gaussian priors for any parameter vector $\boldsymbol{\beta}_{d',s}$ that relates to $\mu_{d',s}$ linearly. This will be crucial in allowing us to extend the basic building block to longitudinal functional mixed model settings in Section 4 next, where we will be modeling time varying $\mu_{d',s}(t)$ as flexible mixtures of B-splines with associated coefficients $\boldsymbol{\beta}_{d',s}$.

3.2.4 Justification as a Proper Bayesian Procedure

Define the constrained conditional posterior distribution, $\tilde{\pi}^{(n)}_{\tilde{\boldsymbol{\mu}}}$, of the drift parameters $\boldsymbol{\mu}$ as

$$\tilde{\pi}^{(n)}_{\tilde{\boldsymbol{\mu}}}(B \mid \boldsymbol{\zeta}) = \pi^{(n)}_{\boldsymbol{\mu}}(\{\boldsymbol{\mu} : \text{Proj}(\boldsymbol{\mu}) \in B\} \mid \boldsymbol{\zeta}), \quad B \subseteq \mathcal{S}_{\varepsilon,k},$$

where $\pi^{(n)}_{\boldsymbol{\mu}}$ is the unconstrained conditional posterior of $\boldsymbol{\mu}_{1:d_0,s}$, given the other variables $\boldsymbol{\zeta}$. The analytic form of the constrained conditional posterior is not available.

Sen *et al.* (2018) established a proper Bayesian justification for the posterior projection approach by showing the existence of a prior $\tilde{\pi}(\boldsymbol{\mu}_{1:d_0,s})$ on the constrained space $\mathcal{S}_{\varepsilon,k}$ such that the resulting posterior is the same as the projected posterior $\tilde{\pi}^{(n)}_{\tilde{\boldsymbol{\mu}}}$. When $\mathcal{S}_{\varepsilon,k}$ is non-empty, closed, and convex, i.e., the projection operator is measurable, such a prior exists if the unconstrained posterior is absolutely continuous with respect to the unconstrained prior (Sen *et al.*, 2018, Corollary 1). As the unconstrained induced prior and posterior of the drift parameters are both Gaussian, this result holds in our case as well.

4 Extension to Longitudinal Mixed Models

In this section we adapt the inverse probit model discussed in Section 3 to complex longitudinal design of our motivating PTC1 data set described in the Introduction. Let $s_{i,\ell,t}$ denote the input tone for the i^{th} individual in the ℓ^{th} trial of block t. Likewise, let $d_{i,\ell,t}$ denote the output tone selected by the i^{th} individual in the ℓ^{th} trial of block t. Setting the offsets at zero, and boundary parameters to a fixed constant b, we now

have

$$P\{d_{i,\ell,t} = d \mid s_{i,\ell,t} = s, \boldsymbol{\mu}^{(i)}_{1:d_0,s}(t)\} = \int_0^\infty g\{\tau \mid \mu^{(i)}_{d,s}(t)\} \prod_{d' \neq d} \left[1 - G\{\tau \mid \mu^{(i)}_{d',s}(t)\}\right] d\tau, \quad (4)$$

$$\text{where} \quad g\{\tau \mid s_{i,\ell,t} = s, \mu^{(i)}_{d',s}(t) = \mu\} = \frac{b}{\sqrt{2\pi}\tau^{3/2}} \exp\left[-\frac{\{b - \mu\tau\}^2}{2\tau}\right].$$

The drift rates $\mu^{(i)}_{d',s}(t)$ now vary with the blocks t. In addition, we accommodate random effects by allowing $\mu^{(i)}_{d',s}(t)$ to also depend on the subject index i. We let $\mathbf{d} = \{d_{i,\ell,t}\}_{i,\ell,t}$, and d_0 be the number of possible decision categories (T1, T2, $\cdots$, Td_0). The likelihood function thus takes the form

$$L(\mathbf{d} \mid \mathbf{s}, \boldsymbol{\theta}) = \prod_{d=1}^{d_0} \prod_{s=1}^{d_0} \prod_{t=1}^{T} \prod_{i=1}^{n} \prod_{\ell=1}^{L} \left[P\{d_{i,\ell,t} \mid s_{i,\ell,t}, \boldsymbol{\mu}^{(i)}_{1:d_0,s}(t)\}\right]^{1\{d_{i,\ell,t}=d, s_{i,\ell,t}=s\}}.$$

We reiterate that in deriving the identifiability conditions and designing their implementation strategy in Section 3.2, we had to make sure that they would be applicable to the complex multi-subject longitudinal design of the PTC1 data set. Following those ideas, we model the time-varying mixed effects drift parameters $\mu^{(i)}_{d',s}(t)$ without any constraints first, then project them to the space satisfying the necessary identifying conditions.

For the unconstrained model, we follow the outline of Paulon *et al.* (2021) with necessary likelihood adjustments. The details are deferred to Section S.1 of the supplementary material. We present here a general outline.

We decompose $\mu^{(i)}_{d',s}(t) = f_{d',s}(t) + u^{(i)}_{d',s}(t)$ where $f_{d',s}(t)$ and $u^{(i)}_{d',s}(t)$ denote, respectively, fixed and random effects components, which are both modeled using flexible mixtures of B-spline bases. This allows us to cluster the fixed effects for different (d', s) combinations with similar shapes by clustering the corresponding B-spline coefficients.

Given posterior samples of $f_{d',s}(t)$ and $u^{(i)}_{d',s}(t)$, unconstrained samples of $\mu^{(i)}_{d',s}(t)$ are obtained. For every input tone s, these unconstrained $\boldsymbol{\mu}^{(i)}_{1:d_0,s}(t)$'s are then projected to the space $\mathcal{S}_{\varepsilon,k}$ following the method described in Section 3.2.1.

5 Posterior Inference

Posterior inference for our proposed inverse-probit mixed model is carried out using samples drawn from the posterior using MCMC algorithm. The algorithm carefully exploits the conditional independence relationships encoded in the model as well as the latent variable construction of the model.

Inference can be greatly simplified by sampling the passage times $\tau_{1:d_0}$ and then

conditioning on them. However, it is not possible to generate $\tau_{1:d_0}$ sequentially, e.g., by generating the passage time of the d-th decision choice τ_d independently, and that of the other decision choices from a truncated inverse-Gaussian distribution, left truncated at τ_d.[1]

We implement a simple accept-reject sampler instead which generates values from the joint distribution of $\tau_{1:d_0}$ and accepts the sample if $\tau_d \leq \tau_{1:d_0}$. It is fast and produces a sample from the desired target conditional distribution. We formalize this result in the following lemma.

Lemma 3. *Let $g\left(\tau_{1:d_0} \mid \mu_{1:d_0}\right)$ be the joint distribution of $\tau_{1:d_0}$. Consider the following accept-reject algorithm:*

Algorithm 1 Generating the passage times $\tau_{1:d_0}$ given $\arg\min_{d' \in \{1:d_0\}} \tau_{d'} = d$

1: Generate $\tau_{1:d_0}$ from the joint distribution $g\left(\tau_{1:d_0} | \mu_{1:d_0}\right)$.

2: Accept $\tau_{1:d_0}$ if $\tau_d \leq \tau_{1:d_0}$.

3: Return to Step 1 otherwise.

Algorithm 1 generates samples from the conditional joint distribution of $\tau_{1:d_0}$, conditioned on the event $\tau_d \leq \tau_{1:d_0}$.

Proof of Lemma 3 is provided in Appendix D.

It can be verified that the acceptance ratio of Algorithm 1 is $M^{-1} = P\left(\tau_d \leq \tau_{1:d_0}\right)$ (see Robert and Casella, 2004) which depends on the drift parameters alone. If the drift parameters are ordered accordingly, so as to satisfy $\mu_d \geq \mu_{1:d_0}$, the acceptance ratios increase. The algorithm thus becomes faster as the sampler converges.

As noted earlier, sampling the latent inverse-gaussian distributed response times $\tau_{1:d_0}$ greatly simplifies computation. Most of the chosen priors, including the priors on the coefficients $\boldsymbol{\beta}$ in the fixed and random effects, are conjugate. Due to space constraints, the details are deferred to Section S.3 in the supplementary material.

6 Simulation Studies

In this section, we discuss the results of a synthetic numerical experiment. We simulate data from a complex longitudinal design that mimics the real PTC1 data set.

[1] We can see this in a simpler example. Suppose we are interested in generating a sample from the conditional distribution of $\boldsymbol{\tau} = (\tau_1, \tau_2)$ given $d = \arg\min_j \tau_j = 1$, where $\tau_i \sim \texttt{Uniform}(0, 1)$, $i = 1, 2$, independently. The conditional density of $\boldsymbol{\tau}$ given $d = 1$ is $f_{\boldsymbol{\tau}|d}(\tau_1, \tau_2) = 2$ if $0 < \tau_1 \leq \tau_2 < 1$, and $= 0$ otherwise. However, if we draw τ_1 from $\texttt{Uniform}(0, 1)$ first and let that realization be $\tau^\star$, and draw τ_2 from the truncated uniform distribution (left truncated at $\tau^\star$), then the pdf of the realization of (τ_1, τ_2) is $\tau^{\star -1}$.

Our generating model contains fixed effects components attributed to different input-response tone combinations and random components attributed to individuals.

We recall that our main objective here is to identify the similarities and differences between the underlying brain mechanisms associated with different input-response category combinations over time while also assessing their individual heterogeneity, as characterized by latent drift-diffusion processes whose parameters can be biologically interpreted. The estimation of the probability curves for different input-response combinations, while a good indicator of our model's fit, is not the main purpose of this endeavor. Traditional categorical probability models, such as multinomial probit or logit, are thus not relevant to the scientific problem we are trying to address here. We are also not aware of any other work in the drift-diffusion literature that attempts to estimate the underlying parameters from category response data alone. In view of this, we restrict our focus to evaluating the performance of the proposed biologically meaningful longitudinal inverse-probit mixed model but do not present comparisons with any other model.

Design. In designing the simulation scenario, we have tried to mimic our motivating category learning data sets. We chose $n = 20$ as the number of participants being trained over $T = 10$ blocks to identify $d_0 = 4$ tones. For each input tone and each block, there are $L = 40$ trials. We set the true $\mu_{d',s}(t)$ values in such a way that they are far from satisfying the constraint $\sum_{d'=1:d_0} \mu_{d',s} = k$, and the decision boundary is set to $b = 2$ for all (d', s). The true drift parameters and the true probabilities, averaged over the participants of each input-response category combination are shown in Figure 3.

There are four true clusters in total, two for correct categorizations, S_1, S_2, and two for incorrect categorizations, M_1, M_2, as follows: $S_1 = \{(1,1),(2,2)\}$, $S_2 = \{(3,3),(4,4)\}$, $M_1 = \{(1,2),(1,3),(2,1),(2,3),(3,4),(4,3)\}$, $M_2 = \{(1,4),(2,4),(3,1),(3,2),(4,1),(4,2)\}$. We may interpret M_1 as the cluster of difficult alternatives, and M_2 as the cluster of easy alternatives. Thus, there are similarities in overall trajectories of $\{T_1, T_2\}$ and $\{T_3, T_4\}$, differentiating between easy and hard category recognition problems. We experimented with 50 synthetic data sets generated according to this design.

Results. As the true drift parameters themselves do not satisfy the constraint, and the estimated drift parameters are on the constrained space, we cannot validate our method by its predictive performance of the drift parameters. Instead, the proposed method is validated in terms of the estimated probabilities.

