frontiers in
HUMAN NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 08 February 2013
doi: 10.3389/fnhum.2012.00359

A model-based approach to trial-by-trial P300 amplitude fluctuations

***Antonio Kolossa*[1], *Tim Fingscheidt*[1]*, *Karl Wessel*[2,3] and *Bruno Kopp*[2,4]**

[1] *Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany*
[2] *Cognitive Neurology, Technische Universität Braunschweig, Braunschweig, Germany*
[3] *Department of Neurology, Braunschweig Hospital, Braunschweig, Germany*
[4] *Department of Neurology, Hannover Medical School, Hannover, Germany*

Edited by:
Sven Bestmann, University College London, UK

Reviewed by:
Francisco Barcelo, University of Illes Balears, Spain
Simon P. Kelly, City College of New York, USA

****Correspondence:***
Tim Fingscheidt, Institute for Communications Technology, Technische Universität Braunschweig, Schleinitzstraße 22, 38106 Braunschweig, Germany.
e-mail: t.fingscheidt@tu-bs.de

It has long been recognized that the amplitude of the P300 component of event-related brain potentials is sensitive to the degree to which eliciting stimuli are surprising to the observers (Donchin, 1981). While Squires et al. (1976) showed and modeled dependencies of P300 amplitudes from observed stimuli on various time scales, Mars et al. (2008) proposed a computational model keeping track of stimulus probabilities on a long-term time scale. We suggest here a computational model which integrates prior information with short-term, long-term, and alternation-based experiential influences on P300 amplitude fluctuations. To evaluate the new model, we measured trial-by-trial P300 amplitude fluctuations in a simple two-choice response time task, and tested the computational models of trial-by-trial P300 amplitudes using Bayesian model evaluation. The results reveal that the new digital filtering (DIF) model provides a superior account of the trial-by-trial P300 amplitudes when compared to both Squires et al.'s (1976) model, and Mars et al.'s (2008) model. We show that the P300-generating system can be described as two parallel first-order infinite impulse response (IIR) low-pass filters and an additional fourth-order finite impulse response (FIR) high-pass filter. Implications of the acquired data are discussed with regard to the neurobiological distinction between short-term, long-term, and working memory as well as from the point of view of predictive coding models and Bayesian learning theories of cortical function.

Keywords: predictive surprise, Bayesian surprise, event-related brain potentials, P300, single trial EEG, digital filtering

1. INTRODUCTION

The notion of a Bayesian brain is increasingly recognized as providing a distinctive framework for investigating cognitive brain functions (Kersten et al., 2004; Knill and Pouget, 2004; Friston, 2005; Doya et al., 2007). Predictive coding theories of cortical function provide a possible route to the Bayesian brain (Friston, 2002). According to the predictive coding approach, constraints from higher levels of a cortical hierarchy provide contextual guidance to lower levels of processing, providing a theory of how bottom-up evidence is combined with top-down priors to compute the most likely interpretation of sensory data. Specifically, predictive coding theory proposes that an internal representation of the world generates predictions that are compared with stimulus-driven activity to calculate the residual error between the predicted and the actual information. The residual error is then used to update the internal representation so as to minimize the residual error imposed by future stimuli (Friston, 2002, 2005; Spratling, 2010).

The general scheme of predictive coding as a ubiquitous mode of cortical processing offers an instrumental framework for analyzing functional correlates of the P300 event-related brain potential (Sutton et al., 1965; Kopp, 2008). It has long been recognized that fluctuations in P300 amplitude reflect the degree of surprise related to the processing of attended, but unforeseeable sensory events. In particular (Donchin, 1981) argued that P300 amplitude is not crucially determined by the inherent attributes of eliciting events. Instead of that, he ascertained that "surprising events elicit a large P300 component" (p. 498). Squires et al. (1976) had presented a model of P300 amplitude fluctuations, based on the concept of expectancy, which was thought to be determined by three factors: "(i) the memory for event frequency within the prior stimulus sequence, (ii) the specific structure of the prior sequence, and (iii) the global probability of the event" (p. 1144).

More recently, Mars et al. (2008) proposed a computational model of processes underlying the generation of the P300 in which trial-by-trial fluctuations in P300 amplitudes were explained in terms of a Bayesian observer keeping track of the global probabilities of sensory events. The subjective estimates of statistical regularities in the environment were thought to depend crucially on the integration of sensory data over long periods of time. However, the adequacy of this Bayesian observer model is limited, because it cannot account appropriately for the well-documented effects of the recent stimulus sequence on P300 amplitudes (e.g., Squires et al., 1976; Leuthold and Sommer, 1993).

Here we tested these two state-of-the-art models against a newly developed computational model of trial-by-trial P300 amplitude fluctuations by Bayesian model selection (Kass and Raftery, 1995; Raftery, 1995). The new model assumes three additive digital filtering processes, thereby integrating aspects of both state-of-the-art models. Specifically, subjective estimates of statistical regularities in sensory data are kept at short-term and long-term decay time parameters. Further it implements an alternation term (as Squires et al., 1976) as well as uniform initial prior probabilities (as Mars et al., 2008). Our findings show that this new approach provides a superior account of parietally distributed trial-by-trial P300 amplitudes compared to these two state-of-the-art models.

2. MATERIALS AND METHODS

2.1. PARTICIPANTS, EXPERIMENTAL DESIGN, AND DATA ACQUISITION

Sixteen healthy participants [fourteen women, mean age: 20 years; age range 18–23 years; mean handedness (Oldfield, 1971): 74; handedness range −76–100], all with normal or corrected-to-normal visual acuity participated in the experiment. All were recruited from introductory courses at the Department of Psychology at the Technische Universität Braunschweig in return for course credit. Experimental procedures were approved by the local ethics committee and in accordance with the Declaration of Helsinki.

Participants performed a simple two-choice response time [RT] task without feedback about response accuracy in which all stimuli had equal behavioral relevance. This feature of the experimental design constitutes an important difference between this and the classical oddball paradigm (Ritter and Vaughan, 1969) in which participants usually discriminate between task-relevant (target) and irrelevant (standard) stimuli.

The experiment was realized using the Presentation® software (Neurobehavioral Systems, Albany, CA, USA). Visual stimuli were presented one at a time for 100 ms each, with a stimulus presentation rate of $f_s = 2/3$ Hz, i.e., one stimulus per 1.5 s. Stimuli were displayed at the center of a CRT monitor (FlexScan T766 19″; Eizo, Hakusan, Ishikawa, Japan) with a refresh rate of 100 Hz at a resolution of 1280 × 1024 pixels against a light gray background. Viewing distance amounted to 1.25 m. Two types of visual stimuli were presented: the stimulus event was either a red or a blue rectangle, each of which subtended approximately 2.75° × 2.25°.

Participants were required to respond to each stimulus with the previously associated button as quickly as possible but not at the expense of accuracy. They used the index finger of both hands (e.g., left button on response to the red rectangle, right button in response to the blue rectangle). Stimulus-response mapping (i.e., [red-left, blue-right] or [red-right, blue-left], respectively) was counterbalanced over participants.

