PNAS
www.pnas.org

Supplementary Information for

Dopamine firing plays a dual role in coding reward prediction errors and signaling motivation in a working memory task.

Stefania Sarno, Manuel Beirán, Joan Falcó-Roget, Gabriel Diaz-deLeon, Román Rossi-Pool, Ranulfo Romo and Néstor Parga

Email: unabuonaora@gmail.com (S.S.); ranulfo.romo@gmail.com (R.R.); nestor.parga@uam.es (N.P.)

This PDF file includes:

Supplementary Information Text

Methods

Discrimination Task

This study was performed on two male monkeys, Macaca mulatta, 5–7 kg. The sensory discrimination task used here has been described previously (1, 2). The schematic representation of this task is depicted in Fig. 1A. The monkey sat on a primate chair with its head fixed. The right hand was restricted through a half-cast and kept in palm-up position. The left hand operated an immovable key (elbow at ~90°) and two push buttons in front of the animal, 25 cm away from the shoulder at eye level. The centers of the switches were located 7 and 10.5 cm to the left of the midsagittal plane. In all trials, the monkey first placed the left hand on the key, and later projected to one of the two switches. Trials began when the mechanical stimulator is lowered, indenting the fingertip of one digit of the restrained hand (Probe Down, PD). The monkey places its free hand on an immovable key (Key Down, KD). The time lag between PD and KD constitutes the reaction time (RT, Fig. 1A). After the KD, a variable delay of 1.5-3s is presented to avoid anticipatory activity before the arrival of the stimulus, followed by the first stimulus (f_1), lasting 0.5s. The second stimulus (f_2) is presented after a 3s delay, also lasting 0.5s. The offset of f_2 signals the monkey to release the key (Key Up, KU), and report its decision by pressing one of two push buttons (PB) with the left hand (lateral push button for $f_2>f_1$, medial push button for $f_2<f_1$). Immediately after the decision report, correct discriminations were rewarded with a few drops of liquid, while incorrect discriminations received a few seconds of delay before the beginning of the next trial. Stimuli were delivered to the skin of the distal segment of one digit of the restrained right hand, via a computer-controlled stimulator (BME Systems; 2 mm round tip). Initial probe indentation was 500 μm. Vibrotactile stimuli were mechanical sinusoids pulses lasting 20ms each. Stimulation amplitudes were adjusted to produce equal subjective intensities (2). Performance was quantified through psychometric techniques (Fig. 1B, D). Animals were handled in accordance with standards of the National Institutes of Health and Society for Neuroscience. All protocols were approved by the Institutional Animal Care and Use Committee of the Instituto de Fisiología Celular (UNAM).

Recordings

Recordings were obtained with quartz-coated platinum-tungsten microelectrodes (2 to 3 MΩ; Thomas Recording) inserted through a recording chamber located over the central sulcus, parallel to the midline. Midbrain dopamine neurons were recorded in and around the substantia nigra, similar to other studies in monkeys (3, 4). DA neurons were identified on the basis of their characteristic regular and low tonic firing rates (1-10 spikes per second) and by their long extracellular spike potential (2.4ms ± 0.4 SD). We furthermore verified that the 22 cells used for the study did show a positive activation to reward delivery in correct (rewarded) trials and with a pause in error (unrewarded) trials. A similar criterion has been adopted in many electrophysiological studies of midbrain DA neurons (3).

Analysis of behavioral data

Animals performed the task for multiple sessions composed of about 120 trials. Behavioral data were obtained on average from 2226 trials per stimulus class (Fig. 1D). To classify trials according to the RT we defined short-RT trials as trials with RT below the median and long-RT trials those with RT above the median (Fig. 1E-F).

Analysis of the firing rate activity

The responses (z-scores) to f_1 and f_2 in Fig. 2A-B were standardized with respect to a temporal window preceding the onset of the base stimulus that lasted 500ms and was centered 1000ms after KD. To estimate the temporal profile of the z-score (Fig. 4, Fig. 5A-B left and center, Fig. 5C, and Fig. 6A) we calculated the firing rate for each neuron in 250ms sliding windows shifted every 50ms and standardized it as in Fig. 2A-B. Finally, the responses (z-scores) during the delay period (Fig. 5A-B right) were measured during its entire duration (3s) and standardized as in Fig. 2A-B.