Figure 4 shows the estimated posterior probability trajectories along with the 95% credible interval and the underlying true probability curves for every combination (d', s) in a typical scenario. The credible interval fails to capture the truth in two

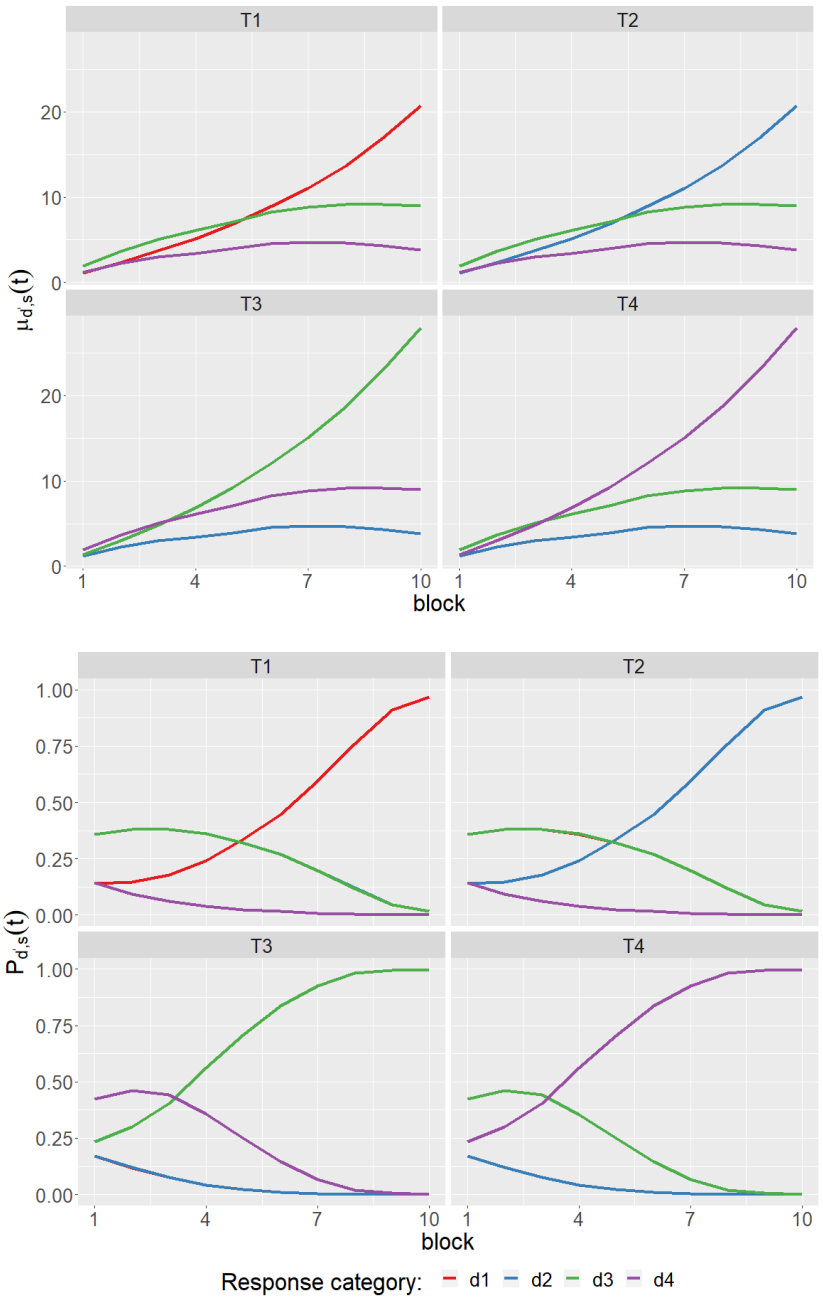


Figure 3: Description of the synthetic data: True values of the drift parameters averaged over the subjects, denoted by $\mu_{d',s}(t)$, and true probabilities $P\{d_{i,\ell,t} \mid s_{i,\ell,t}, \boldsymbol{\mu}^{(i)}_{1:d_0,s}(t)\}$ averaged over the subjects, denoted here by $P_{d',s}(t)$. Here T1, T2, T3, and T4 represent input categories 1 to 4, respectively. Some of the curves overlap according to the true clustering structure described in Section 6.

situations, when the true probability is very close to zero, or it is very close to one. The former case corresponds to classes with very low success probability, resulting in very few observations to estimate. The latter is underestimated as a consequence of the former since the probabilities add up to one.

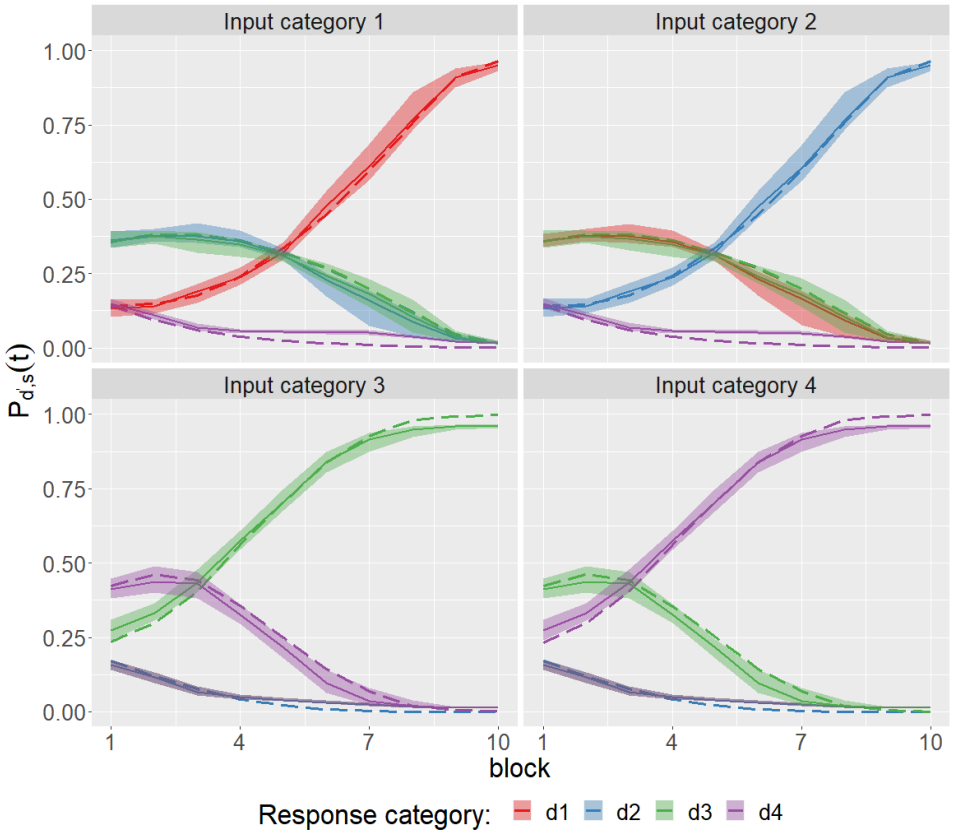


Figure 4: Results for synthetic data: Posterior trajectories of the probabilities for each combination of (d', s) over blocks estimated by the proposed model. The shaded areas represent the corresponding 95% point-wise credible intervals. The thick dashed lines represent underlying true curves some of which overlap according to the true clustering structure described in Section 6.

The results produced by our method are mostly stable and consistent across all synthetic data sets. There are, however, a few cases of incorrect cluster assignments, resulting in some outliers in each boxplot. Note that if an incorrect cluster assignment takes place, the probabilities of all input-response combinations are affected by that. For example, if a component of M_1 is wrongly assigned to M_2, then not only the probabilities of input-output combinations in M_1 and M_2 are affected, since the probabilities add up to one, those of S_1 and S_2 are also affected.

In estimating the probabilities, the overall mean squared error, i.e., the mean squared difference of the estimated and the true probabilities taking all combinations of (d, s, i, t) into account, came out to be 0.0028. Figure 5 provides a detailed description of the estimation of the probabilities for two input categories (one from each similarity group). As described for the individual simulation results, there are cases of under-estimation of the probabilities which are close to one, and consequently, over-estimation of the probabilities close to zero. However, the amount of departure from the true probability in each case is very small which can also be seen in the

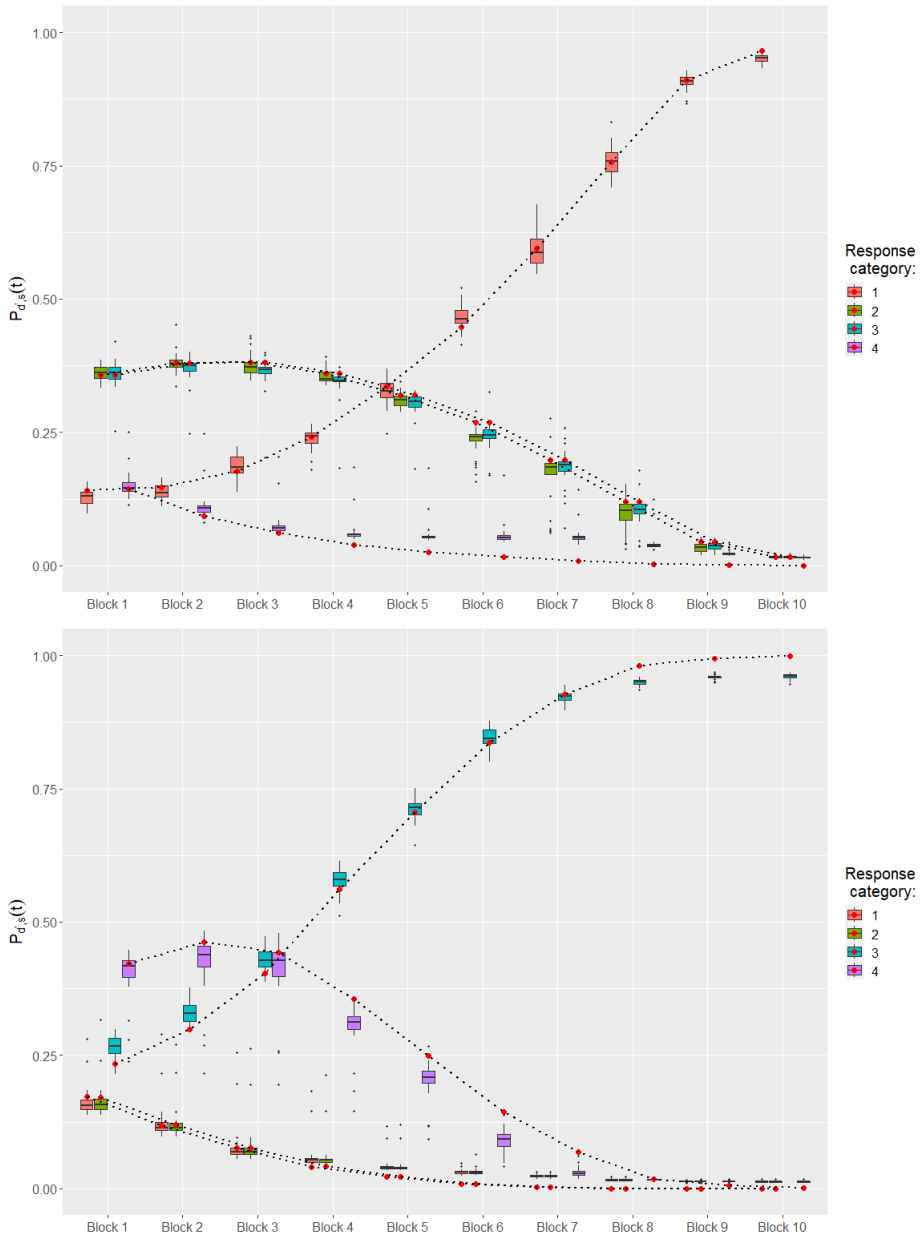


Figure 5: Results of the synthetic data: Boxplots of the estimated probabilities over 50 simulations, and true probabilities (in red dot) of each block and for two panels, one from each similarity group (panel T_1 in the top and T_3 in the bottom).

small overall MSE.

Further, the overall efficiency in identifying the true clustering structure is validated using Rand (Rand, 1971) and adjusted Rand (Hubert and Arabie, 1985) indices. The definitions of Rand and adjusted Rand indices are provided in Section S.6 in the supplementary material. The average Rand and adjusted Rand indices for our proposed method over 50 simulations 0.9105 and 0.8277, respectively, indicating high overall efficacy in correctly clustering the probability curves.

7 Applications

Analysis of the PTC1 data set. We present here the analysis of the PTC1 data set described in Section 2 using our proposed longitudinal inverse-probit mixed model. We first demonstrate the performance of the proposed method in estimating the probabilities associated with different (d, s) pairs. Figure 6 shows the 95% credible intervals for the estimated probabilities for different input tones, along with the average proportions of times an input tone was classified into different tone categories across subjects. The latter serves as the empirical estimate of the probabilities.

We observe that except for the input-response combination (1,1) in block 3 and some cases with a low number of data points, the 95% credible intervals include the corresponding empirical probabilities. An explanation of the occasional underperformance is given later in this section.

