Supplementary Materials for

Bayesian Semiparametric Longitudinal Inverse-Probit Mixed Models for Category Learning

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.

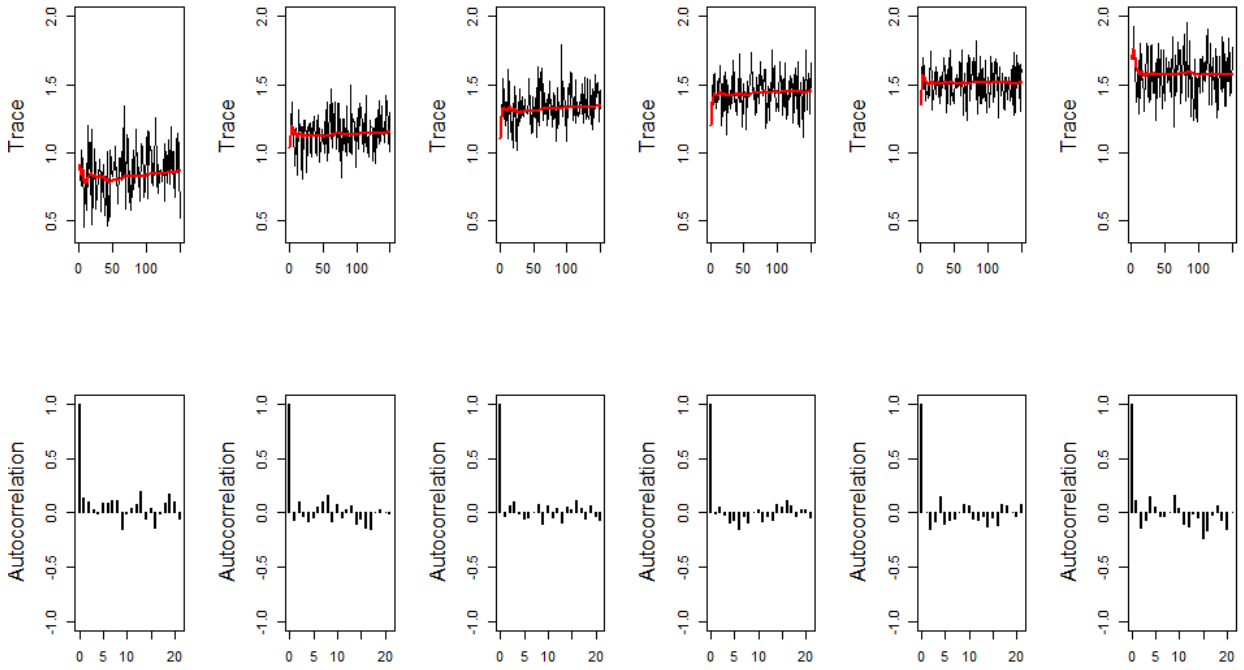

Figure S.1: Analysis of PTC1 data: Trace plots (top) and auto-correlation plots (bottom) of the individual drift rates $\mu^{(1)}_{1,1}(t)$ corresponding to the success categorization of tone T_1 evaluated at each of the training blocks. In each panel, the solid red line shows the running mean. Results 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 $\mu^{(i)}_{d',s}(t)$. Due to the identifiability constraints, the estimates of $\mu^{(i)}_{d',s}(t)$ can only be observed on a relative