Latency values

The firing rate time-course of the responses to f_2 depended on trial uncertainty and trial outcome (Fig. 4). To determine the time of divergence between the two time-courses, we applied a receiver operating characteristic (ROC) curve analysis in each sliding temporal window. This was done within the period lasting from 300ms before f_2 onset to 200ms after its offset. For each neuron we obtained the normalized firing rate (z-score) in sliding windows of 250ms shifted in 10ms steps. We used the z-scores of all neurons and trials to calculate the ROC curve at each time bin. The area under the ROC curve (AUROC) was used as the index indicating differential neuronal activity across different trial types. Values of the AUROC higher or lower than 0.5 indicated that, at the population level, one type of trial evoked a higher or lower DA response than the other. To determine the statistical significance of the computed AUROCs, we used a permutation test with 1000 resamples. Significance was determined with $p<0.05$ in 5 consecutive windows and the latency was defined as the first window that met this criterion. A similar analysis was used to determine significant differences in the temporal profile of the normalized activity for trials sorted according to the RT.

Dependence of the DA activity on f_1 and class number

In order to search for f_1-dependent activity we performed two different tests. We first used a linear and a sigmoidal regression analysis (Fig. 6E) to assess whether the activity was monotonic with respect to f_1. Then, we used a one-way analysis of variance (ANOVA) test to identify any general, non-monotonic relationship between firing rate activity and f_1. We focused both analyses on the f_1-stimulation period and WM delay between f_1 and f_2. We calculated a mean time-dependent z-score (standardized as in Fig. 2A-B) using a sliding window of 250ms moving in steps of 10ms, from 0.5s before f_1 onset up to 0.5s after f_2 offset. Window times with a significant monotonic signal (slope different from zero, $p<0.01$, for either a linear or sigmoidal fit with $Q>0.05$) were marked as "significantly monotonic." Window times where the ANOVA was significant ($p<0.05$) were marked as "significantly dependent" (1, 5). We then divided the 5s period (from 0.5s before f_1 onset up to 0.5s after f_2 offset) into 10 non-overlapping intervals of 500ms and counted the number of windows that were significantly linear in each interval. For each interval, we said that the f_1 dependence was significantly linear if more than the 40% of windows were significantly linear, and significantly dependent if more than 40% of windows gave a significant ANOVA p-value. Fig. S3A shows the temporal evolution of the p-value resulting from these multiple ANOVA tests. Fig. S3B shows the z-scores in windows in which the dependence was significant. A similar procedure based on the ANOVA test was employed to calculate how the responses depended on the class number. The temporal evolution of the p-value resulting from the ANOVA tests and the z-scores in windows with significant dependence are shown in Fig. S3C-D, respectively.

Correlations between DA and RT

To obtain the correlations between RTs and DA activity (z-scores) in Fig. 6B-C we employed data from all correct trials independently of the class presented. The z-scores were obtained as in Fig. 2A-B with specific temporal windows for each analyzed event. 250ms after cue presentation (PD), the z-score was obtained using a window of 300ms of width. Similarly, correlations at f_1 presentation were computed using a window of 480ms of width centered at 240ms after the onset of the stimulus. The z-score used to obtain the correlations during the delay period (Fig. 6C left) was in a window centered 2.5 seconds after the first stimulus offset and of 1 second width. The correlation coefficient between RT and DA was calculated using the MATLAB function "corrcoef". To verify the significance of the correlations, we performed a permutation test (Fig. 6B-C left) using 10000 randomly shuffled samples. Corrcoef function also provides a p-value for the coefficient. This p-value coincided up to the second decimal point with the one obtained with the permutation test. During the 3s delay period, the temporal evolution of correlations (Fig. 6C right) was studied by obtaining the z-score in 17 windows of width 300ms evenly spaced between 0.5 and 3 seconds after the offset of f_1. Correlations, which became significant towards the end of the delay, were consistently negative regardless of the number of windows as well as the width of them. Significance was assessed using corrcoef function in MATLAB.