Next, we examine the clusters identified by the proposed model. Apart from the two clusters obtained for the success combinations ($d = s$), three clusters are additionally identified in the incorrect input-response combinations ($d \neq s$). The clusters of success combinations are $S_1 = \{(1,1),(2,2),(4,4)\}$ and $S_2 = \{(3,3)\}$, and of wrong allocations are $M_1 = \{(1,2),(1,4),(2,1),(2,4),(3,2),(4,1),(4,2)\}$, $M_2 = \{(1,3),(2,3),(3,4),(4,3)\}$, and $M_3 = \{(3,1)\}$. Figure 7 shows the input-response tone combinations color-coded as per cluster identity, and the proportion of times each pair of input-response tone combinations appeared in the same cluster after burnin. Figure 7 indicates that, while the clusters S_1, S_2, M_1 are stable, there is some instability among the other two clusters, namely M_2 and M_3.

Key findings. The clustering structure reveals that the low-dipping (T_3) response trajectories are different from the other three response categories. While for correct input-output tone combinations, S_2 forms a separate singleton cluster, for incorrect combinations, M_2 contains all the low-dipping trajectories, indicating their similarities across the panels. Also for T_3, faster increase of the probabilities of correct identification, as well as faster decay of probabilities of incorrect identification indicate that T_3 is easily distinguishable from other alternatives.

On the other hand, the trajectories of high-flat (T_1), low-rising (T_2) and high-falling (T_4) response categories are quite similar across panels. While for correct input-response combinations, these three form the cluster S_1, the corresponding incorrect tone combinations are clustered in M_1. The slower rise of the observed empirical probabilities for the elements in S_1 and the slower decay of the same for M_1 indicate that T_1, T_2 and T_4 are difficult to distinguish. However, in block 3 the empirical probabilities of correct input-response combinations differ moderately. While T_2 and T_4 show a relative drop in the empirical probabilities at block 3, T_1 shows a sudden pick in the same. This local dissimilarity of the trajectories at block 3, leads to a departure of the empirical probability of T_1 from the estimated credible band.

Next, we consider the results concerning the estimation of the drift parameters

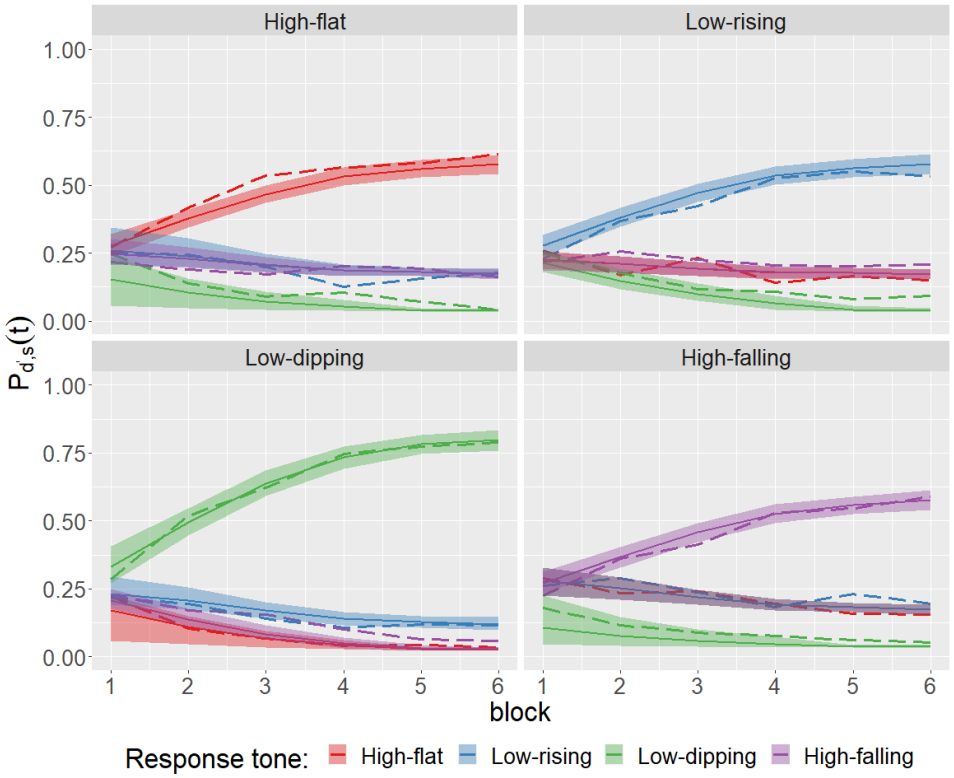


Figure 6: Results for PTC1 data: Estimated probability trajectories compared with average proportions of times an input tone was classified into different tone categories across subjects (in dashed line). The means across subjects are indicated by thick lines and the shaded regions indicate corresponding 95% coverage regions.

$\mu^{(i)}_{d',s}(t)$. As discussed in Section 3.2, given the identifiability constraints, the estimates of $\mu^{(i)}_{d',s}(t)$ can only be interpreted on a relative scale. Figure 8 shows the posterior mean trajectories and associated 95% credible intervals for the projected drift rates.

Importantly, our proposed mixed model also allows us to assess individual-specific parameter trajectories. Figure 9 shows the posterior mean trajectories and the associated 95% credible intervals for the drift rates $\mu^{(i)}_{d',s}$ estimated by our method for the different success combinations (d', s) for two participants - one with the best accuracy averaged across all blocks, and the other with the worst accuracy averaged across all blocks. These results suggest significant individual-specific heterogeneity. For the well-performing participant, the drift parameters are much higher than those for the poorly performing individual, indicating their ability to more quickly accumulate evidence compared to the poorly-performing adult. These differences persisted over all blocks with a small gradual increase over time.

Analysis of benchmark data. To validate the proposed method, we also analyzed tone learning data which, in addition to response accuracies, included accurate measurements of the response times. It was previously analyzed in Paulon *et al.* (2021)

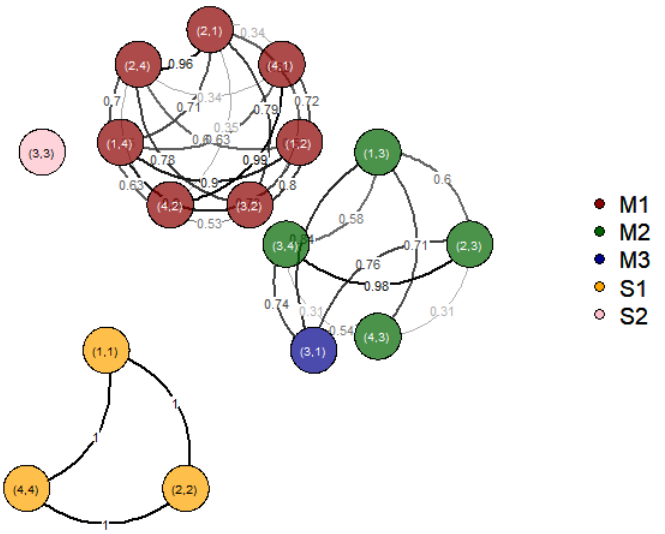


Figure 7: Results for PTC1 data: Network plot of similarity groups showing the intra and inter-cluster similarities of tone recognition problems. Each node is associated with a pair indicating the input-response tone category, (s, d). The number associated with each edge indicates the proportion of times the pair in the two connecting nodes appeared in the same cluster after burnin.

using the drift-diffusion model (2) which allowed inference on both the drift and the boundary parameters. For our analysis with the method proposed here, however, we ignored the response times. We observed that the estimates of the drifts produced by our proposed methodology match well with the estimates obtained by Paulon *et al.* (2021). A description of this 'benchmark' data set and other details of our analyses are provided in Section S.5 of the supplementary material.

8 Discussion, Conclusion, Broader Utility, and Future Work

Summary. In this article, we developed a novel longitudinal inverse-probit mixed categorical probability model. Our research was motivated by category-learning experiments where scientists are interested in using drift-diffusion models to understand how the decision-making mechanisms evolve as the participants get more training and experience. However, unlike traditional drift-diffusion analyses which require data on both response categories and response times, we only had usable records of response categories but no response times. To our knowledge, biologically interpretable latent drift-diffusion process-based categorical probability models had never been considered for such scenarios in the literature before. We addressed this need. Building on a previous work on longitudinal drift-diffusion mixed joint models for response categories

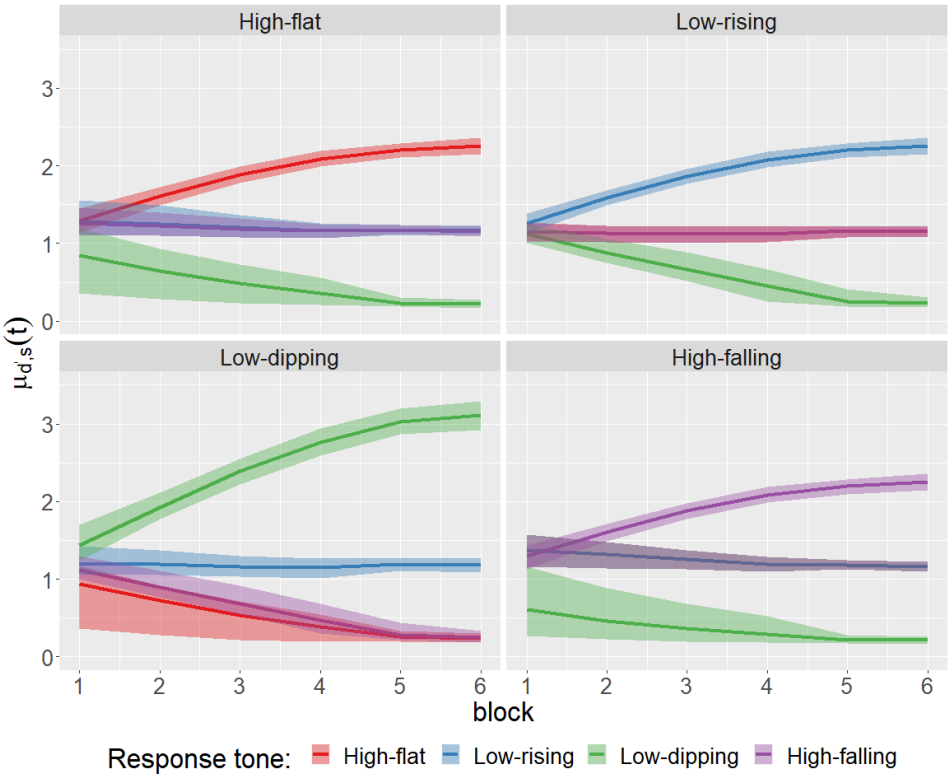


Figure 8: Results for PTC1 data: Estimated posterior mean trajectories of the population level drifts $\mu_{d',s}(t)$ for the proposed model. The shaded areas represent the corresponding 95% point-wise credible intervals.

and response times but now integrating out the response times, we obtained a new class of category probability models which we referred to here as the inverse-probit model. We explored parameter recoverability in such models, showing, in particular, that the offset parameters can not be recovered and drifts and boundaries both can not be recovered from data only on response categories. In our analyses, we thus focused on estimating the biologically more important drift parameters but kept the offsets and the boundaries fixed. We showed that with careful domain knowledge informed choices for the boundaries, the general trajectories of the drift parameters can be recovered by our proposed approach even in the complete absence of response times.