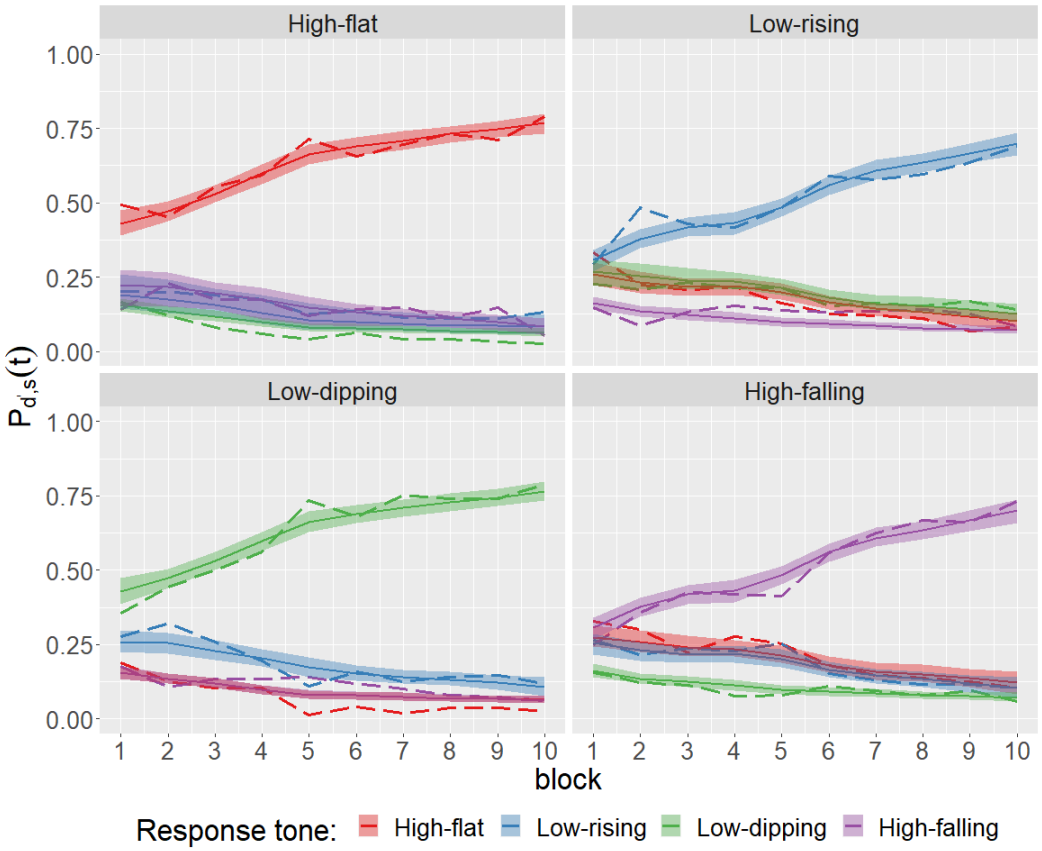

Figure S.2: Results for the benchmark data: Estimated probability trajectories compared with average proportions of times an input tone was classified into different tone categories across subjects (in dashed line). High-flat tone responses are shown in red; low-rising in blue; low-dipping in green;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

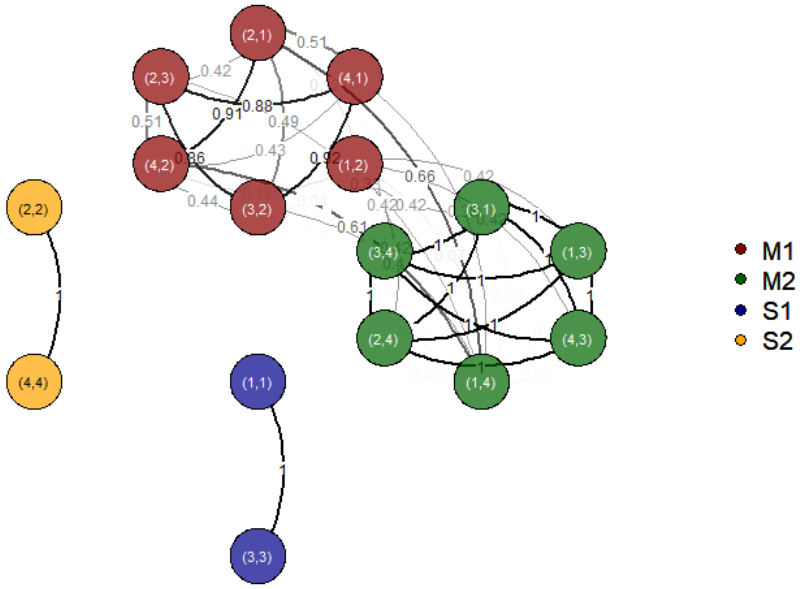

Figure S.3: Results for the benchmark data: Network plot of similarity groups showing the intra and inter-cluster similarities. Each node is associated with a pair indicating the input-response tone category (s, d). The number associated with each edge indicates the proportion of times the pair in the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