Participants performed twelve blocks of $N = 192$ trials of the two-choice RT task. The probability of the occurrence of each stimulus event was manipulated between blocks such that the relative probabilities of events were either 0.5 for each event, across six consecutive blocks (1152 trials overall), or [0.3, 0.7], across the remaining six consecutive blocks (1152 trials overall). Stimulus-probability mapping was counterbalanced over participants (i.e., a stimulus color identified the rare (0.3) stimulus in fifty percent of the participants but the frequent stimulus (0.7) in the remaining participants).

The order of the probability manipulation was counterbalanced over participants (probability category [0.5, 0.5] prior to [0.3, 0.7] or vice versa) *who were not informed about these probabilities.* Participants were informed that the two different stimuli were randomly distributed across blocks. Between the blocks a break was scheduled, participants were free to initiate the subsequent block at their own pace.

A continuous electroencephalogram (EEG) was recorded using a QuickAmps-72 amplifier (Brain Products, Gilching, Germany) and the BrainVision Recorder® Version 1.02 software (Brain Products, Gilching, Germany) from frontal (F7, F3, Fz, F4, F8), central (T7, C3, Cz, C4, T8), parietal (P7, P3, Pz, P4, P8), occipital (O1, O2), and mastoid (M1, M2) sites. Ag-AgCl EEG electrodes were used which were mounted on an EasyCap (EasyCap, Herrsching-Breitbrunn, Germany). Electrode impedance was kept below 10 kΩ. All EEG electrodes were referenced to average reference during the recording.

For each participant, the actual stimulus sequence of each probability category [0.5, 0.5], and [0.3, 0.7], respectively, was randomized only once in order to enhance the reliability of the sequential trial-by-trial P300 estimates (see below). Thus, each participant received solely one truly random arrangement of trials in each probability category. This arrangement was repeatedly presented across all six blocks of each probability category, unbeknownst to participants. In consequence, sequential P300 estimates could be averaged over the six sequence repetitions per probability category, thereby improving the notoriously low signal-to-noise ratio of single-trial EEG data. Task-related brain activity of a single trial is much more obscured by task-unrelated brain activity than is task-related activity averaged across trials (Blankertz et al., 2002).

Participants were informed about the problem of non-cerebral artifacts, and they were encouraged to reduce the occurrence of movement artifacts (Picton et al., 2000). Ocular artifacts were monitored by means of bipolar pairs of electrodes positioned at the sub and supraorbital ridges (vertical electrooculogram, vEOG) and at the external ocular canthi (horizontal electrooculogram, hEOG). The EEG and EOG channels were subject to a bandpass of 0.01–30 Hz and digitized at 250 Hz sampling rate.

Off-line analysis of the EEG data was performed by means of the BrainVision Analyzer® Version 2.0.1 software (Brain Products, Gilching, Germany). Careful manual artifact rejection was performed before averaging to discard trials during which eye movements, or any other non-cerebral artifact except blinks, had occurred. Deflections in the averaged EOG waveforms were small indicating that fixation was well maintained in those trials that survived the manual artifact rejection process. Semi-automatic blink detection and the application of an established method for blink artifact removal were employed for blink correction (Gratton et al., 1983). A digital high-pass filter was applied to the data (0.75 Hz cutoff frequency, 48 db/oct) in order to eliminate low-frequency variations in the EEG signal which were associated with the occasional occurrence of electro-dermal artifacts.

The EEG was then divided into epochs of 1000 ms duration, starting 100 ms before stimulus onset. Epochs were corrected using the interval [−100, 0 ms] before stimulus presentation as the baseline. As a start, event-related potential (ERP) waveforms were created (Luck, 2005). ERP waveforms were calculated as trial averages for each participant and for each event probability [i.e., 0.5, 0.3, 0.7], with the exception that those trials in which the participant selected the wrong behavioral response were excluded from averaging.

Thereafter, trial-by-trial P300s were estimated from the EEG data at electrode Pz, where this ERP component is traditionally reported to be maximal (Duncan-Johnson and Donchin, 1977). To estimate trial-by-trial P300 amplitudes, for each participant, the time point at which the averaged P300 waveforms at Pz were modulated maximally by relative stimulus frequency in the [0.3, 0.7] probability category was determined ($M = 344$ ms, $SD = 48$ ms; range 280–464 ms). Identifying the P300 in single trials is a notoriously difficult problem, due to the low signal-to-noise ratio of single-trial EEG data (Blankertz et al., 2002). In our study, for each event probability, trial-by-trial P300 estimates were extracted over a temporal window of ±60 ms around the individual time point of maximal modulation (Barceló et al., 2008), thereby completely ignoring latency variability across single trials (Luck, 2005). Albeit this drawback of the method, it was nevertheless chosen in order (1) to keep the testing environment as similar as possible to the procedures employed by Mars et al. (2008), and (2) to improve the reliability of trial-by-trial amplitude measures, in comparison to peak detection measures, akin to previous studies (Debener et al., 2005).

2.2. CONVENTIONAL DATA ANALYSIS

Trial-by-trial P300 estimates, RTs, and error rates were averaged according to the three event probabilities [i.e., 0.3, 0.5, 0.7]. In the [0.5, 0.5] probability category, trial-by-trial P300 estimates were additionally averaged according to eight third-order stimulus sequences (denoted as *aaaa*, *baaa*, *abaa*, *aaba*, *bbaa*, *abba*, *baba*, *bbba*), four second-order stimulus sequences (*aaa*, *baa*, *aba*, *bba*), and two first-order sequences (*aa*, *ba*), with up to four consecutive trials ($xxxx$) = (trial $n-3$, trial $n-2$, trial $n-1$, trial n = eliciting event). Please note that the symbol *a* simply denotes one of the two possible stimulus events while symbol *b* signifies the other one in this notation. For example, if *a* signifies the red rectangle, then *b* signifies the blue rectangle (and vice versa). In the [0.5, 0.5] probability category, sequential analysis could be collapsed across the two possible stimulus events since both stimuli were equally probable and task-relevant. The same kind of sequential analysis was performed in the [0.3, 0.7] probability category. However, in this experimental condition, *a* consistently denoted the rare stimulus in the [0.3, 0.7] probability category, whereas *b* signified the frequent stimulus in the [0.3, 0.7] probability category. We refrained from analyzing the eight third-order stimulus sequences in the [0.3, 0.7] probability category in order to obtain a sufficient number of trials entering the sequential P300 estimates. Thus, trial-by-trial P300 estimates were solely averaged according to eight second-order stimulus sequences (*aaa*, *baa*, *aba*, *bba*, *bbb*, *abb*, *bab*, *aab*) and four first-order sequences (*aa*, *ba*, *bb*, *ab*) in the [0.3, 0.7] probability category, separately for rare and frequent stimuli.

Individual medians of trial-by-trial P300 estimates, RTs, and error rates over the three event probabilities [i.e., 0.3, 0.5, 0.7] as well as the sequential P300 estimates, generated as described above, were submitted to repeated measures analysis of variance (ANOVAs), using the Greenhouse–Geisser correction. The results of the univariate tests are provided, using a format which gives the uncorrected degrees of freedom, and ϵ in order to compensate for violations of sphericity or equal covariance among all pairs of levels of the repeated measures (Picton et al., 2000). A measure of effect size, η_p^2 (partial eta squared), is also provided.