Conclusion. Overall, when it comes to making scientific inferences about drift-diffusion model parameters in the absence of data on response times, our work implies a mixed promise. On the downside, our work shows that the detailed interplay between drifts and boundaries cannot be captured. On the positive side, our results also suggest that, with our carefully designed model, and the fixed value of the boundary parameters appropriately chosen by experts, the general longitudinal trends in the drifts can still be estimated well. Caution should still be exercised not

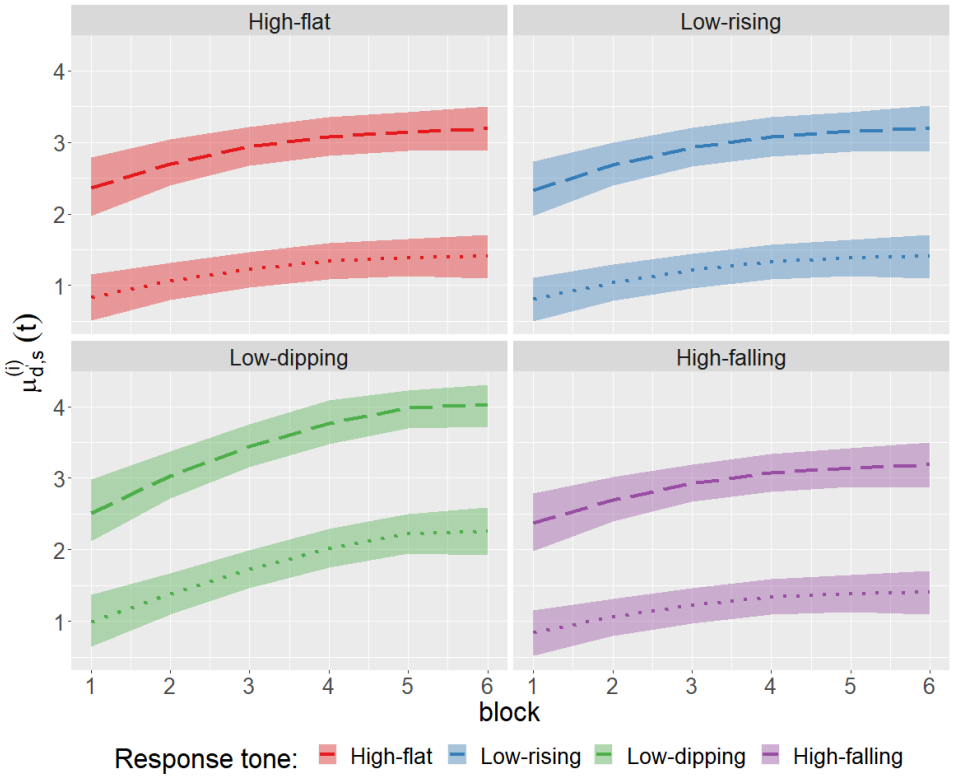


Figure 9: Results for PTC1 data: Estimated posterior mean trajectories for individual specific drifts $\mu^{(i)}_{d',s}(t) = \exp\{f_{d',s}(t) + u^{(i)}_C(t)\}$ for successful identification ($d' = s$) for two different participants - one performing well (dashed line) and one performing poorly (dotted line). The shaded areas represent the corresponding 95% point-wise credible intervals.

to over-interpret the results.

Broader utility in auditory neuroscience. The proposed model, we believe, has significant implications for auditory neuroscience. We focused here specifically on a pupillometry study for which the experimental paradigms need to be adapted to prioritize slow pupillary response, rendering the behavioral response times useless. However, as discussed in the Introduction, there could be many other situations where usable data on response times may not be available. The proposed model can be useful in such scenarios to understand the perceptual mechanisms underlying auditory decision-making.

Broader utility beyond auditory neuroscience. While we focused here on studying auditory category learning, the method proposed is applicable to other domains of behavioral neuroscience research studying categorical decision-making when the response times measurements are either not available or not reliable.

Broader utility in statistics. On the statistical side, the projection-based approach proposed here to impose non-standard identifiability conditions and address clustering problems within and between different panels is not restricted to inverse-probit models introduced here. They can be easily adapted to other classes of generalized linear models such as the widely popular logit and probit models and hence may also be of interest to a much broader statistical audience.

Future directions: The models and the analyses of the PTC1 data set presented here excluded the pupillometry measurements themselves. An important and challenging problem being pursued separately elsewhere is to see how those measurements relate to drift-diffusion model parameters.

Appendix

Appendix A Proof of Lemma 1

Proof. It is easy to check that the offset parameters δ_s are not identifiable since

$$P(d \mid s, \delta_s, \boldsymbol{\mu}_{1:d_0,s}, \mathbf{b}_{1:d_0,s}) = \int_{\delta_s}^{\infty} g(\tau \mid \delta_s, \mu_{d,s}, b_{d,s}) \prod_{d' \neq d} \{1 - G(\tau \mid \delta_s, \mu_{d',s}, b_{d',s})\}\, d\tau$$

$$= \int_{0}^{\infty} g(\tau \mid 0, \mu_{d,s}, b_{d,s}) \prod_{d' \neq d} \{1 - G(\tau \mid 0, \mu_{d',s}, b_{d',s})\}\, d\tau = P(d \mid s, 0, \boldsymbol{\mu}_{1:d_0,s}, \mathbf{b}_{1:d_0,s}).$$

Next we will show that the drift parameters and decision boundaries are not separately identifiable, even if we fix offset parameters to a constant.

First note that equation (3) can also be represented as

$$\int_{\delta_s}^{\infty} \cdots \int_{\delta_s}^{\infty} \prod_{d' \neq d} g(\tau_{d'} \mid \boldsymbol{\theta}_{d',s}) \int_{\delta_s}^{\wedge_{d \neq d'} \tau_{d'}} g(\tau_d \mid \boldsymbol{\theta}_{d,s}) d\tau_d \prod_{d' \neq d} d\tau_{d'}. \tag{A.1}$$

First observe that $\tau^\star = \wedge_{d' \neq d} \tau_{d'} = \tau^\star_{-1} \wedge \tau_1$, where $\tau^\star_{-1} = \wedge_{d' \neq \{1,d\}} \tau_{d'}$. Thus the integral above can be written as

$$\int_{\delta_s}^{\infty} \cdots \int_{\delta_s}^{\infty} \prod_{d' \neq \{1,d\}} g(\tau_{d'} \mid \boldsymbol{\theta}_{d',s}) \left\{ \int_{\delta_s}^{\infty} g(\tau_1 \mid \boldsymbol{\theta}_{1,s}) \int_{\delta_s}^{\tau^\star_{-1} \wedge \tau_1} g(\tau_d \mid \boldsymbol{\theta}_{d,s}) d\tau_d \right\} \prod_{d' \neq \{1,d\}} d\tau_{d'}$$

$$= \int_{\delta_s}^{\infty} \cdots \int_{\delta_s}^{\infty} \prod_{d' \neq \{1,d\}} g(\tau_{d'} \mid \boldsymbol{\theta}_{d',s}) \left\{ \int_{\delta_s}^{\tau^\star_{-1}} g(\tau_d \mid \boldsymbol{\theta}_{d,s}) \int_{\tau_d}^{\infty} g(\tau_1 \mid \boldsymbol{\theta}_{1,s}) d\tau_1 d\tau_d \right\} \prod_{d' \neq \{1,d\}} d\tau_{d'}.$$

Proceeding sequentially one can show that the integral above is the same as in (3).

Using the above we express the probability in (3) as in (A.1). As the offset parameter δ_s is already shown to be not identifiable, we need to fix the same. Without loss of generality, we fix the offset parameter at 0. The probability density function

of inverse Gaussian distribution, with parameters $\boldsymbol{\theta}_{d',s} = (\mu_{d',s}, b_{d',s})$ evaluated at $\tau_{d'}$, $g(\tau_{d'} \mid \boldsymbol{\theta}_{d',s})$ can be obtained from (1) by replacing $\delta_s = 0$ and $d = d'$.

Consider the transformation of $\tau_{d'}$ to $\tau^\star_{d'}$ as $\tau_{d'} = c^2\tau^\star_{d'}$, for some constant $c > 0$, and for all d'. Further, define $b^\star_{d',s} = b_{d',s}/c$ and $\mu^\star_{d',s} = c\mu_{d',s}$, for all d'. Then observe that

$$g(\tau_{d'} \mid \boldsymbol{\theta}_{d',s})d\tau_{d'} = (2\pi)^{-1/2} b^\star_{d',s}(\tau^\star_{d'})^{-3/2} \exp\left\{-(2\tau^\star_{d'})^{-1}\left(b^\star_{d',s} - \mu^\star_{d',s}\tau^\star_j\right)^2\right\} d\tau^\star_{d'} = g(\tau^\star_{d'} \mid \boldsymbol{\theta}^\star_{d',s})d\tau^\star_{d'},$$

where $g(\tau^\star_{d'} \mid \boldsymbol{\theta}^\star_{d',s})$ is the pdf of inverse Gaussian distribution with parameters $\mu^\star_{d',s}$ and $b^\star_{d',s}$, evaluated at the point $\tau^\star_{d'}$.

Applying the transformation on $\tau_{d'}$ for all d' we get that the integral in (A.1) with $\delta_s = 0$ is same as

$$\int_0^\infty \cdots \int_0^\infty \prod_{d' \neq d} g(\tau^\star_{d'} \mid \boldsymbol{\theta}^\star_{d',s}) \int_0^{\wedge_{d' \neq d}\tau^\star_{d'}} g(\tau^\star_d \mid \boldsymbol{\theta}^\star_{d,s})d\tau^\star_d \prod_{d' \neq d} d\tau^\star_{d'}.$$

As c is arbitrary, this shows that the drifts and boundaries are not separately estimable. □

Appendix B Proof of Theorem 1

Proof. Let $\mathbf{P}(\boldsymbol{\mu}_{1:d_0,s}) = \{p_1(\boldsymbol{\mu}_{1:d_0,s}), \ldots, p_{d_0}(\boldsymbol{\mu}_{1:d_0,s})\}^{\mathrm{T}}$ be the function, given by (4), from $\mathcal{S}_{0,k}$ to unit probability simplex Δ^{d_0-1}. For notational simplicity, we write $\boldsymbol{\mu}_{1:d_0,s} = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_{d_0})^{\mathrm{T}}$. We first find the matrix of partial derivative $\nabla\mathbf{P}$ with respect to $\boldsymbol{\mu}$.

For $\boldsymbol{\mu} \in \mathcal{S}_{0,k}$, $\mathbf{1}^T\boldsymbol{\mu} = k$, and hence the probability reduces to

$$p_d(\boldsymbol{\mu}) = \frac{\left(be^b\right)^{d_0}}{(2\pi)^{d_0/2}} \int_0^\infty \int_{\tau_d}^\infty \cdots \int_{\tau_d}^\infty |\boldsymbol{\tau}|^{-3/2} \exp\left\{-\frac{1}{2}\left(\mathbf{1}^T\boldsymbol{\tau}^{-1}\mathbf{1} + \boldsymbol{\mu}^T\boldsymbol{\tau}\boldsymbol{\mu}\right)\right\} d\boldsymbol{\tau}_{-d}d\tau_d,$$

for $d = 1, \ldots, d_0$, where $\boldsymbol{\tau} = \mathrm{diag}(\tau_1, \ldots, \tau_{d_0})$, and $\boldsymbol{\tau}_{-d}$ is the sub-vector of $\boldsymbol{\tau}$ excluding the d-th element. Next, differentiating $p_d(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$, we get

$$\begin{aligned}\frac{\partial p_d(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} &= \frac{\left(be^b\right)^{d_0}}{(2\pi)^{d_0/2}} \int_0^\infty \int_{\tau_d}^\infty \cdots \int_{\tau_d}^\infty |\boldsymbol{\tau}|^{-3/2}(-\boldsymbol{\tau}\boldsymbol{\mu}) \exp\left\{-\frac{1}{2}\left(\mathbf{1}^T\boldsymbol{\tau}^{-1}\mathbf{1} + \boldsymbol{\mu}^T\boldsymbol{\tau}\boldsymbol{\mu}\right)\right\} d\boldsymbol{\tau}_{-d}d\tau_d, \\ &= \begin{bmatrix}\mu_1\eta_2 & \cdots & \mu_{d-1}\eta_2 & \mu_d\eta_1 & \mu_{d+1}\eta_2 & \cdots & \mu_{d_0}\eta_2\end{bmatrix}^T,\end{aligned}$$

where $\eta_1 = -E\left\{\tau_1 \mathbb{I}\left(\tau_2 > \tau_1, \cdots, \tau_{d_0} > \tau_1\right) | \boldsymbol{\mu}\right\}$, and $\eta_2 = -E\left\{\tau_2 \mathbb{I}\left(\tau_2 > \tau_1, \cdots, \tau_{d_0} > \tau_1\right) | \boldsymbol{\mu}\right\}$, and $\mathbb{I}(A)$ is the indicator function of the event A. Here the expectation is considered under the joint distribution of $(\tau_1, \ldots, \tau_d)$, which is a product of independent inverse Gaussians. Clearly $\eta_1 > \eta_2 > 0$.