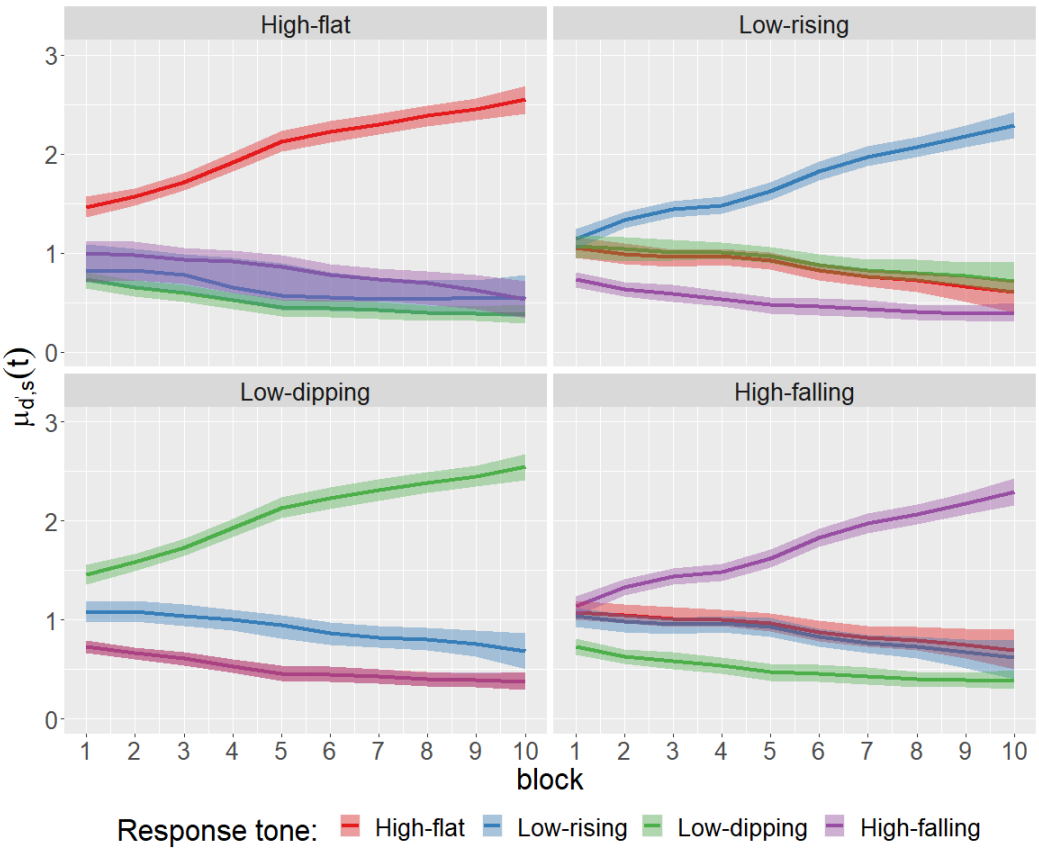

Figure S.4: Results for the benchmark data: Estimated posterior mean trajectories of the population level drifts $\mu_{d',s}(t)$ for the proposed model. The shaded areas represent the corresponding 95% pointwise credible intervals. Parameters for the high-flat tone response category are shown in red;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

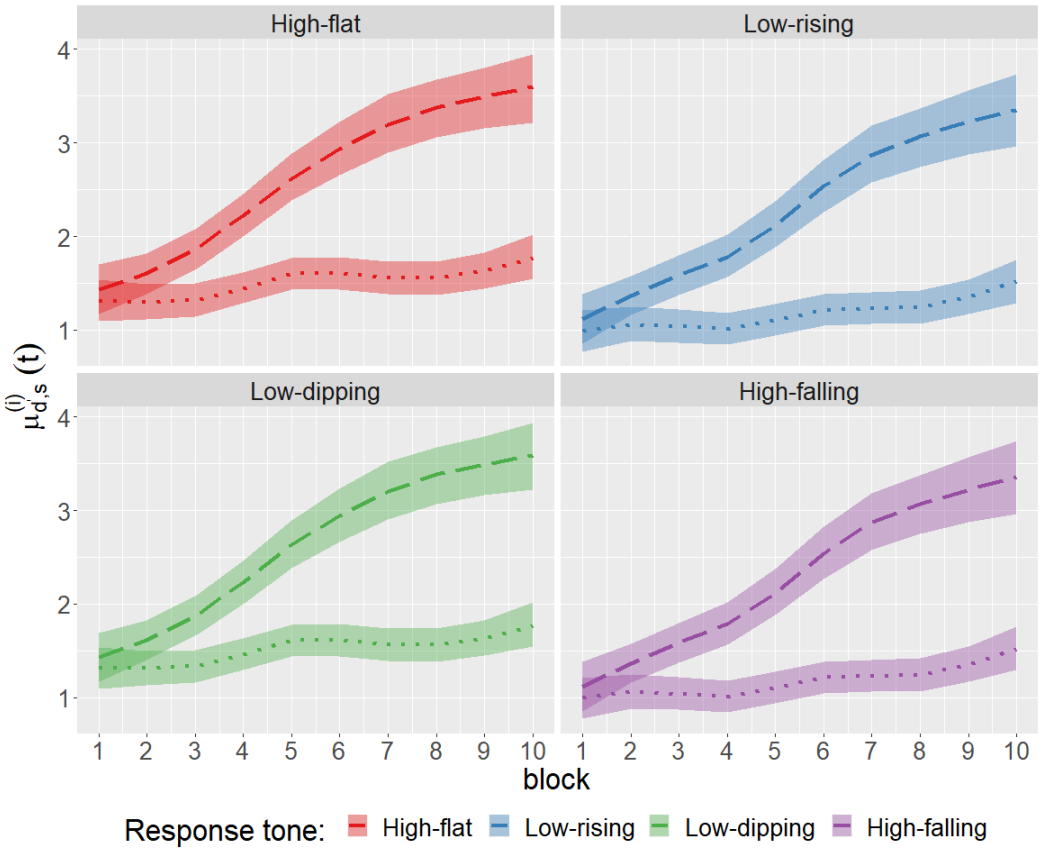

Figure S.5: Results for the benchmark data: Estimated posterior mean trajectories for individual specific drifts $\mu^{(i)}_{d',s}(t) = \exp\{f_{d',s}(t) + u^{(i)}_C(t)\}$ for correct identification ($d' = s$) for two different participants - one performing well (dashed line) and one performing poorly (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.

References

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

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

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

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

Paulon, G., Llanos, F., Chandrasekaran, B., and Sarkar, A. (2021). Bayesian semiparamet-

ric longitudinal drift-diffusion mixed models for tone learning in adults. *Journal of the American Statistical Association,*, **116**, 1114–1127.

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

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.