2.3. STATE OF THE ART MODELS

Let us call

$$P_k(n) = P\left(s(n) = k|\mathrm{s}_1^{n-1}\right) \quad \text{with } k \in \{1, \ldots, K\} \qquad (1)$$

an *estimated subjective probability* (henceforth simply called *subjective probability*) that event $k \in \{1, \ldots, K\}$ on trial $n \in \{1, \ldots, N\}$ will be observed, given a sequence $\mathrm{s}_1^{n-1} = (s(1), s(2), \ldots, s(n-1))$ of $n-1$ former stimulus observations. While n is the discrete time index of the consecutive trials, the value N denotes the total number of trials in a block within an experimental probability category for one subject. Note that in (1) stimulus $s(n)$ has not yet been observed, therefore, a subjective probability distribution $P_k(n)$ for *all* possible stimuli $k \in \{1, \ldots, K\}$ on trial n is of interest. However, once the stimulus $k = s(n)$ on trial n has been observed (which is only a *single* value k out of set $\{1, \ldots, K\}$), the respective subjective probability $P_k(n)$ can be used to calculate the degree of *surprise* (Shannon and Weaver, 1948; Strange et al., 2005)

$$I(n) = -\log_2 P_{k=s(n)}(n)\,. \qquad (2)$$

Following Mars et al. (2008), we assume the trial-by-trial P300 estimate $Y(n)[\mu\mathrm{V}]$ to be proportional to the surprise $I(n)$[bit]:

$$Y(n) \propto I(n) \qquad (3)$$

Note that Squires et al. (1976) assumed direct proportionality between the so-called *expectancy* $E_k(n)$ and the trial-by-trial P300 estimate:

$$Y(n) \propto E_{k=s(n)}(n) \qquad (4)$$

In the following we briefly recapitulate these two well-known state-of-the-art approaches to compute the subjective probability $P_k(n)$ or expectancy $E_k(n)$, which play the role of a *dynamically updated prior probability* for learning statistical parameters of the stimulus sequence.

2.3.1. *Approach by Mars et al. (MAR)*

Mars et al. (2008) proposed a Bayesian observer model (henceforth called MAR) without forgetting according to

$$P_k(n) = \frac{\tilde{c}_{\mathrm{L},k}(n) + 1}{(n-1) + K}, \qquad (5)$$

where

$$\tilde{c}_{\mathrm{L},k}(n) = \sum_{\nu=1}^{n-1} d_k(\nu) \qquad (6)$$

counts the number of occurrences of event k until trial $n-1$. The time sequence $d_k(\nu)$ holds $d_k(\nu) = 1$ if $s(\nu) = k$, $\nu = 1,2, \ldots,$

otherwise $d_k(\nu) = 0$. Note that $\sum_{k=1}^{K} \tilde{c}_{\mathrm{L},k}(n) = n - 1$. As can be easily seen in (5), the subjective probability for event k on trial $n = 1$ equals a uniform initial prior $P_k(1) = 1/K$. After many trials ($n \gg K$, and $\tilde{c}_{\mathrm{L},k}(n) \gg 1$), the subjective probability approximates $P_k(n) \approx \frac{\tilde{c}_{\mathrm{L},k}(n)}{n-1}$, i.e., the relative frequency of event k until trial $n - 1$. Note that the index "L" of the count function $\tilde{c}_{\mathrm{L},k}(n)$ expresses the *long-term memory* character of Mars' model.

2.3.2. Approach by Squires et al. (SQU)

Unlike Mars et al. (2008), Squires et al. (1976) did not formulate a strict computational model to compute the subjective probability $P_k(n)$. Moreover, having investigated solely a $K = 2$ case, they use the notion of *expectancy*[1] $E_k(n)$ for stimulus k on trial n. While Squires et al. (1976) have described their model, hence called SQU, partly in math, partly in words, in the following we present a complete analytical formulation of their approach, which is straightforward to implement in software. Their empirical formulation of expectancy that event $k \in \{1, 2\}$ will be observed on trial $n \in \{1, \ldots, N\}$ is given as

$$E_k(n) = 0.505 \cdot P_k + 0.235 \cdot \check{c}_{\mathrm{S},k}(n) + 0.033 \cdot \check{c}_{\Delta,k}(n) - 0.027 \quad (7)$$

with three expectancy contributions, namely the assumed-to-be-known global probability P_k, a count function for the *short-term memory* "S", and a count function for the *alternation expectancy* "Δ" (and an additive constant).

2.3.2.1. Short-term memory. The short-term count function is defined as

$$\check{c}_{\mathrm{S},k}(n) = \sum_{\nu=n-N_{\mathrm{depth}}}^{n-1} \gamma_{\mathrm{S}}^{n-\nu} d_k(\nu), \quad (8)$$

which is different to (6), since only a limited memory span of $N_{\mathrm{depth}} = 5$ is covered, and an exponential forgetting factor $\gamma_{\mathrm{S}} = e^{-\frac{1}{\beta_{\mathrm{S}}}}$ with $0 \leq \gamma_{\mathrm{S}} \leq 1$ and time constant $0 \leq \beta_{\mathrm{S}} < \infty$ is introduced, with $\gamma_{\mathrm{S}} = 0.6$ for all probability categories (i.e., $\beta_{\mathrm{S}} = 1.96$). Note that the count function (8) depends only on stimulus observations in the recent past.

2.3.2.2. Global probability. The term $P_k = P(s(n) = k)$ in (7) denotes the true global probability of the stimulus being k. It is nothing else but the relative frequency of the stimulus in the current experimental probability category which must be made known to this model.

2.3.2.3. Alternation expectancy. In contrast, the term $\check{c}_{\Delta,k}(n) \in \{-3, -2, 0, 2, 3\}$ denotes the expectancy w.r.t. alternating stimuli, and how this expectancy is met by the present stimulus $s(n)$. The sign of $\check{c}_{\Delta,k}(n)$ is negative if the stimulus $s(n)$ violates the alternation expectation [i.e., $s(n)$ and $s(n-1)$ are identical] and positive if the alternation expectation is met [i.e., $s(n)$ and $s(n-1)$ differ from each other]. The amplitude of $\check{c}_{\Delta,k}(n)$ depends on the number of *previous* alternations *in a row*. The formulas for calculating $\check{c}_{\Delta,k}(n)$ are detailed in the Appendix.

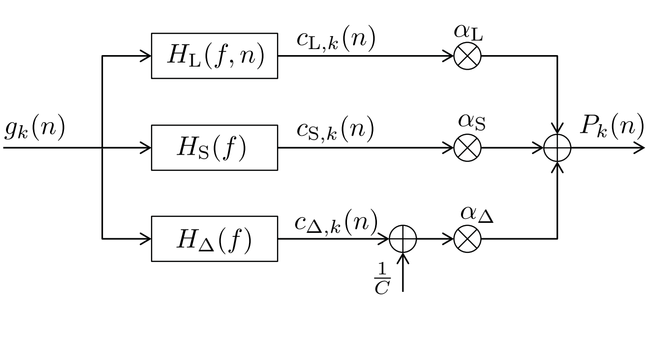


FIGURE 1 | Block diagram of the new digital filter (DIF) model with input $g_k(n)$ in (10) and output $P_k(n)$ in (9), digital filter transfer functions $H(f)$ with $0 \leq f \leq f_s/2$, a stimulus presentation rate of f_s (= 2/3 Hz), and probability normalizing constant $1/C$.