From the above derivation it is easy to obtain that

$$\nabla \mathbf{P}(\boldsymbol{\mu}) = \begin{bmatrix} \mu_1\eta_1 & \mu_2\eta_2 & \cdots & \mu_{d_0}\eta_2 \\ \mu_1\eta_2 & \mu_2\eta_1 & \cdots & \mu_{d_0}\eta_2 \\ \vdots & \vdots & \cdots & \vdots \\ \mu_1\eta_2 & \mu_2\eta_2 & \cdots & \mu_{d_0}\eta_1 \end{bmatrix} = \mathbf{M}\left\{(\eta_1 - \eta_2) I + \eta_2 \mathbf{1}\mathbf{1}^T\right\},$$

where $\mathbf{M} = \mathrm{diag}(\mu_1, \ldots, \mu_{d_0})$.

Now, suppose there exists $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ in $\mathcal{S}_k$ such that $\boldsymbol{\mu} \neq \boldsymbol{\nu}$ and $\mathbf{P}(\boldsymbol{\mu}) = \mathbf{P}(\boldsymbol{\nu})$. Define $\boldsymbol{\gamma} : [0,1] \to \mathbb{R}^{d_0}$ such that $\boldsymbol{\gamma}(t) = \boldsymbol{\mu} + t(\boldsymbol{\nu} - \boldsymbol{\mu})$, $t \in [0,1]$. Further, define $h(t) = \langle \mathbf{P}(\boldsymbol{\gamma}(t)) - \mathbf{P}(\boldsymbol{\mu}), \boldsymbol{\nu} - \boldsymbol{\mu} \rangle$, as the cross product of $\mathbf{P}(\boldsymbol{\gamma}(t)) - \mathbf{P}(\boldsymbol{\mu})$ and $\boldsymbol{\nu} - \boldsymbol{\mu}$. Then $h(1) = h(0) = 0$ under the proposition that $\mathbf{P}(\boldsymbol{\mu}) = \mathbf{P}(\boldsymbol{\nu})$. Therefore, by the Mean Value Theorem, as $\boldsymbol{\mu} \neq \boldsymbol{\nu}$, there exists some point $c \in (0,1)$ such that $\partial h(t)/\partial t|_{t=c} = 0$. Now,

$$\begin{aligned} \frac{\partial h(t)}{\partial t} &= \sum_{d'=1}^{d_0} (\nu_{d'} - \mu_{d'}) \frac{\partial}{\partial t} \left[p_{d'}\{\boldsymbol{\gamma}(t)\} - p_{d'}(\boldsymbol{\mu}) \right] \\ &= \sum_{d'=1}^{d_0} (\nu_{d'} - \mu_{d'}) \left\{ \frac{\partial}{\partial \boldsymbol{\gamma}} p_{d'}(\boldsymbol{\gamma}) \right\}^T \frac{\partial \boldsymbol{\gamma}(t)}{\partial t} \\ &= (\boldsymbol{\nu} - \boldsymbol{\mu})^T \nabla \mathbf{P}\{\boldsymbol{\gamma}(t)\} (\boldsymbol{\nu} - \boldsymbol{\mu}) \\ &= (\eta_1 - \eta_2)(\boldsymbol{\nu} - \boldsymbol{\mu})^T \boldsymbol{\Gamma}(t)(\boldsymbol{\nu} - \boldsymbol{\mu}) + \eta_2 (\boldsymbol{\nu} - \boldsymbol{\mu})^T M \mathbf{1}\mathbf{1}^T (\boldsymbol{\nu} - \boldsymbol{\mu}) \\ &= (\eta_1 - \eta_2)(\boldsymbol{\nu} - \boldsymbol{\mu})^T \boldsymbol{\Gamma}(t)(\boldsymbol{\nu} - \boldsymbol{\mu}), \end{aligned}$$

as $\mathbf{1}^T(\boldsymbol{\nu} - \boldsymbol{\mu}) = 0$, where $\boldsymbol{\Gamma}(t) = \mathrm{diag}\{\boldsymbol{\gamma}(t)\}$.

As every component of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is positive, for any $c \in (0,1)$, the matrix $\boldsymbol{\Gamma}(c)$ is positive definite. Further, as $\eta_1 > \eta_2$, $\partial h(t)/\partial t|_{t=c} = 0$ only if $\boldsymbol{\mu} = \boldsymbol{\nu}$, which contradicts the proposition. □

Appendix C Algorithm for Minimal Distance Mapping

The problem of finding projection of a point $\boldsymbol{\mu}$ onto the space $\mathcal{S}_{k,\varepsilon}$ is equivalent to the following non-linear optimization problem:

$$\text{minimize}_{\mathbf{w}} \|\mathbf{w} - \boldsymbol{\mu}\|^2 \qquad \text{such that} \quad \sum_{i=1}^{d_0} w_i = k,\ w_i \geq \varepsilon.$$

Duchi *et al.* (2008, Algorithm 1) provides a solution to the problem of projection of a given point $\boldsymbol{\mu}$ onto the space $\mathcal{S}_{k,\varepsilon}$ for $\varepsilon = 0$, which is modified for any given ε below.

Algorithm 2

INPUT: A vector $\boldsymbol{\mu}$, and a pair (k, ε).

1: Sort $\boldsymbol{\mu}$ into $\boldsymbol{\mu}^\star$ such that the elements of $\boldsymbol{\mu}^\star$ are in descending order.
2: Find $\rho = \max\left\{j : \mu_j^\star - j^{-1}\left(\sum_{l=1}^{j}\mu_l^\star - k\right) > \varepsilon\right\}$.
3: Define $\theta = \rho^{-1}\left\{\sum_{l=1}^{j}\mu_l^\star - k + (k-\rho)\varepsilon\right\}$.

OUTPUT: w such that $w_i = \max\{\mu_i - \theta, \varepsilon\}$.

Appendix D Proof of Lemma 3

Proof. We consider the unconditional distribution of $\tau_{1:d_0}$, given the parameters $\mu_{1:d_0}$ as the proposal distribution, g. Clearly, the proposal distribution g and the target conditional joint distribution f satisfies $f(\tau_{1:d_0}|\mu_{1:d_0})/g(\tau_{1:d_0}|\mu_{1:d_0}) \leq M$, where $M^{-1} = P(\tau_d \leq \tau_{1:d_0})$. Therefore, for any random sample $U \sim U(0,1)$, $f(\tau_{1:d_0}|\mu_{1:d_0}) \geq MUg(\tau_{1:d_0}|\mu_{1:d_0})$ if the sample satisfies the condition $\tau_d \leq \tau_{1:d_0}$, and $f(\tau_{1:d_0}|\mu_{1:d_0}) < MUg(\tau_{1:d_0}|\mu_{1:d_0})$ otherwise. Hence by Lemma 2.3.1 of Robert and Casella (2004), the algorithm above produces samples from the target distribution. □

Supplementary Materials

The supplementary materials detail the choice of the prior hyper-parameters, the MCMC algorithm used to sample from the posterior and some performance diagnostics, and the analysis of a real benchmark data set. Separate files additionally include R programs implementing the longitudinal inverse-probit mixed model developed in this article and the PTC1 data set analyzed in Section 7.

References

Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, **88**, 669–679.

Ashby, F. G., Noble, S., Filoteo, J. V., Waldron, E. M., and Ell, S. W. (2003). Category learning deficits in parkinson's disease. *Neuropsychology*, **17**, 115.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., and Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, **33**, 10–16.

Borooah, V. K. (2002). *Logit and probit: ordered and multinomial models*. Sage.

Brody, C. D. and Hanks, T. D. (2016). Neural underpinnings of the evidence accumulator. *Current Opinion in Neurobiology*, **37**, 149–157.

Brown, S. D. and Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, **57**, 153–178.

Burgette, L. F. and Nordheim, E. V. (2012). The trace restriction: An alternative identification strategy for the Bayesian multinomial probit model. *Journal of Business & Economic Statistics*, **30**, 404–410.

Burgette, L. F., Puelz, D., and Hahn, P. R. (2021). A symmetric prior for multinomial probit models. *Bayesian Analysis*, **16**, 991–1008.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., and Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, **14**, 1462.

Chandrasekaran, B., Yi, H.-G., and Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, **21**, 488–495.

Chandrasekaran, B., Yi, H.-G., Smayda, K. E., and Maddox, W. T. (2016). Effect of explicit dimensional instruction on speech category learning. *Attention, Perception, & Psychophysics*, **78**, 566–582.

Chhikara, R. (1988). *The inverse Gaussian distribution: Theory, methodology, and applications*. CRC Press.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

Cox, D. R. and Miller, H. D. (1965). *The theory of stochastic processes*. CRC Press.

de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag.

Deo, S. (2018). *Algebraic Topology*, volume 27 of *Texts and Readings in Mathematics*. Hindustan Book Agency, New Delhi.

Ding, L. and Gold, J. I. (2013). The basal ganglia's contributions to perceptual decision making. *Neuron*, **79**, 640–649.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279.

Dufau, S., Grainger, J., and Ziegler, J. C. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**, 1117.

Dunson, D. B. and Neelon, B. (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics*, **59**, 286–295.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, **11**, 89–102.

Filoteo, J. V., Lauritzen, S., and Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. *Psychological science*, **21**, 415–423.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian statistics, 4 (Peniscola, 1991)*, pages 169–193. Oxford Univ. Press, New York.

Glimcher, P. W. and Fehr, E. (2013). *Neuroeconomics: Decision making and the brain*. Academic Press.

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

Gunn, L. H. and Dunson, D. B. (2005). A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics*, **6**, 434–449.

Heekeren, H. R., Marrett, S., Bandettini, P. A., and Ungerleider, L. G. (2004). A general mechanism for perceptual decision-making in the human brain. *Nature*, **431**, 859.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.

Johndrow, J., Dunson, D., and Lum, K. (2013). Diagonal orthant multinomial probit models. In *Artificial Intelligence and Statistics*, pages 29–38.

Kim, S., Potter, K., Craigmile, P. F., Peruggia, M., and Van Zandt, T. (2017). A Bayesian race model for recognition memory. *Journal of the American Statistical Association*, **112**, 77–91.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, **16**, 526–558.

Leite, F. P. and Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, **72**, 246–273.

Llanos, F., McHaney, J. R., Schuerman, W. L., Yi, H. G., Leonard, M. K., and Chandrasekaran, B. (2020). Non-invasive peripheral nerve stimulation selectively enhances speech category learning in adults. *NPJ science of learning*, (1), 1–11.

Lu, J. (1995). *Degradation processes and related reliability models*. Ph.D. thesis, McGill University, Montreal, Canada.

McHaney, J. R., Tessmer, R., Roark, C. L., and Chandrasekaran, B. (2021). Working memory relates to individual differences in speech category learning: Insights from computational modeling and pupillometry. *Brain and Language*, **22**, 1–15.

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., and Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, **5**, 437–449.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, **2**, 321–359.

Parthasarathy, A., Hancock, K. E., Bennett, K., DeGruttola, V., and Polley, D. B. (2020). Bottom-up and top-down neural signatures of disordered multi-talker speech perception in adults with normal hearing. *Elife*, **9**, e51419.

Paulon, G., Llanos, F., Chandrasekaran, B., and Sarkar, A. (2021). Bayesian semiparametric longitudinal drift-diffusion mixed models for tone learning in adults. *Journal of the American Statistical Association,*, **116**, 1114–1127.

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, **39**, 204–214.

Purcell, B. A. (2013). *Neural mechanisms of perceptual decision making*. Vanderbilt University.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: Methods and case studies*. Springer.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

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

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

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

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, **20**, 260–281.

Reetzke, R., Xie, Z., Llanos, F., and Chandrasekaran, B. (2018). Tracing the trajectory of sensory plasticity across different stages of speech learning in adulthood. *Current Biology*, **28**, 1419–1427.

Roark, C. L., Smayda, K. E., and Chandrasekaran, B. (2021). Auditory and visual category learning in musicians and nonmusicians. *Journal of Experimental Psychology: General*.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Robison, M. K. and Unsworth, N. (2019). Pupillometry tracks fluctuations in working memory performance. *Attention, Perception, & Psychophysics*, **81**, 407–419.

Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. (1996). *Stochastic processes*. Wiley New York.

Rudin, W. (1991). *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience*, **2**, 33.

Sen, D., Patra, S., and Dunson, D. (2018). Constrained inference through posterior projections. *arXiv preprint arXiv:1812.05741*.

Smayda, K. E., Chandrasekaran, B., and Maddox, W. T. (2015). Enhanced cognitive and perceptual processing: A computational basis for the musician advantage in speech learning. *Frontiers in Psychology*, pages 1–14.

Smith, P. L. and Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, **27**, 161–168.

Smith, P. L. and Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, **32**, 135–168.