Bayesian model for the discrimination task

The discrimination task was modeled using a Bayesian framework. The prior probabilities of f_1 and f_2 were taken to be uniform and denoted as $P(f_1^i)$ (i = 1, …, 6) and $P(f_2^j)$ (j = 1, …, 10), respectively. It was assumed that the animal knew the class structure used in the experiment (Fig. 1B-C), but it had access only to noisy representations (observations) of the two frequencies presented in the trial (denoted by $f_{1,0}$ and $f_{2,0}$). At the onset of the second stimulus, an observation o_2 of the frequency $f_{2,0}$ was obtained from a Gaussian distribution with mean $f_{2,0}$ and standard deviation σ_2. This noisy information was combined with knowledge of the prior distribution $P(f_2^j)$ to obtain the belief, or posterior distribution about the value of the second frequency, $b_2(f_2^j) = P(f_2^j|o_2) \propto P(o_2|\, f_2^j)\, P(f_2^j)$. The observation of the first frequency $f_{1,0}$, made at the end of the delay period, had to be retrieved from working memory and was indicated by o_1^*. This observation was also taken from a Gaussian distribution, but with mean $f_{1,0}$ and standard deviation σ_1. The belief about the value of the first frequency was denoted by $b_1(f_1^i) = P(f_1^i|\, o_1^*)$.

The belief state $B(k|\, o_1^*, o_2)$ about the class $c_k = (f_1^k, f_2^k)$ (k=1, …, 12; class labels, Fig. 1C) was defined as the set of the posterior probabilities $P(f_1 = f_1^k, f_2 = f_2^k \mid o_1^*, o_2)$ that the class c_k had been presented in this trial, conditioned to the observations o_2 and o_1^*. It can be written as:

$$B(k|o_1{}^*, o_2) = \frac{P(o_2)}{P(o_2|o_1*)} \frac{P(f_2=f_2^k|f_1=f_1^k)}{P(f_2^k)} b(f_1^k) b(f_2^k) \qquad (1)$$

The first factor in the above equation is a normalization factor. The second factor is a transition matrix relating the first and the second stimulation frequencies divided by the prior probability of the second stimulation frequency. Since we assumed that the animal had perfect knowledge of the class structure, the only non-zero matrix elements correspond to the 12 classes c_k of the experiment. Furthermore, since f_2 can only take two possible values for a given value of f_1, all non-zero transition probabilities are 0.5. The last two factors then become the beliefs $b_1(f_1^k)$ and $b_2(f_2^k)$ about the first and second frequencies being those in class c_k.

The sums of the $B(k|o_1^*,o_2)$'s over classes k for each choice, gives the two belief values ($f_1>f_2$ or $f_1<f_2$). These two sums were denoted by b(H) (higher choice, $f_1>f_2$) and b(L) = 1 − b(H) (lower choice, $f_1<f_2$), and choices were made according to the larger of these two, denoted as b_c. The performance can be measured by the fraction of trials of a given class in which the decision is correct.

The two unknown model parameters, the standard deviations σ_1 and σ_2, were fitted by minimizing the mean squared error between the model performance and the animal's performance. We found the optimal parameter values: σ_1 =5.5Hz and σ_2 =3.2Hz (see next section for more details about the model fitting procedure).

The statistical uncertainty U of a given trial is U=1-max[b(H), b(L)]=1-b_c, which is bounded between 0 and 0.5. The uncertainty of a given class was defined as the average of the value U across trials of each class. The uncertainty in hits (or errors) of a given class was defined as the average over the correct (or wrong) trials from each class.

Bayesian model fitting procedure

Parameter fitting in Fig. 3B and S2 were made using the Simulating Annealing solver in MATLAB. For each RT condition in Fig. S2, 200 fits were performed. The two fits that yielded the lowest error were used to model the monkey's performance in Fig. 1. Given that 200 adjustments were performed, a distribution for each of the two parameters was available. The mean value for σ_1 obtained in the short-RT condition was found to be significantly lower than the noise parameter for the long-RT group (one-tailed two sample t-test; $p<0.001$). The same result was found for the noise parameter σ_2 (one-tailed two sample t-test; $p<0.001$).