2.3.3. Explanatory notes

In summary, (7) provides a model for expectancy that is linearly composed of three contributions: Firstly, the relative frequency P_k of event k which equals the correct global probability throughout probability categories. Note that, in contrast to Mars et al. (2008), the relative frequency P_k is not learned sequentially by experience but assumed to be known by participants. Secondly, a purely predictive limited length ($N_{\mathrm{depth}} = 5$) exponentially decaying short-term memory [cf. count function $\check{c}_{\mathrm{S},k}(n)$ in (8)]. Thirdly, an expectancy contribution in the range $-3 \leq \check{c}_{\Delta,k}(n) \leq +3$ depending on the extent to which a first-order alternation (*aa* or *ba*) expectation has been build up and then met/violated within the latest observed $N_{\mathrm{depth}} = 5$ trials.

2.4. PROPOSED DIGITAL FILTER MODEL

In this section we present our newly proposed model, inspired by both Mars et al. (2008) and Squires et al. (1976). Our aim is to unify the learned relative frequency estimation of Mars et al. (2008; long-term) with the exponentially decaying short-term memory and alternation expectation capabilities of Squires et al. (1976), and to express the result in terms of a simple new *digital filter (DIF) model*. Besides an additive probability-normalizing constant $1/C$ (see Appendix for details), it consists of three additive contributions to subjective probability: a long-term contribution ("L"), a short-term one ("S"), and one term capturing alternations ("Δ") as depicted in **Figure 1**:

$$P_k(n) = \alpha_\mathrm{L} \cdot c_{\mathrm{L},k}(n) + \alpha_\mathrm{S} \cdot c_{\mathrm{S},k}(n) + \alpha_\Delta \cdot \left[c_{\Delta,k}(n) + \frac{1}{C} \right]. \quad (9)$$

There are three different count functions used, each represented by a digital filter transfer function $H(f)$ applied to the common input signal $g_k(n)$, which is given as:

$$g_k(n) = \begin{cases} \frac{1}{K}, & \text{if } n \leq 0 \text{ (uniform initial prior)} \\ 1, & \text{if } n > 0 \text{ and } s(n) = k \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

[1] Squires et al. (1976) denote expectancy also as *subjective probability*.

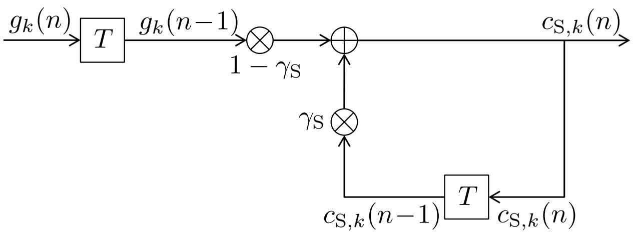


FIGURE 2 | Block diagram of a first-order infinite impulse response length (IIR) filter with transfer function $H_S(f)$ equivalent to (11) or (12). Elements $\boxed{T}$ denote a delay of one trial. At the adder element ⊕, updating of the weighted output $\gamma_S \cdot c_{S,k}(n-1)$ of the last trial $n-1$ with the weighted input $(1-\gamma_S) \cdot g_k(n-1)$ of the last trial results in the current output $c_{S,k}(n)$. Note that via $\gamma_S \cdot c_{S,k}(n-1)$, all preceding inputs (and outputs) influence the current output, though for the short-term memory, the influence of trials not in the recent past is negligible.

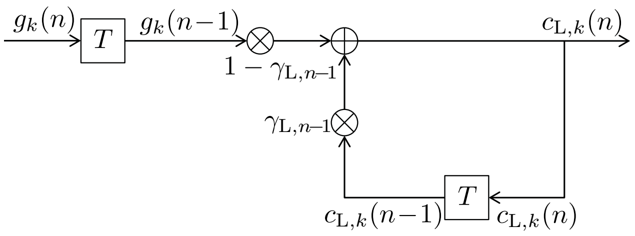


FIGURE 3 | Block diagram of a first-order infinite impulse response length (IIR) filter with transfer function $H_L(f,n)$ equivalent to (13) or (14). Elements $\boxed{T}$ denote a delay of one trial. At the adder element ⊕, updating of the weighted output $\gamma_{L,n-1} \cdot c_{L,k}(n-1)$ of the last trial $n-1$ with the weighted input $(1-\gamma_{L,n-1}) \cdot g_k(n-1)$ of the last trial results in the current output $c_{L,k}(n)$. Note that via $\gamma_{L,n-1} \cdot c_{L,k}(n-1)$, all preceding inputs (and outputs) influence the current output.

which implicitly contains an initial prior of $1/K$ at the start of a block of trials[2], and a "1" wherever a past stimulus $s(\nu)$ equals the current stimulus $s(n) = k$, otherwise a "0." Note that in contrast to the sequence $d_k(\nu)$ as used in Mars et al. (2008) (6) and Squires et al. (1976) (8) we define a model-exciting infinite length *signal* $g_k(\nu)$, $\nu \in \{-\infty, \ldots, n-2, n-1\}$. The digital filter model yields an output signal $P_k(n)$ as given in (9). The weighting parameters $\alpha_L, \alpha_S, \alpha_\Delta$ hold $\alpha_L + \alpha_S + \alpha_\Delta = 1$ and $0 \le \alpha_i \le 1, i \in \{L,S,\Delta\}$.

2.4.1. Short-term memory

The block diagram of the infinite impulse response (IIR) digital filter is shown in **Figure 2**. The respective short-term memory count function can be expressed as

$$c_{S,k}(n) = \frac{1}{C_S} \sum_{\nu=-\infty}^{n-1} \gamma_S^{n-\nu} g_k(\nu), \tag{11}$$

with some normalizing constant C_S and an exponential forgetting factor $\gamma_S = e^{-\frac{1}{\beta_S}}$ with $0 \le \beta_S < \infty$, as with count function $\breve{c}_{S,k}(n)$ in (8). The transfer function of the short-term digital filtering process as described by (11) is depicted in **Figure 1** as $H_S(f)$, and is plotted in **Figure 4** as dashed curve, revealing a smooth (i.e., weak) low-pass characteristic. Note that the short-term memory count function (11) can be expressed mathematically equivalent in a recursive form according to (see the Appendix)

$$c_{S,k}(n) = (1-\gamma_S) \cdot g_k(n-1) + \gamma_S \cdot c_{S,k}(n-1), \tag{12}$$

initialized with the uniform initial prior $c_{S,k}(0) = 1/K$, which in (11) was contained in the values $g_k(\nu) = 1/K$ for $\nu \le 0$ of (10). The recursive character of (12) becomes apparent by substituting the right hand side of (12) in itself for calculating $c_{S,k}(n-1)$. **Figure 2** further illustrates the updating process inherent in (12). The input signal defined in (10) and the weights $(1-\gamma_S)$ and γ_S guarantee $0 \le c_{S,k}(n) \le 1$. The equivalence of (11) and (12) and the derivation of $C_S = \gamma_S/(1-\gamma_S)$ are shown in the Appendix.