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

Wade, S. (2023). Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A*, **381**, 1–20.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, **106**, 3649–3658.

Wang, Y., Jongman, A., and Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, **113**, 1033–1043.

Whitmore, G. and Seshadri, V. (1987). A heuristic derivation of the inverse gaussian distribution. *The American Statistician*, **41**, 280–281.

Winn, M. B., Wendt, D., Koelewijn, T., and Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing*, **22**, 1–32.

Zekveld, A. A., Kramer, S. E., and Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, **32**, 498–510.

Supplementary Materials for

Bayesian Semiparametric Longitudinal Inverse-Probit Mixed Models for Category Learning

Minerva Mukhopadhyay[1](minervam@iitk.ac.in)
Jacie McHaney[2] (j.mchaney@pitt.edu)
Bharath Chandrasekaran[2](b.chandra@pitt.edu)
Abhra Sarkar[3] (abhra.sarkar@utexas.edu)

[1]Department of of Mathematics and Statistics,
Indian Institute of Technology,
Kanpur, UP 208016, India

[2]Department of Communication Science and Disorders,
University of Pittsburgh,
4028 Forbes Tower, Pittsburgh, PA 15260, USA

[3]Department of Statistics and Data Sciences,
University of Texas at Austin,
105 East 24th Street D9800, Austin, TX 78712, USA

The supplementary materials detail the choice of the prior hyper-parameters, the MCMC algorithm used to sample from the posterior and some performance diagnostics, and the analysis of a real benchmark data set. Separate files additionally include R programs implementing the longitudinal inverse-probit mixed model developed in this article and the PTC1 data set analyzed in Section 6 in the main paper.

S.1 Modelling the Drift Parameters

S.1.1 Functional Fixed Effects

We model the fixed effects functions $f_x(t)$ using flexible mixtures of B-spline bases (de Boor, 1978) that allow them to smoothly vary with time t while also depending locally on the indexing variable x as

$$f_x(t) = \sum_{k=1}^{K} \beta_{x,k} B_k(t) = \mathbf{B}(t)\boldsymbol{\beta}_x. \tag{S.1}$$

Here $\mathbf{B}(t) = \{B_1(t), \ldots, B_K(t)\}$ are a set of known locally supported basis functions spanning $[1, T]$, $\boldsymbol{\beta}_x = (\beta_{x,1}, \ldots, \beta_{x,K})^{\mathrm{T}}$ are associated unknown coefficients to be estimated from the data. Allowing the $\boldsymbol{\beta}_x$'s to flexibly vary with x can generate widely different shapes for different input-response category combinations.

Towards clustering the fixed effects curves, we introduce a set of latent variables z_x for each input-response category combination x with a shared state space $\{1, \ldots, z_{\max}\}$ and associated coefficient atoms $\boldsymbol{\beta}^\star_z = (\beta^\star_{z,1}, \ldots, \beta^\star_{z,K})^{\mathrm{T}}$, we let

$$(\boldsymbol{\beta}_x \mid z_x = z) = \boldsymbol{\beta}^\star_z, \quad \text{implying} \quad \{f_x(t) \mid z_x = z\} = f^\star_z(t) = \sum_{k=1}^{K} \beta^\star_{z,k} B_k(t), \tag{S.2}$$

To probabilistically cluster the $\boldsymbol{\beta}_x$'s, we next let

$$\begin{aligned} z_x &\sim \mathrm{Mult}(\boldsymbol{\pi}_z) = \mathrm{Mult}(\pi_1, \ldots, \pi_{z_{\max}}), \\ \boldsymbol{\pi}_z &\sim \mathrm{Dir}(\alpha/z_{\max}, \ldots, \alpha/z_{\max}). \end{aligned} \tag{S.3}$$

We next consider priors for the atoms $\boldsymbol{\beta}^\star_z$. We let

$$\boldsymbol{\beta}^\star_z \sim \mathrm{MVN}_K\{\boldsymbol{\mu}_{\beta,0}, (\sigma_a^{-2}\mathbf{I}_K + \sigma_s^{-2}\mathbf{P})^{-1}\}, \tag{S.4}$$

where $\mathrm{MVN}_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a K dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and $\mathbf{P} = \mathbf{D}^{\mathrm{T}}\mathbf{D}$, where the $(K-1) \times K$ matrix $\mathbf{D}$ is such that $\mathbf{D}\boldsymbol{\beta}$ computes the first order differences in $\boldsymbol{\beta}$. The model thus penalizes $\sum_{k=1}^{K}(\nabla\beta^\star_{z,k})^2 = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{P}\boldsymbol{\beta}$, the sum of squares of first order differences in $\boldsymbol{\beta}^{(i)}_u$ (Eilers and Marx, 1996). The variance parameter σ_s^2 models the smoothness of the functional atoms, smaller σ_s^2 inducing smoother $f^\star_z(t)$'s. Additional departures from $\boldsymbol{\mu}_{\beta,0}$ are explained by the other variance component σ_a^2. We assign half Cauchy priors on the variance parameters as

$$\sigma_s^2 \sim \mathrm{C}^+(0, 1), \qquad \sigma_a^2 \sim \mathrm{C}^+(0, 1).$$

S.1.2 Functional Random Effects

We allow different random effects $u^{(i)}_C(t)$ and $u^{(i)}_I(t)$ for correct (C) (when $d = s$) and incorrect (I) (when $d \neq s$) identifications, respectively, as

$$u^{(i)}_{d,s}(t) = u^{(i)}_C(t) \quad \text{when } d = s, \qquad u^{(i)}_{d,s}(t) = u^{(i)}_I(t) \quad \text{when } d \neq s.$$

Suppressing the subscripts to simplify notation, we model the time-varying random effects components $u^{(i)}(t)$ as

$$\begin{aligned} u^{(i)}(t) &= \textstyle\sum_{k=1}^{K} \beta_{u,k}^{(i)} B_k(t) = \mathbf{B}(t)\boldsymbol{\beta}_u^{(i)}, \\ \boldsymbol{\beta}_u^{(i)} &\sim \mathrm{MVN}_K\{\mathbf{0}, (\sigma_{u,a}^{-2}\mathbf{I}_K + \sigma_{u,s}^{-2}\mathbf{P})^{-1}\}, \end{aligned} \tag{S.5}$$

where $\boldsymbol{\beta}_u^{(i)} = (\beta_{1,u}^{(i)}, \ldots, \beta_{K,u}^{(i)})^{\mathrm{T}}$ are subject-specific spline coefficients. We assign non-informative half-Cauchy priors on the variance parameters as

$$\sigma_{u,s}^2 \sim \mathrm{C}^+(0,1), \qquad \sigma_{u,a}^2 \sim \mathrm{C}^+(0,1).$$

S.2 Prior Hyper-parameters and Initialization

The random effects of the inverse-probit mixed model are all initialized at zero. The variance and smoothing parameters are initially set to 0.1 each. The location parameter of the prior on $\boldsymbol{\beta}_z^\star$, $\boldsymbol{\mu}_{\beta,0}$ is set to $(1, \ldots, 1)$. This choice of $\boldsymbol{\beta}_z^\star$ would set the expected value of $\mu_x^{(i)}(t)$ to 1, which is supported empirically. The value of the parameter α is set to 1.

S.3 Posterior Inference

Posterior inference for the longitudinal drift-diffusion mixed model, described in Section 3 in the main paper, is based on samples drawn from the posterior using an MCMC algorithm. The algorithm carefully exploits the conditional independence relationships encoded in the model as well as the latent variable construction of the model. Sampling the latent inverse-Gaussian distributed response times, in particular, greatly simplifies computation.

In what follows, $\boldsymbol{\zeta}$ denotes a generic variable that collects all other variables not explicitly mentioned, including the data points. Also, p_0 will sometimes be used as a generic for a prior distribution without explicitly mentioning its hyper-parameters. The notation x is used to abbreviate (d', s). The sampler for the inverse-probit mixed model of Section 3 iterates between the following steps.

1. **Sampling $\tau_{1:d_0}^{(i,l)}(t)$:** Suppose the i-th individual selects the output tone d, in the t-th block, l-th trial, given the input tone s. Then $\tau_1^{(i,l)}(t), \ldots, \tau_{d_0}^{(i,l)}(t)$ is generated as in Algorithm 1 (see Section 4) from the joint distribution of $\tau_1^{(i,l)}(t), \ldots, \tau_{d_0}^{(i,l)}(t)$ given $\mu_{1,s}^{(i)}(t), \ldots, \mu_{d_0,s}^{(i)}(t)$, followed by an accept-reject step.

2. **Updating the components of fixed effects $f_x(t)$:**

   (a) The latent variable $\mathbf{z}_x$, indicating the group identities of $\boldsymbol{\beta}_x$, follows multinomial distribution with $z_{\max}$ labels and probabilities $P(z_x = z|\boldsymbol{\zeta})$, $z = 1, \ldots, z_{\max}$ a posteriori. The probability $P(z_x = z|\boldsymbol{\zeta}) \propto \pi_z \times l_z$, where l_z is the likelihood of $\boldsymbol{\beta}_x$

evaluated at $\boldsymbol{\beta}_z$. Let $L^\star$ be the set of all trials corresponding to input-output tones $x = (s, d)$ for i-th individual and t-th block, and $n_x^{(i)}(t)$ be the cardinality of $L^\star$. Furthermore, let $\tau_x^{(i)}(t) = \sum_{l \in L^\star} \tau_x^{(i,l)}(t)$, $\tau_x(t) = \sum_i \tau_x^{(i)}(t)$, and $n_x(t) = \sum_i n_x^{(i)}(t)$. A little algebra shows that the likelihood of $\boldsymbol{\beta}_x$ is Gaussian with variance matrix $\boldsymbol{\Sigma}_{\beta,x} = \left\{\sum_t \tau_x(t)\mathbf{B}(t)^T\mathbf{B}(t)\right\}^{-1}$, and mean vector $\boldsymbol{\mu}_{\beta,x} = \boldsymbol{\Sigma}_{\beta,x}\left\{\sum_t \mathbf{B}(t)M_x(t)\right\}$, where $M_x(t) = 2n_x(t) - \sum_i u_x^{(i)}(t)\tau_x^{(i)}(t)$. Therefore, l_z is the Gaussian likelihood with mean $\boldsymbol{\mu}_{\beta,x}$ and variance $\boldsymbol{\Sigma}_{\beta,x}$, evaluated at $\boldsymbol{\beta}_z$.

(b) Let $N_z = \sum_x 1(z_x = z)$, $z = 1, \ldots, z_{\max}$, where $1(\cdot)$ is the indicator function. Then the conditional posterior of $\boldsymbol{\pi}_z$ is Dirichlet with parameters $\alpha/z_{\max}+N_1, \ldots, \alpha/z_{\max}+N_{z_{\max}}$.

(c) The full conditional posterior distribution of the coefficient atoms $\boldsymbol{\beta}_z^\star$ is Gaussian with variance-covariance matrix $\boldsymbol{\Sigma}_{\beta,z}^\star$ and $\boldsymbol{\mu}_{\beta,z}^\star$, where $\boldsymbol{\Sigma}_{\beta,z}^{\star,-1} = \sum_{x:z_x=z} \boldsymbol{\Sigma}_{\beta,x}^{-1} + \boldsymbol{\Sigma}_{\beta,0}^{-1}$, and $\boldsymbol{\mu}_{\beta,z}^\star = \boldsymbol{\Sigma}_{\beta,z}^\star\left[\sum_{x:z_x=z} \boldsymbol{\Sigma}_{\beta,x}^{-1}\boldsymbol{\mu}_{\beta,x} + \boldsymbol{\Sigma}_{\beta,0}^{-1}\boldsymbol{\mu}_{\beta,0}\right]$, where $\boldsymbol{\Sigma}_{\beta,0}^{-1} = (\sigma_a^{-2}\mathbf{I}_K + \sigma_s^{-2}\mathbf{P})$.