Reinforcement Learning Model based on Belief States

To test whether the phasic responses can be attributed to dopamine reward prediction errors (RPEs) we constructed a reinforcement learning model and checked if it was able to reproduce the observed responses. Given that in the task the relevant stimuli are only partially observable (the animal is not aware of the true value of f_1 and f_2) we used a belief-state temporal difference (TD) model (similar to that proposed by (6)) to compute reward expectations and simulate the RPEs signaling. This was first implemented in a Bayesian module that works similarly to the Bayesian model described above and uses the same fitted values of the two noise parameters, σ_1 and σ_2. This module yields the belief $b_1(f_1^i)$ (i=1, …, 6) about the value of the first frequency, the belief state $B(k \mid o_1^*, o_2)$ about the class $c_k = (f_1^k, f_2^k)$ (k=1, …, 12) at the time when the second frequency is presented and the belief b_c = [b(H), 1-b(H)] about which of the two frequencies is higher. Then it transmits these inference results to a TD module that selects actions and generates reward prediction errors (RPEs).

The RL model also uses a fully observable variable *pm* (push movement) that represents the movement towards one of the two buttons. This variable has two possible states: *py* ("push yes") when the decision is $f_1>f_2$ and *pn* ("push no") when the decision is $f_1<f_2$. Finally, the reward function, denoted by r, is a scalar function that takes two different values for correct and incorrect decisions (see Equation 8).

At the beginning of each trial the TD module calculates the value of the first stimulus as:

$$V_1(b_1) = \sum_{i=1}^{6} Q_1(i) \cdot b_1(i) \qquad (2)$$

where Q_1(i) is a set of adaptable weights. The RPE at the first stimulus is $\delta(f_1) = V_1(b_1)$.

At the second stimulus the TD module computes the value of the class:

$$V_B(B) = \sum_{k=1}^{12} Q_B(k) \cdot B(k|o_1{}^*, o_2). \quad (3)$$

Here the Q_B(k) (k=1, …,12) are another set of adaptable weights. At the second stimulus the TD module also estimates the value resulting from the comparison of the two frequencies

$$V(b_c) = \sum_i^G g_i(b_c) * \nu_i. \quad (4)$$

where the ν_i are a set of adaptable weights. The g_i (i=1, …, G) are convenient functions that account for the contribution of the belief b_c to this value. The functions g_i were taken as the following basis functions:

$$g_i(b_c) = \left(\frac{1}{2}\right)[cos(\mathrm{a}\ (b_c - \mathrm{c}_i)) + 1], \quad (5)$$

if $\mathrm{c}_i - 0.1 < b_c < \mathrm{c}_i + 0.1$ and $g_i(b_c) = 0$ otherwise. The c_i's are G=11 equally spaced centroids in [0,1] and a = π /0.1.

Given the values $V_B(B)$ and $V(b_c)$ the RPE at the second stimulus is:

$$\delta(f_2) = V_B(B) + V(b_c) - V_1(b_1). \quad (6)$$

As in the Bayesian model, we assumed that decisions were made according to the larger of the beliefs b(H) and b(L) about which of the two frequencies was the higher. The value of the response movement when action j is selected is indicated with V_{rm}(j) to highlight the correspondence between the action selected and the subsequent movement.

The RPE at the response movement is:

$$\delta(rm) = V_{rm}(j) - V_B(B) - V(b_c) \quad (7)$$

and the RPE at the delivery of reward is:

$$\delta(r) = r - V_{rm}(j), \quad (8)$$

where r =1 for correct discrimination and r = -0.5 otherwise.

The RPEs $\delta(f_2)$, $\delta(rm)$ and $\delta(r)$ are used to update the adaptable weights Q_1(i), Q_B(k), ν_i, the value of the movement V_{rm}(j).