2.4.2. Long-term memory

The long-term memory count function can be expressed as

$$c_{L,k}(n) = \frac{1}{C_{L,n}} \sum_{\nu=-\infty}^{n-1} \gamma_{L,n}(\nu)\, g_k(\nu), \tag{13}$$

with the time-dependent (i.e., *dynamic*) exponential forgetting factor $\gamma_{L,n}(\nu) = \prod_{\upsilon=\nu+1}^{n} \gamma_{L,\upsilon} \frac{1-\gamma_{L,\upsilon-1}}{1-\gamma_{L,\upsilon}}$ (using $\gamma_{L,\upsilon} = e^{-\frac{1}{\beta_{L,\upsilon}}}$), the dynamic normalizing value $C_{L,n}$, and the same model-exciting signal $g_k(\nu)$ as before (10). The formulas for calculating $\gamma_{L,n}(\nu)$ and $C_{L,n}$ are derived in the Appendix. The transfer function of the long-term digital filtering process as described by (13) is depicted in **Figure 1** as $H_L(f,n)$. Analog to (12) a recursive function with the same behavior as (13) can be defined as

$$c_{L,k}(n) = \left(1-\gamma_{L,n-1}\right) \cdot g_k(n-1) + \gamma_{L,n-1} \cdot c_{L,k}(n-1) \tag{14}$$

with the same initial value $C_{L,k}(0) = 1/K$. The dynamics of the forgetting factor of the long-term memory are detailed in the Appendix.

Figure 3 illustrates the updating process inherent in (14). The long-term transfer function $H_L(f,n)$ is plotted in **Figure 4** as dash-dotted curve (for $n = 1$) and as solid curve (for $n = N = 192$), respectively. Inspection of **Figure 4** reveals an initially moderate low-pass characteristic which becomes much sharper when the number of trials increases.

2.4.3. Alternation expectation

Finally, our model comprises a count function capturing alternations:

$$c_{\Delta,k}(n) = \frac{1}{C_\Delta} \sum_{\nu=n-4}^{n-1} \gamma_{\Delta,n-\nu} \cdot g_k(\nu) \tag{15}$$

with some normalizing constant C_Δ and the same model-exciting signal $g_k(\nu)$ as before (10). In contrast to the short- and long-term

[2] Participants were not informed about event probabilities, therefore a uniform initial prior distribution $P_k(1) = 1/K$, $k \in \{1,\ldots,K\}$ is a reasonable model assumption before any stimulus has been observed.

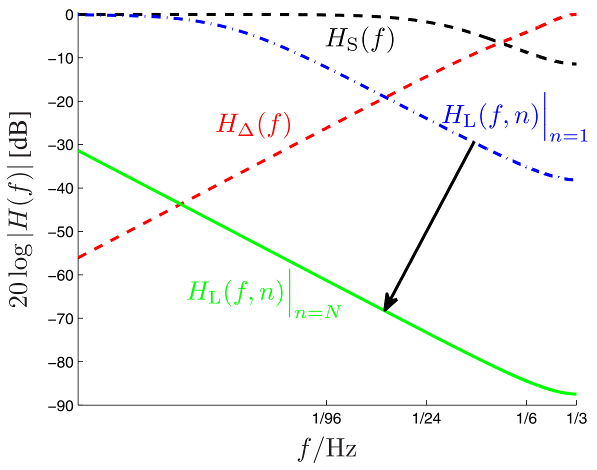


FIGURE 4 | Amplitude responses of the long-term ($H_{\mathrm{L}}(f,n)$, dash-dotted and solid low-pass curves), short-term ($H_{\mathrm{S}}(f)$, dashed low-pass curve) and alternation ($H_{\Delta}(f)$, dashed high-pass curve) filters of the DIF model as a function of the input signal frequency f (logarithmic scale!). The dynamic long-term filter (dash-dotted and solid curves) is shown for $n=1$ and $n=N=192$.

IIR filters which both have a low-pass characteristic, this finite impulse response (FIR) filter reveals a high-pass characteristic. Its transfer function is plotted in **Figure 4** as ascending dashed curve. The coefficients $\gamma_{\Delta,n-\nu}$ are specified in more detail in the Appendix.

2.4.4. Explanatory notes

There are two important differences to Squires et al. (1976): We propose to use *two* terms with different time parameters β_{S} and $\beta_{\mathrm{L},n}$, one accounting for the short-term memory [as in (8)], the other one accounting for a dynamically adapted long-term memory. Secondly, we allow for $N_{\mathrm{depth}} \to \infty$. Moreover, the role of negative trial indices in our model is to define the initial subjective probability distribution, which is $P_k(1) = 1/K$, $k \in \{1, \ldots, K\}$ in (9), reflecting that participants were not informed about the actual relative frequencies of events over a block of trials.

In summary, the DIF model expresses both a long-term memory contribution, and a short-term memory contribution by exponential decay processes, with uniform initial subjective prior probabilities, and a contribution of alternation expectation. Though there are similarities between (7) and (9), the model of Squires et al. (1976) uses information about the experimental design (P_k) which was actually unknown to the participants. In contrast, the DIF model uses only the information contained in the stimulus sequence as observed by participants, and it always starts with a uniform initial subjective probability distribution $P_k(1) = 1/K$, $k \in \{1, \ldots, K\}$, regardless of the actual relative frequencies of events over a block of trials. Finally it should be noted that our new model yields a conceptually well-defined subjective probability, as opposed to an expectancy as in Squires et al. (1976).

2.4.5. Evaluation methods

Following Mars et al. (2008) we compared the DIF to the MAR and SQU models using the log-Bayes factor based on the model evidences. The evidences were approximated using the variational free energy which consists of an accuracy and complexity term, thus enabling the comparison and selection of competing models (Penny et al., 2004; Friston et al., 2007; Penny, 2012). We employed the same three-level hierarchical general linear model (GLM) as Mars et al. (2008). For model fitting and calculation of the model evidences we used parametric empirical Bayes (PEB) from the `spm_PEB.m` function of the Statistical Parametric Mapping (SPM8) software (Friston et al., 2002, 2007).