3. **Updating the components of random effects:** We use the generic notation U to indicate the correct (C, i.e., $d = s$) or incorrect (I, i.e, $d \neq s$) cases. Define $\tau_U^{(i)}(t) = \sum_{x:x\in U} \tau_x^{(i)}(t)$, $n_U^{(i)}(t) = \sum_{x:x\in U} n_x^{(i)}(t)$, $f\tau_U^{(i)}(t) = \sum_{x:x\in U} \tau_x^{(i)}(t)f_x(t)$, $\boldsymbol{\Sigma}_{U,0}^{-1} = \sigma_{U,a}^{-2}\mathbf{I}_K + \sigma_{U,s}^{-2}\mathbf{P}$, and $\boldsymbol{\Sigma}_U^{(t)-1} = \sum_t \tau_U^{(i)}(t)\mathbf{B}(t)^T\mathbf{B}(t)$. The conditional posterior of $\boldsymbol{\beta}_U^{(i)}$ is Gaussian with covariance $\boldsymbol{\Sigma}_{U,\text{post}}^{(i)} = \left(\boldsymbol{\Sigma}_{U,0}^{-1} + \boldsymbol{\Sigma}_U^{(i)-1}\right)^{-1}$, and location parameter $\boldsymbol{\mu}_C^{(i)} = \boldsymbol{\Sigma}_{U,\text{post}}^{(i)}\boldsymbol{\Sigma}_U^{(i)-1}\left[\sum_t \left\{2n_U^{(i)}(t) - f\tau_U^{(i)}(t)\right\}\mathbf{B}(t)\right]$, respectively.

4. **Updating the precision and smoothing parameters:** The precision and smoothness parameters involved in the fixed effects part are σ_a^2 and σ_s^2, and those involved in the random effects part are $\sigma_{U,a}$ and $\sigma_{U,s}^2$, $U = C, I$. We update these variance components using Metropolis-Hastings algorithm with log-normal proposal distributions centered on the previous sample values.

5. **Estimation of probability:** For each (s, i, t), we calculate the probability of selecting the d-th response in the following way: Let $g\{\cdot \mid \mu_{d',s}^{(i)}(t)\}$ be the pdf of inverse Gaussian distribution of the form (1) with parameters $\delta_s = 0$, $b_{d',s} = 2$ and $\mu_{d',s} = \mu_{d',s}^{(i)}(t)$. We generate $M = 2000$ independent samples $\boldsymbol{\tau}_m = [\tau_{1,m}, \ldots, \tau_{d_0,m}]^T$, $m = 1, \ldots, M$, where $\tau_{d',m}$ is generated independently from $g\{\cdot \mid \mu_{d',s}^{(i)}(t)\}$. Among these M independent samples, the proportion of occurrences of $\{\tau_{d,m} \leq \wedge_{d'=1:d_0}\tau_{d',m}\}$ is considered as the estimated probability of selecting d^{th} response.

The results reported in this article are all based on $5,000$ MCMC iterations with the initial $2,000$ iterations discarded as burn-in. The remaining samples were further thinned by an interval of 5. We programmed in `R`. The codes are available as part of the supplementary material. A 'readme' file, providing additional details for a practitioner, is also included in the supplementary material. In all experiments, the posterior samples produced very stable estimates of the population and individual level parameters of interest. MCMC diagnostic checks were not indicative of any convergence or mixing issues.

S.4 MCMC Diagnostics

This section presents some convergence diagnostics for the MCMC sampler described in the main manuscript. The results presented here are for the PTC1 data set. Diagnostics for the simulation experiments and the benchmark data were similar and hence omitted.

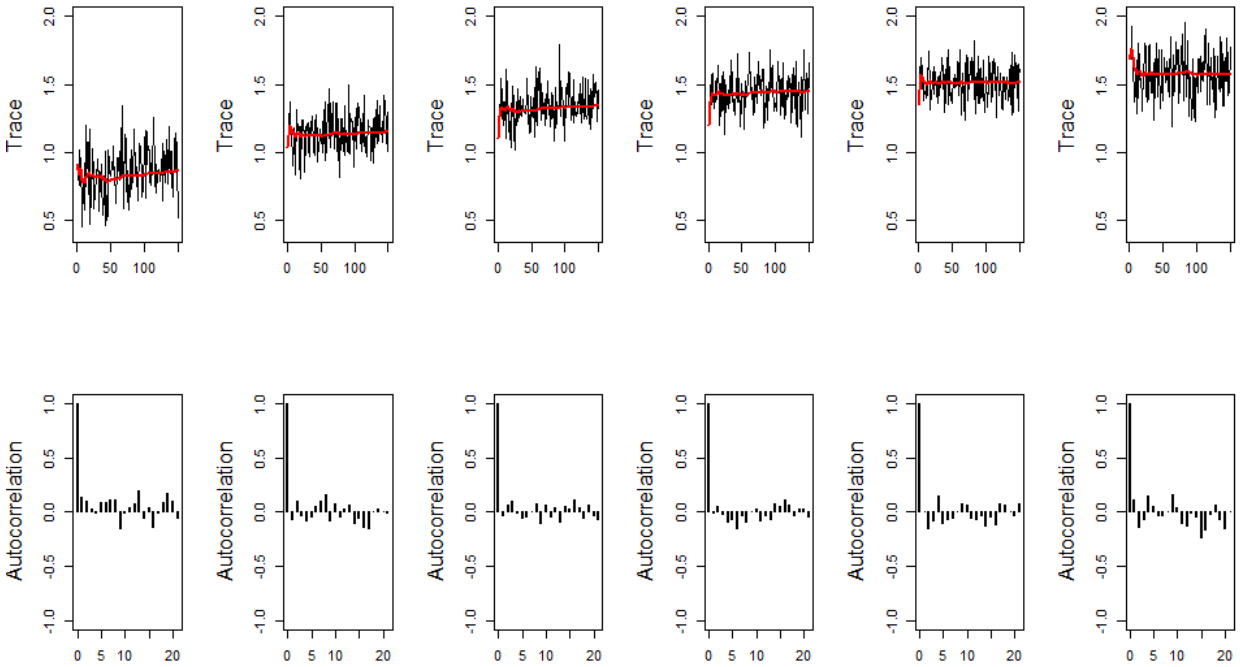


Figure S.1: Analysis of PTC1 data: Trace plots (top) and auto-correlation plots (bottom) of the individual drift rates $\mu_{1,1}^{(1)}(t)$ corresponding to the success categorization of tone T_1 evaluated at each of the training blocks. In each panel, the solid red line shows the running mean. Results for other drift parameters were very similar.

Figure S.1 shows the trace plots and auto-correlation of some individual level parameters at different training blocks. These results are based on the MCMC thinned samples. As these figures show, the running means are very stable and there seems to be no convergence issues. Additionally, the Geweke test (Geweke, 1992) for stationarity of the chains, which formally compares the means of the first and last part of a Markov chain, was also performed. If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke statistic has an asymptotically standard normal distribution. The results of the test, reported in Table S.1, indicate that convergence was satisfactory.

	$t=1$	$t=2$	$t=2$	$t=2$	$t=2$	$t=2$
Geweke statistics	-1.233	-0.392	-0.678	-0.136	0.440	0.339
p-value	0.217	0.695	0.498	0.892	0.660	0.734

Table S.1: Geweke statistics and associated p-values assessing convergence of the of the individual level drift parameters $\mu_{1,1}^{(1)}(t)$ corresponding to the success categorization of tone T_1 evaluated at each of the training blocks. Results for other drift parameters were very similar.

S.5 Analysis of Benchmark Data

Description of the data. The data set we consider next is a multi-day longitudinal speech category training study reported previously in Reetzke *et al.* (2018) and analyzed previously in Paulon *et al.* (2021). In this study, $n = 20$ participants were trained to learn 4 tones, namely, high-level (T1), low-rising (T2), low-dipping (T3), or high-falling (T4) tone, respectively. The trials were administered in blocks, each comprising 40 categorization trials. Participants were trained across several days, with five blocks on each day. On each trial, participants indicated the tone category they heard via button press on a computer keyboard. Following the button press, they were given corrective feedback. The data consist of tone responses and associated response times for different input tones for the 20 participants. We focus here on the first two days of training (10 blocks in total) as they exhibited the steepest improvement in learning as well as the most striking individual differences relative to any other collection of blocks.

Analysis. We first demonstrate the performance of the proposed method in estimating the probabilities associated with different (d, s) pairs. Figure S.2 shows the 95% credible intervals for the estimated probabilities for different input tones along with the average proportions of times an input tone was classified into different tone categories across subjects.

Observe that, except in situations with a very small number of data points the 95% credible intervals include the empirical probabilities. Further, the estimated credible region is narrow enough implying high precision of the inference.

Next, consider the clustering results. We obtained two clusters each in pairs of success combinations ($d = s$) and in the wrong allocations ($d \neq s$). The clusters of success combinations are $S_1 = \{(1,1),(3,3)\}$ and $S_2 = \{(2,2),(4,4)\}$, and that in wrong allocations are $M_1 = \{(1,2),(2,1),(2,3),(3,2),(4,1),(4,2)\}$, and $M_2 = \{(1,3),(1,4),(2,4),(3,1),(3,4),(4,3)\}$. The network plot in Figure S.3 shows the stability of the clusters over the MCMC iterations.

From an overall perspective, the trajectory of 'High-level' (T_1) and 'Low-dipping' (T_3) are similar with two wrong allocations from M_2 and one from M_1, and that of 'Low-rising' (T_2) and 'High-failing' (T_4) are similar with two wrong allocations from M_1 and one from M_2. These similarities in the overall trajectories of $\{T_1, T_3\}$ and $\{T_2, T_4\}$ were also noted by Paulon *et al.* (2021).

Next, we consider the estimation of the underlying drift parameters $\mu_{d',s}^{(i)}(t)$. Due to the identifiability constraints, the estimates of $\mu_{d',s}^{(i)}(t)$ can only be observed on a relative

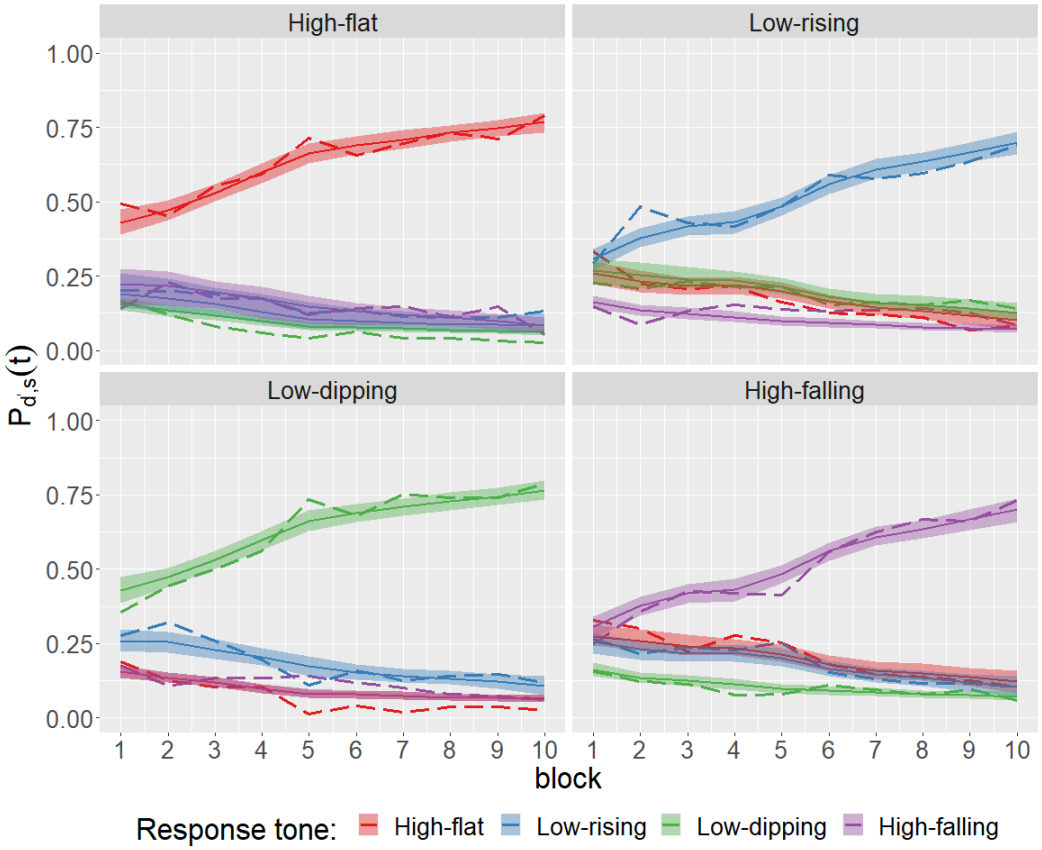


Figure S.2: Results for the benchmark data: Estimated probability trajectories compared with average proportions of times an input tone was classified into different tone categories across subjects (in dashed line). High-flat tone responses are shown in red; low-rising in blue; low-dipping in green; and high-falling in purple.

scale. Figure S.4 shows the posterior mean trajectories and associated 95% credible intervals for the projected drift rates estimated by our method for different combinations of (d', s). In comparison with the previous analysis of Paulon *et al.* (2021), the trajectories of our estimated drift rates show significant similarity throughout.