We assumed a discount factor $\gamma = 1$ and used the TD(λ) algorithm to update the weights at the end of each trial. The updating rule is the following:

$$Q_1(i) = Q_1(i) + \alpha \cdot \left(\lambda^{\Delta_{12}} \cdot \delta(f_2) + \lambda^{\Delta_{1rm}} \cdot \delta(rm) + \lambda^{\Delta_{1r}} \cdot \delta(r)\right) \cdot b_1(i) \quad (9)$$

$$Q_B(k) = Q_B(k) + \alpha \cdot \left(\lambda^{\Delta_{2rm}} \cdot \delta(rm) + \lambda^{\Delta_{2r}} \cdot \delta(r)\right) \cdot B(k) \quad (10)$$

$$V_{rm}(j) = V_{rm}(j) + \alpha \cdot \left(\lambda^{\Delta_{rmr}} \cdot \delta(r)\right) \quad (11)$$

In all the equations above α represents the learning rate. The value of Δ is the temporal interval between the relevant task events, i.e. the onset of f_1, the onset of f_2, the response movement and the reward. To mimic the task temporal structure we took $\Delta_{12} = 30, \Delta_{1rm} = 40, \Delta_{1r} = 45, \Delta_{2rm} = 10, \Delta_{2r} = 15, \Delta_{rmr} = 5$ (The temporal intervals are expressed in units of the time step, $dt = 0.1$ s). For the simulation, we use $\lambda = 0.95$ and $\alpha = 0.05$.

REFERENCES

1. R. Romo, C. D. Brody, A. Hernández, L. Lemus, Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).

2. R. Romo, R. Rossi-Pool, Turning Touch into Perception. *Neuron* **105**, 16–33 (2020).

3. E. S. Bromberg-Martin, M. Matsumoto, O. Hikosaka, Distinct Tonic and Phasic Anticipatory Activity in Lateral Habenula and Dopamine Neurons. *Neuron* **67**, 144–155 (2010).

4. M. Matsumoto, O. Hikosaka, Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837–841 (2009).

5. C. D. Brody, A. Hernández, A. Zainos, R. Romo, Timing and Neural Encoding of Somatosensory Parametric Working Memory in Macaque Prefrontal Cortex. *Cereb. Cortex* **13**, 1196–1207 (2003).