The different models (DIF, MAR, and SQU) generate the model-specific surprise[3] $I_\ell(n)$ or expectancy $E_{k,\ell}(n)$ values as regressors, with the subscript $\ell \in \{1, \ldots, L = 16\}$ denoting the individual participants and $n \in \{1, \ldots, N = 192\}$ being the discrete time index of the consecutive trials within one block. The first level of the GLM models the *measured trial-by-trial P300 estimates* $Y_\ell(n)$ [μV] as a linear function of the surprise $I_\ell(n)$ [bit] with the intercept $\theta_\ell^{(1)}$[μV], the slope $\vartheta_\ell^{(1)}$ [μV/bit], and an error $\epsilon_{n,\ell}^{(1)}$[μV] :

$$Y_\ell\,(n) = \theta_\ell^{(1)} + \vartheta_\ell^{(1)} I_\ell\,(n) + \epsilon_{n,\ell}^{(1)}. \tag{16}$$

Note that the fitted model-based P300 estimates then follow

$$\widehat{Y}_\ell(n) = \theta_\ell^{(1)} + \vartheta_\ell^{(1)} I_\ell(n). \tag{17}$$

The second level models the *participant-specific parameters* $\theta_\ell^{(1)}$ and $\vartheta_\ell^{(1)}$ as *deviations from the corresponding group parameters* $\theta^{(2)}$ [μV] and $\vartheta^{(2)}$[μV/bit]:

$$\theta_\ell^{(1)} = \theta^{(2)} + \epsilon_{\theta,\ell}^{(2)} \tag{18}$$

$$\vartheta_\ell^{(1)} = \vartheta^{(2)} + \epsilon_{\vartheta,\ell}^{(2)}. \tag{19}$$

The third level functions as a shrinkage prior on the group parameters. In matrix notation the GLM structure is

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}^{(1)}\boldsymbol{\Theta}^{(1)} + \mathbf{E}^{(1)} \\ \boldsymbol{\Theta}^{(1)} &= \mathbf{X}^{(2)}\boldsymbol{\Theta}^{(2)} + \mathbf{E}^{(2)} \\ \boldsymbol{\Theta}^{(2)} &= \mathbf{X}^{(3)}\boldsymbol{\Theta}^{(3)} + \mathbf{E}^{(3)}. \end{aligned} \tag{20}$$

$\mathbf{Y} = [\mathbf{Y}_{\ell=1,\ldots,}\ \mathbf{Y}_{\ell=L}]^T \in \mathbb{R}^{2NL \times 1}$ is a vector concatenating the trial-by-trial P300 estimates for all participants, with $[\,]^T$ being the transpose, making $\mathbf{Y}$ a column vector. The participant-specific vector $\mathbf{Y}_\ell = [Y_\ell(n=1), \ldots, Y_\ell(n=2N)] \in \mathbb{R}^{1 \times 2N}$ contains the trial-by-trial P300 estimates for one participant, averaged over the six sequence repetitions, for both probability categories. The first level design matrix $\mathbf{X}^{(1)} \in \mathbb{R}^{2NL \times 2L}$ is block-diagonal with L partitions $\mathbf{X}_\ell^{(1)} = [\mathbf{1}_{2N}\,\mathbf{I}_\ell] \in \mathbb{R}^{2N \times 2}$, each of which contains an all-one column vector $\mathbf{1}_{2N}$ of length $2N$, and surprise values $\mathbf{I}_\ell = [I_\ell(n=1), \ldots, I_\ell(n=2N)]^T \in \mathbb{R}^{2N \times 1}$

[3] In the following, whenever $I_\ell(n)$ is used, $E_{k=s(n),\ell}(n)$ can be used interchangeably.

as explanatory variables. The second level design matrix $\mathbf{X}^{(2)} = \mathbf{1}_L \otimes \mathbf{I}_{2\times 2} = [\mathbf{I}_{2\times 2,\ell=1,\ldots,}\ \mathbf{I}_{2\times 2,\ell=L}]^T \in \mathbb{R}^{2L\times 2}$ is the Kronecker product of an all-one column vector $\mathbf{1}_L$ of length L and an identity matrix $\mathbf{I}_{2\times 2}$. The third level design matrix $\mathbf{X}^{(3)}$ shall have all-zero elements. The unknown level-one parameters $\theta_\ell^{(1)}$ and $\vartheta_\ell^{(1)}$ are assembled in the parameter vector $\mathbf{\Theta}^{(1)} = [\theta_{\ell=1}^{(1)}, \vartheta_{\ell=1}^{(1)}, \ldots, \theta_{\ell=L}^{(1)}, \vartheta_{\ell=L}^{(1)}]^T \in \mathbb{R}^{2L\times 1}$. Likewise, the second level parameters $\theta^{(2)}$ and $\vartheta^{(2)}$ are assembled in the vector $\mathbf{\Theta}^{(2)} \in \mathbb{R}^{2\times 1}$. All errors are assumed to be normally distributed $\mathbf{E}^{(j)} \sim \mathcal{N}(0, \Sigma_\epsilon^{(j)})$. The covariance is parameterized following $\Sigma_\epsilon^{(j)} = \lambda^{(j)}\mathbf{I}^{(j)}$, with $\mathbf{I}^{(j)}$ as an identity matrix with the same dimension as the number of rows of the design matrix of the corresponding level $\mathbf{X}^{(j)}$. The hyperparameters $\lambda^{(j)}$ are the free parameters of the hierarchical linear model and are estimated using an EM algorithm for maximum likelihood estimation.

The conditional means of the first level parameters $\boldsymbol{\mu}_{\Theta|Y}^{(1)}$ of the posterior densities $\mathcal{N}(\boldsymbol{\mu}_{\Theta|Y}^{(j)}, \mathbf{\Sigma}_{\Theta|Y}^{(j)})$ were used as maximum *a posteriori* point estimates of the parameters for the model fitting for the **Figures 6–8** and the calculation of the mean squared error (MSE) and fraction of variance explained (FVE) between $\hat{Y}_\ell(n)$ and $\mathrm{Y}_\ell(n)$ in **Table 4** (Friston et al., 2002). The log-evidences or marginal log-likelihoods of the models $F = \ln(p(\mathbf{Y}\,|\,\mathcal{M}))$, with $p(\mathbf{Y}\,|\,\mathcal{M})$ being the likelihood of the data Y given the model $\mathcal{M}$, were used for model comparison via the log-Bayes factor ln(BF) which is the natural logarithm of the quotient of the model likelihoods or the difference in log-evidence (Kass and Raftery, 1995; Penny et al., 2004; Friston et al., 2007):

$$\ln\left(\mathrm{BF}_{\mathrm{DIF-XXX}}\right) = \ln\left(\frac{p\left(\mathbf{Y}|\mathcal{M}_{\mathrm{DIF}}\right)}{p\left(\mathbf{Y}|\mathcal{M}_{\mathrm{XXX}}\right)}\right) = F_{\mathrm{DIF}} - F_{\mathrm{XXX}} \qquad (21)$$

If the log-evidence is calculated this way, positive values reflect evidence in favor of the DIF model and negative values in favor of the XXX model (being SQU or MAR), respectively. Values larger than five are considered "very strong" evidence (Kass and Raftery, 1995; Penny et al., 2004). To summarize, in our evaluation method we closely followed (Mars et al., 2008).

2.4.6. *DIF model parameter identification*

The values for the free model parameters, namely α_L in (9), τ_1 and τ_2 (both (A12)), β_S [for (11)], α_Δ in (9), and $\gamma_{\Delta,2}$ in (15), have to be trained on the measured data, i.e., on the trial-by-trial P300 estimates. The model-based P300 estimates are calculated with the *same model parameters for all participants* and then used for maximization of the DIF model evidence, which is our optimization criterion, as described in Section 2.4.5.