Figure S.5 shows the posterior mean trajectories and associated 95% credible intervals for the drift rates $\mu^{(i)}_{d',s}(t)$ for the different correct combinations (d', s) with $d' = s$ for two participants - the one with the best accuracy averaged, and the one with the worst accuracy averaged across all blocks. For the well-performing participant, the drift trajectories increase rapidly and for the poorly performing candidate, on the other hand, the drift trajectories increase very slowly. Once again, in spite of the limitation of inferring on a relative scale, the relative differences of the best and worst performing participants across blocks show great similarity with the inference of Paulon *et al.* (2021).

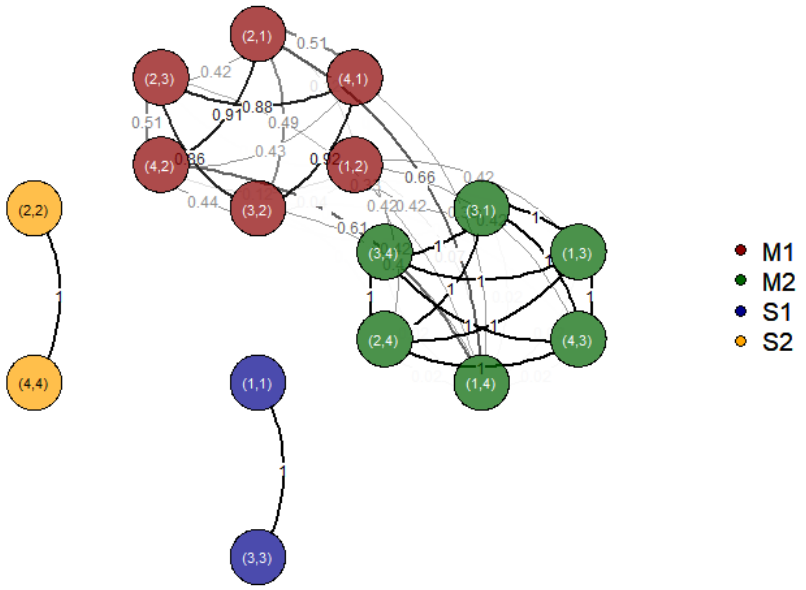


Figure S.3: Results for the benchmark data: Network plot of similarity groups showing the intra and inter-cluster similarities. Each node is associated with a pair indicating the input-response tone category (s, d). The number associated with each edge indicates the proportion of times the pair in the two connecting nodes appeared in the same cluster after burning.

S.6 Rand and Adjusted Rand Indices

Rand Index. Given a set of n objects $S = \{s_1, \ldots, s_n\}$, let $\mathbf{U} = \{U_1, \ldots, U_R\}$ and $\mathbf{V} = \{V_1, \ldots, V_C\}$ represent two different partitions of the objects in S such that $\cup_{i=1}^{R} U_i = S = \cup_{j=1}^{C} V_j$ and $U_i \cap U_{i'} = \emptyset = V_j \cap V_{j'}$ for $1 \leq i \neq i' \leq R$ and $1 \leq j \neq j' \leq C$. Rand index estimates the similarity between the allocations of S in $\mathbf{U}$ and $\mathbf{V}$.

Let a be the number of pairs of objects that are placed in the same partition in $\mathbf{U}$ and the same partition in $\mathbf{V}$, and b be the number of pairs of objects that are in different partitions of $\mathbf{U}$, as well as in different partitions of $\mathbf{V}$. Here a and b can be interpreted as agreements in $\mathbf{U}$ and $\mathbf{V}$, and the total number of pairs is $\binom{n}{2}$. The Rand index (Rand, 1971) is

$$\texttt{RI} = (a + b)/\binom{n}{2}.$$

The Rand index lies between 0 and 1. When the two partitions agree perfectly, the `RI` takes the value 1.

Adjusted Rand Index. The expected value of the Rand index of two random partitions does not take a constant value. The adjusted Rand index (Hubert and Arabie, 1985) assumes generalized hypergeometric distribution as the model of randomness, and makes a base and scale change of the quantity $(a+b)$, defined above, so that the resultant quantity is bounded by $[-1, 1]$ and has expected value 0 under completely random allocation.

Let $n_{i,j}$ be the number of object that are both in i^{th} partition of $\mathbf{U}$ and j^{th} partition of $\mathbf{V}$, n_i and n_j be the total number of components in i^{th} partition of $\mathbf{U}$, and j^{th} partition of $\mathbf{V}$, respectively.

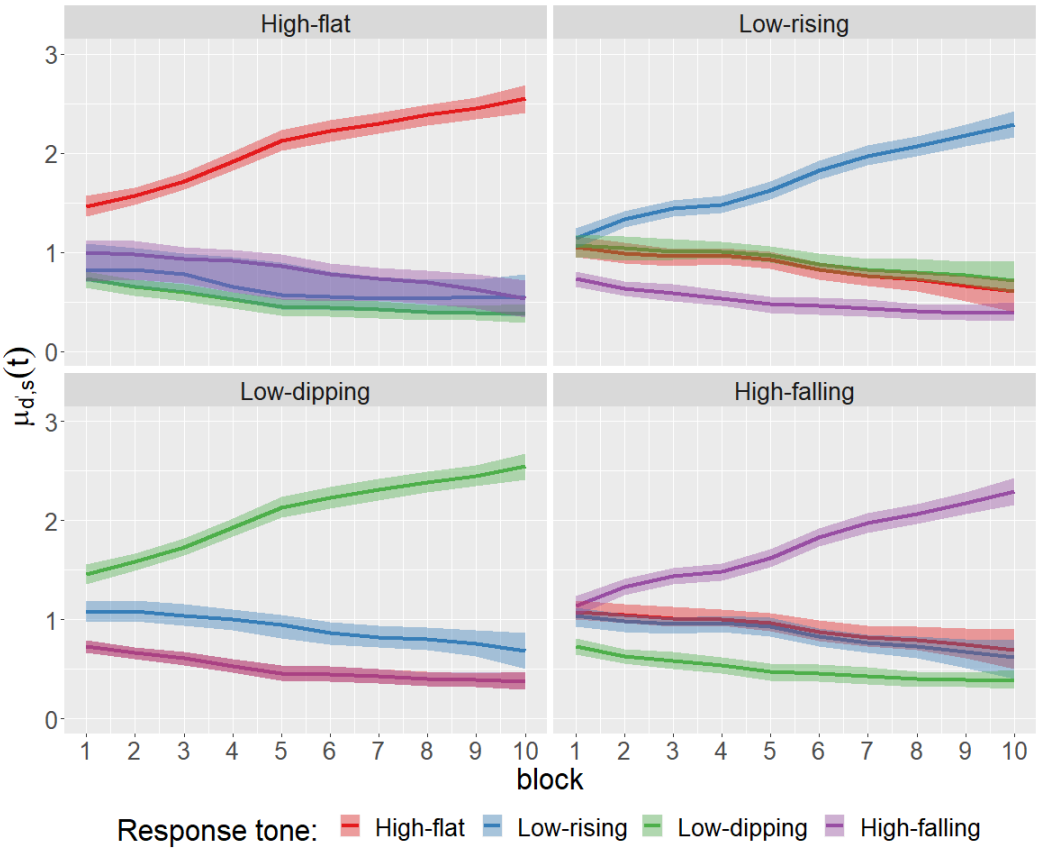


Figure S.4: Results for the benchmark data: Estimated posterior mean trajectories of the population level drifts $\mu_{d',s}(t)$ for the proposed model. The shaded areas represent the corresponding 95% pointwise credible intervals. Parameters for the high-flat tone response category are shown in red; low-rising in blue; low-dipping in green; and high-falling in purple.

The expression $a + d$ can be simplified to a linear transformation of $\sum_{i,j} \binom{n_{i,j}}{2}$. Further, under the generalized hypergeometric model, it can be shown that

$$\mathbb{E}\left[\sum_{i,j}\binom{n_{ij}}{2}\right] = \left[\sum_{i}\binom{n_i}{2}\sum_{j}\binom{n_j}{2}\right] / \binom{n}{2}.$$

Therefore, scaled the difference of linear transformed $(a + b)$ and its expectation is the adjusted Rand index, defined as:

$$\texttt{ARI} = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - \left[\sum_i \binom{n_i}{2}\sum_j \binom{n_{i,j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \left[\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}\right] / \binom{n}{2}}.$$

The expected value of `ARI` index is zero and the range is $[-1, 1]$. Like the `RI`, the `ARI` also takes the value 1, when the two partitions agree perfectly.

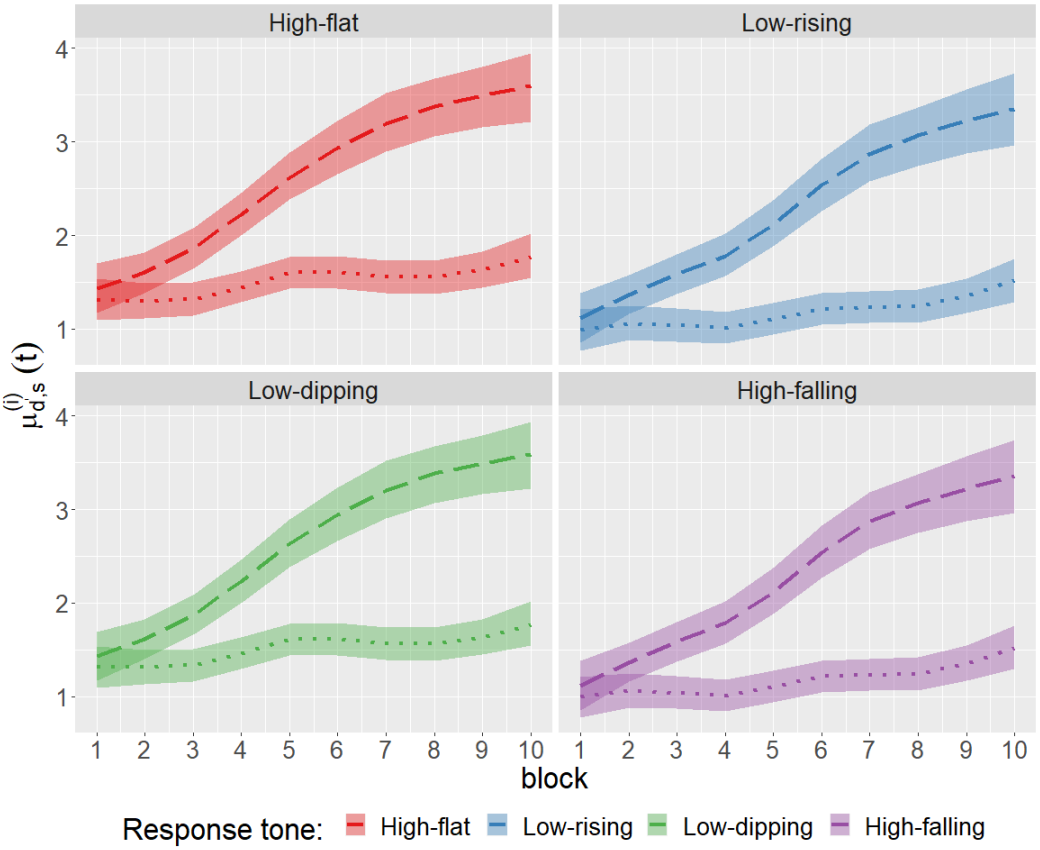


Figure S.5: Results for the benchmark data: Estimated posterior mean trajectories for individual specific drifts $\mu^{(i)}_{d',s}(t) = \exp\{f_{d',s}(t) + u^{(i)}_C(t)\}$ for correct identification ($d' = s$) for two different participants - one performing well (dashed line) and one performing poorly (dotted line). The shaded areas represent the corresponding 95% point-wise credible intervals. Parameters for the high-flat tone response category are shown in red; low-rising in blue; low-dipping in green; and high-falling in purple.