6. R. P. N. Rao, Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Front. Comput. Neurosci.* **4**, 1–18 (2010).

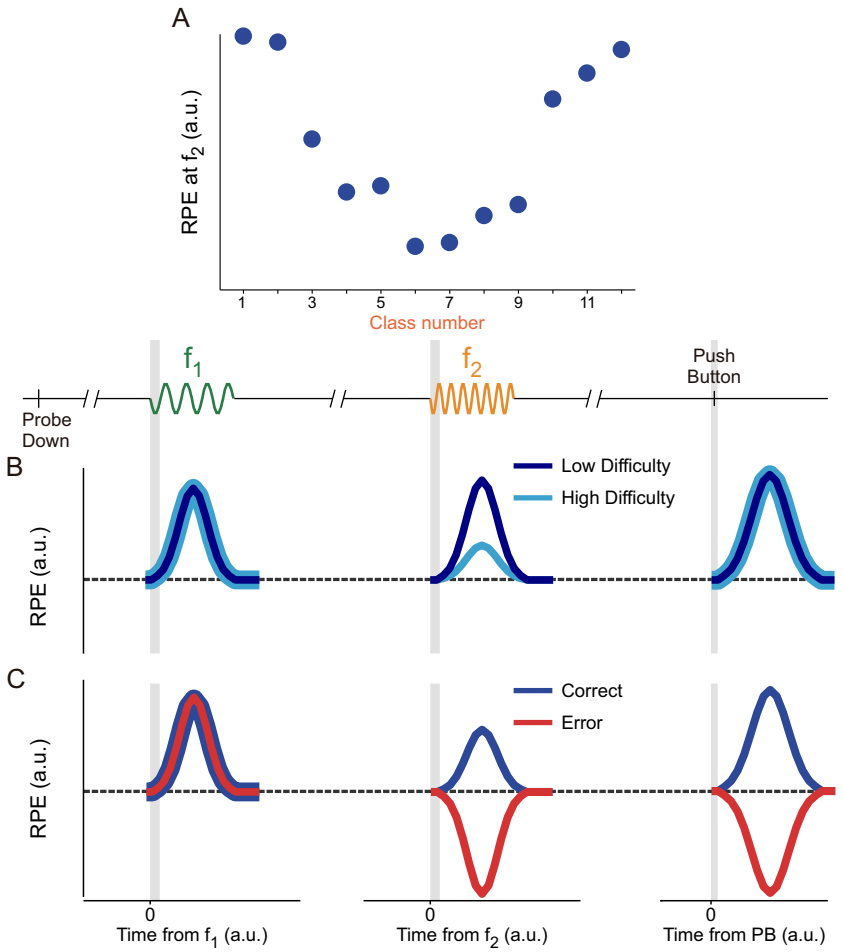

Figure S1. Reward prediction errors predicted by the reinforcement learning model are similar to DA phasic responses (related to Figures 2 and 4). (A) RPE as a function of class number in correct trials, taken at the onset of f_2. A Gaussian filter was applied to the model predictions. Noise parameters σ_1=5.50 Hz for f_1 sampling and σ_2=3.2 Hz for f_2 sampling. (B) RPE signal calculated during 3 different periods with emulated data: f_1 presentation, f_2 presentation, and reward delivery after Push Button (PB, decision report). Dark blue curve represents the simulated low difficulty group. Light blue curve represents the simulated high difficulty group. Left: RPE after the onset of f_1 does not depend on the difficulty level, observable in the overlap between both of our difficulty group curves. Center: RPE after the onset of f_2 depends on choice difficulty, favoring the low difficulty group. The high difficulty group (light blue) reaches less than half the max value observed for the low difficulty group. Right: RPE after reward delivery peaks at a value independent of the difficulty level since the curves overlap perfectly. (C) RPE calculated during 3 different periods for simulated correct trials (blue curve) and simulated error trials (red curve). Left: RPE after the onset of f_1 is similar in correct and error trials, since the two curves overlap. Center: RPE after f_2 onset shows activation in correct trials and a strong depression in error trials (red curve). Right: RPE generated by the model after PB. Correct trials show an increase in RPE, while the error trials show a depression in RPE. The amount of change between error and correct trials is approximately equivalent.

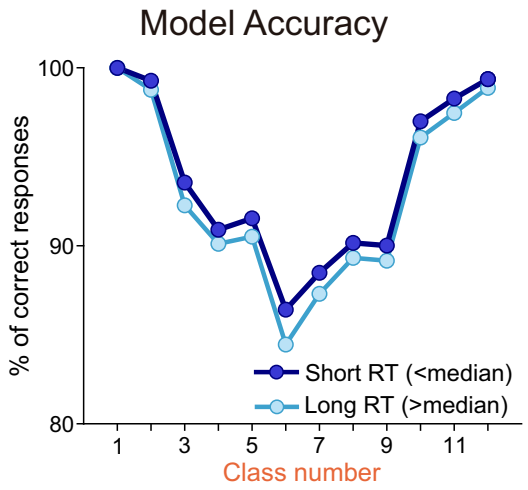

Figure S2. Percentage of correct responses as a function of class number obtained with the Bayesian model for short- and long-RT trials (light and dark blue circles and line, respectively; related to Fig. 1F, left). Model best-fit parameter values were σ_1=5.28 Hz and σ_2=3.01 Hz for the short-RT group, and σ_1=5.385 Hz and σ_2=3.0 Hz for the long-RT group.