The calculation of the model evidence for the whole range of possible combinations of the parameters with a reasonable resolution is computationally too expensive. For this reason only subsets of parameter combinations were optimized simultaneously with a resolution of 100 values per parameter, which results in 30,000 possible parameter combinations for one iteration. In the first iteration, while optimizing one set of parameters, the not yet optimized parameters were fixed to the center of their respective intervals. In the following iteration, parameters not currently optimized were fixed to the optimal values from the last iteration. In these two iterations a total of 60,000 parameter combinations have been evaluated, and the set with the highest evidence was considered optimal. Note that only a locally optimal parameter combination can be found using this procedure, as many iterations may be necessary for convergence toward the global optimum, if it can be found at all. **Table 1** gives an overview over the searched parameter space.

Table 1 | The ranges of the free model parameters, with a resolution of 100 values per parameter.

Parameter	Min	Max
α_L	0.5	0.9
τ_1	10	100
τ_2	0.1	1
β_S	1	10
α_Δ	0.001	0.1
$\gamma_{\Delta,2}$	0.5	1

Note that α_S is not a free parameter due to the restriction $\alpha_S = 1 - \alpha_L - \alpha_\Delta$ and thus not optimized independently.

3. RESULTS

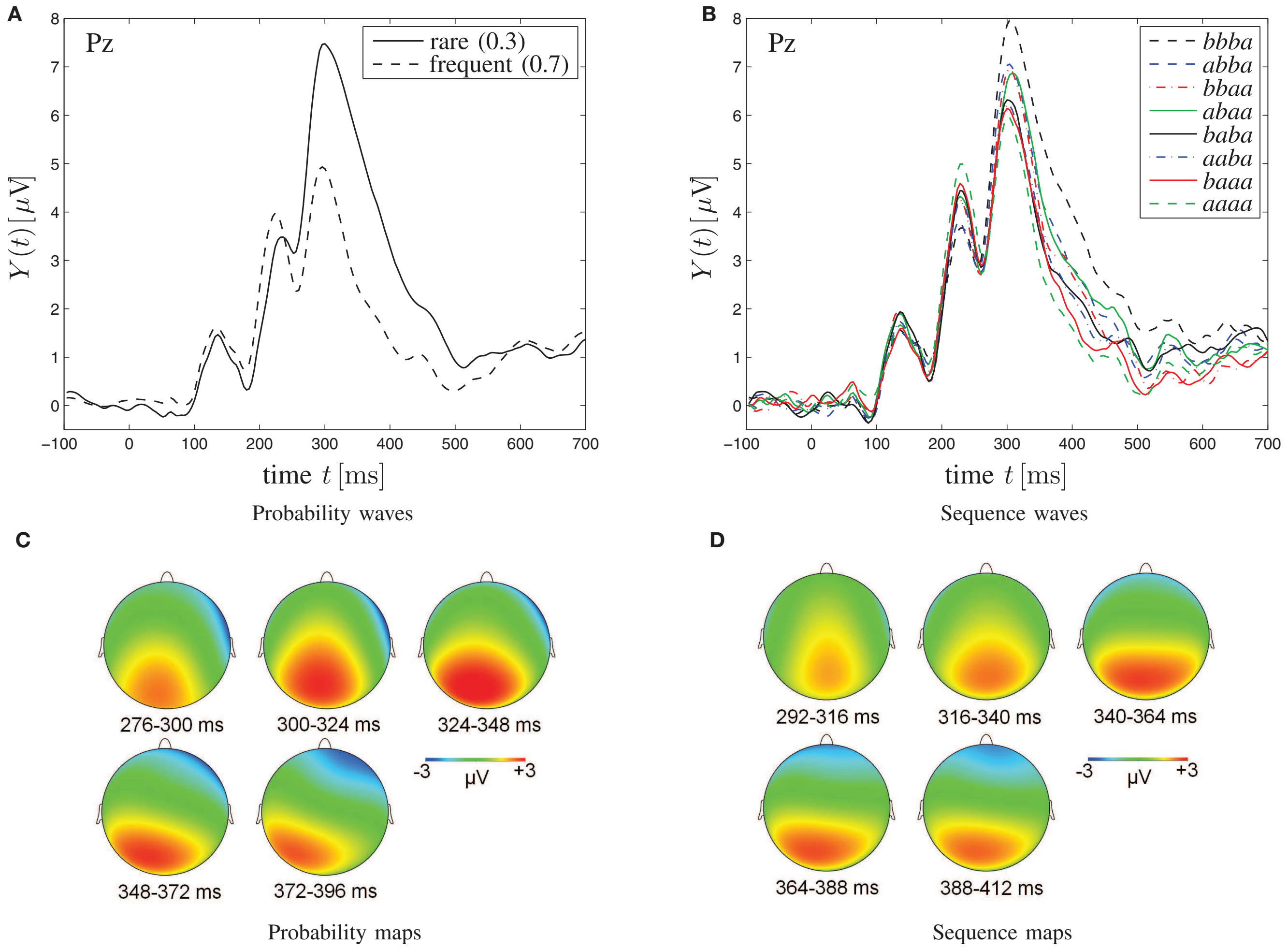
3.1. BEHAVIORAL RESULTS

RTs showed clear dependence on stimulus probability (0.3, $M = 373.4$ ms, $SE = 6.9$ ms; 0.5, $M = 353.5$ ms, $SE = 7.7$ ms; 0.7, $M = 319.8$ ms, $SE = 6.4$ ms). The slowdown of responding to less probable stimuli was confirmed by an ANOVA on RTs as a function of probability, $F_{(2,30)} = 80.98$, $p < 0.001$, $\eta_p^2 = 0.84$, $\epsilon = 0.95$. Polynomial contrasts revealed a linear trend, $F_{(1,15)} = 166.79$, $p < 0.001$, $\eta_p^2 = 0.92$, in the absence of a quadratic trend, $F_{(1,15)} = 3.37$, $p > 0.05$.

Error rates similarly showed clear dependence on stimulus probability (0.3, $M = 9.6\%$, $SE = 1.6\%$; 0.5, $M = 4.7\%$, $SE = 0.9\%$; 0.7, $M = 2.1\%$, $SE = 0.4\%$). The enhanced error proneness in response to less probable stimuli was confirmed by an ANOVA on arcsin-transformed error rates as a function of probability, $F_{(2,30)} = 25.19$, $p < 0.001$, $\eta_p^2 = 0.63$, $\epsilon = 0.65$. Polynomial contrasts revealed a linear trend, $F_{(1,15)} = 28.94$, $p < 0.001$, $\eta_p^2 = 0.66$ as well as a quadratic trend, $F_{(1,15)} = 4.86$, $p < 0.05$, $\eta_p^2 = 0.25$.

3.2. CONVENTIONAL ERP RESULTS

Figure 5 depicts grand-average ERP waveforms (upper panels) and topographic maps (lower panels). Left panels illustrate ERP waveforms at Pz that were obtained in the $[0.3, 0.7]$ probability category. Right panels show third-order sequence effects on ERP waveforms at Pz that were obtained in the $[0.5, 0.5]$ probability category. Note that sequences of four successive stimuli are illustrated, in temporal order (trial $n-3$, trial $n-2$, trial $n-1$, trial $n =$ eliciting event); *a* signifies a particular stimulus, *b* the other one. For example, *aaaa* gives a description of stimulus *a* being repeated across four consecutive trials (shown as green dashed curve), whereas *bbba* represents the presentation of stimulus *a* after having stimulus *b* repeated across the three immediately preceding trials (shown as black dashed curve).