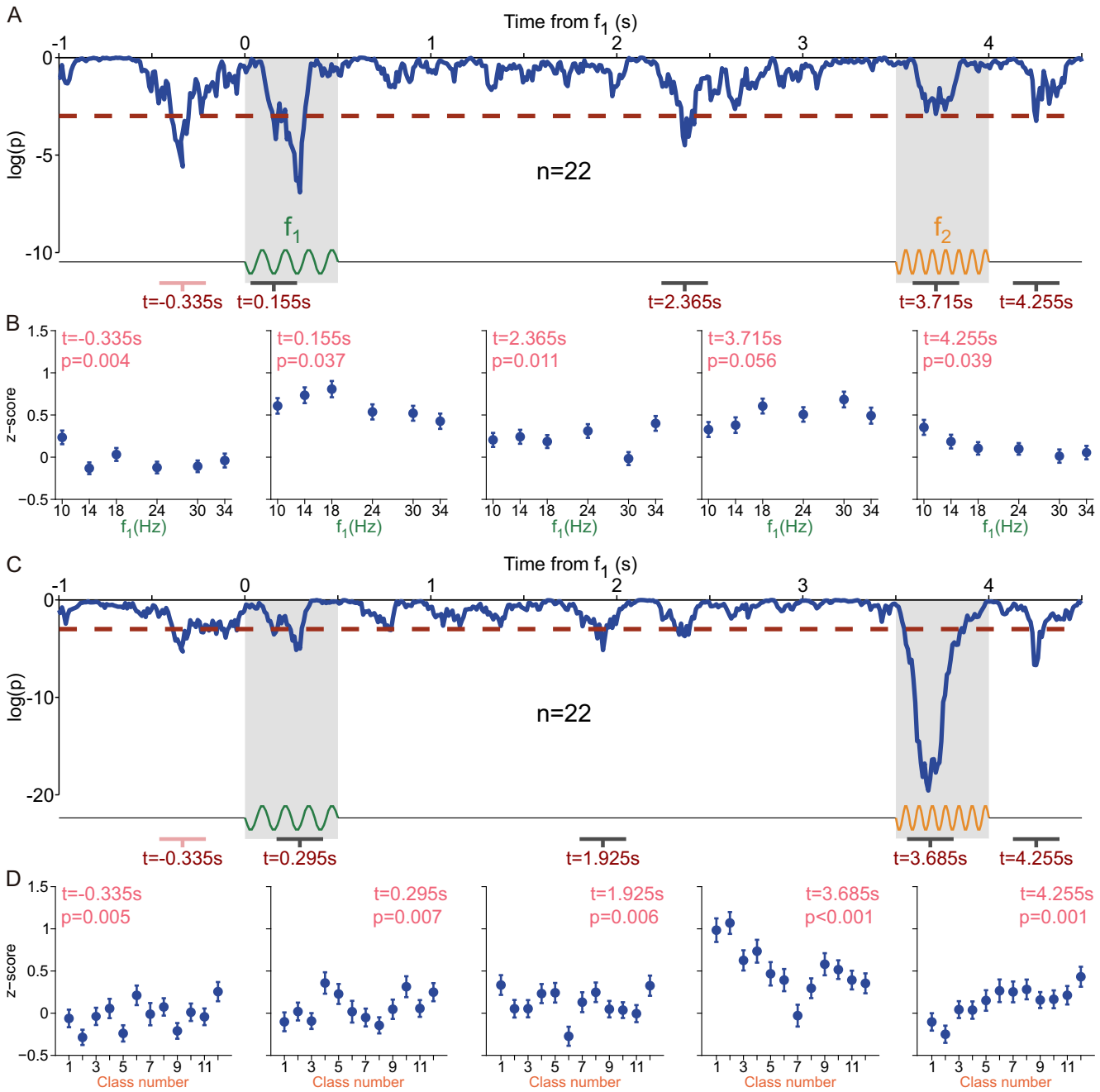

Figure S3. Analysis of the existence of general dependencies on f_1 and on class number during the trial (related to Fig. 5E). (A) Temporal evolution of the logarithm of p-value (log(p)) resulting from multiple ANOVA tests performed to check the dependence of the z-score on the f_1 values. We calculated a mean time-dependent z-score (standardized as in Fig. 2A-B) using a sliding window of 250ms shifting in steps of 10ms and performed multiple ANOVA tests sorting the z-score according to the values of f_1. Values below the red dotted lines are considered as significant (significance is assessed as $p<0.05$). The p-value intermittently crossed the significance threshold. However, it consistently remained significant only during the presentation of the first stimulus. (B) Average z-score for each f_1 stimulus (error bars for ±1 SEM) calculated in each of the 5 indicated regions from panel (A). First graphic for the basal period before f_1 onset, proceeding in increasing chronological order from left to right. Times in the top left corner (pink) and p values (pink) are the results of the ANOVA tests and indicate dependencies on f_1 when below threshold (red dotted line in panel A). (C) Similar to panel (A) but the p-value was obtained by sorting the z-score according to the class number and running multiple ANOVA tests. (D) Similar to panel (B) but averaging the z-score according to the class number.