Probability waves

Sequence waves

Probability maps

Sequence maps

FIGURE 5 | Grand-average waveforms (A,B) and topographic maps (C,D) of P300 amplitudes. (A,C) Probability effect on P300 amplitudes in the [0.3, 0.7] probability category. **(C)** The probability maps show the scalp topography of the rare-frequent difference wave in the [0.3, 0.7] probability category at various points in time (276–396 ms, divided into five windows of 24 ms each). **(B,D)** Sequence effect on P300 amplitudes in the [0.5, 0.5] probability category. Note that sequences of four successive stimuli are illustrated; *a* signifies a particular stimulus (*b* the other one). Note further that the two solid traces, originating from the *abaa* and the *baba* sequences, respectively, show reversed P300 amplitudes. Specifically, for the single -*b*- sequence *abaa*, the P300 waveform lies amongst those from dual -*bb*- sequences, whereas for the dual -*bb*- sequence *baba*, the P300 waveform appears indistinguishable from the waveforms from single -*b*- sequences. As further detailed in the Discussion, this amplitude reversal is attributed to the disconfirmation of alternation expectation in the *abaa* sequence, was well as to the confirmation of alternation expectation in the *baba* sequence. **(D)** Sequence maps show the scalp topography of the *bbba-aaaa* difference wave in the [0.5, 0.5] probability category at various points in time (292–412 ms, divided into five time windows of 24 ms each).

As can be seen from **Figure 5A**, trial-by-trial P300 estimates showed clear dependence on stimulus probability (0.3, $M = 4.84\,\mu$V, $SE = 0.72\,\mu$V; 0.5, $M = 3.51\,\mu$V, $SE = 0.58\,\mu$V; 0.7, $M = 2.00\,\mu$V, $SE = 0.52\,\mu$V). P300 augmentation over stimulus improbability was confirmed by an ANOVA on P300 amplitudes as a function of probability, $F_{(2,30)} = 39.88$, $p < 0.001$, $\eta_p^2 = 0.73$, $\epsilon = 0.82$. Polynomial contrasts revealed a linear trend, $F_{(1,15)} = 55.49$, $p < 0.001$, $\eta_p^2 = 0.79$, in the absence of a quadratic trend, $F_{(1,15)} = 0.17$, $p > 0.05$.

Sequential P300 estimates in the [0.5, 0.5] probability category (**Figure 5B**) yielded main effects of first-, $F_{(1,15)} = 6.72$, $p < 0.05$, $\eta_p^2 = 0.31$, second-, $F_{(1,15)} = 21.04$, $p < 0.001$, $\eta_p^2 = 0.58$, and third-order sequences, $F_{(1,15)} = 6.89$, $p < 0.05$, $\eta_p^2 = 0.32$, as well as a significant three-way first- by second- by third-order sequence interaction, $F_{(1,15)} = 6.70$, $p < 0.05$, $\eta_p^2 = 0.31$. First-order alternations (*ba*; $M = 3.75\,\mu$V, $SE = 0.57\,\mu$V) were associated with enhanced P300 amplitudes compared to first-order repetitions (*aa*; $M = 3.26\,\mu$V, $SE = 0.59\,\mu$V). Likewise, second-order alternations (*bxa*; $M = 3.98\,\mu$V, $SE = 0.64\,\mu$V) were associated with enhanced P300 amplitudes compared to second-order repetitions (*axa*; $M = 3.03\,\mu$V, $SE = 0.51\,\mu$V). Finally, third-order alternations (*bxxa*; $M = 3.64\,\mu$V, $SE = 0.58\,\mu$V) were associated with enhanced P300 amplitudes compared to third-order repetitions (*axxa*; $M = 3.37\,\mu$V, $SE = 0.57\,\mu$V). Separate ANOVAs on

Table 2 | The maximum log-Bayes factors $\ln(\mathrm{BF}_{\mathrm{DIF}-\mathrm{XXX}})$, left panel $\ln(\mathrm{BF}_{\mathrm{DIF}-\mathrm{SQU}})$, right panel $\ln(\mathrm{BF}_{\mathrm{DIF}-\mathrm{MAR}})$.

$\ln(\mathrm{BF}_{\mathrm{DIF}-\mathrm{SQU}})$	$\ln(\mathrm{BF}_{\mathrm{DIF}-\mathrm{MAR}})$
35	170

Table 3 | The optimized model parameters.

α_{L}	τ_1	τ_2	α_{S}	β_{S}	α_Δ	$\gamma_{\Delta,2}$
0.83	33.6	0.27	0.12	1.82	0.05	0.94

Note that τ_1 and τ_2 yield $\beta_{L,1} = 40.3$ and $\beta_{L,192} = 11787$, respectively.

sequential P300 estimates were performed in each second-order sequence condition to further parse the three-way interaction. These ANOVAs revealed that the two-way first- by third-order sequence interaction was not significant when the second-order sequence consisted of stimulus repetitions (i.e., when *xaxa* sequences were included), $F_{(1,15)} = 0.94$, $p > 0.05$, whereas the two-way first- by third-order sequence interaction was significant when the second-order sequence consisted of stimulus alternations (i.e., when *xbxa* sequences were included), $F_{(1,15)} = 6.34$, $p < 0.05$, $\eta_p^2 = 0.30$. Further comments on these data are deferred to the Discussion.

Sequential P300 estimates in the [0.3, 0.7] probability category yielded significant main effects of stimulus probability, $F_{(1,15)} = 44.08$, $p < 0.001$, $\eta_p^2 = 0.75$ (0.3, $M = 4.59\,\mu$V, $SE = 0.70\,\mu\mathrm{V} > 0.7$, $M = 2.20\,\mu$V, $SE = 0.53\,\mu$V), of first-order sequence, $F_{(1,15)} = 5.80$, $p < 0.05$, $\eta_p^2 = 0.28$ (*ba*, $M = 3.68\,\mu$V, $SE = 0.58\,\mu\mathrm{V} >$ *aa*, $M = 3.10\,\mu$V, $SE = 0.63\,\mu$V), and of second-order sequence, $F_{(1,15)} = 11.20$, $p < 0.01$, $\eta_p^2 = 0.43$ (*bxa*, $M = 3.81\,\mu$V, $SE = 0.66\,\mu\mathrm{V} >$ *axa*, $M = 2.98\,\mu$V, $SE = 0.54\,\mu$V), but without interactions between these factors.

3.3. MODEL-BASED TRIAL-BY-TRIAL ANALYSIS

The maximum log-Bayes factors in favor of the DIF model over the MAR and SQU models are shown in **Table 2**. In both cases, this is considered a *very strong* evidence in favor of the DIF model (Kass and Raftery, 1995; Penny et al., 2004). **Table 3** shows the free parameters of the DIF model which were used for calculating the log-Bayes factors in **Table 2**. **Figure 6** illustrates how the log-Bayes factors vary in dependence on the model parameters, with the tip of the "V" marking the parameter combination with the highest evidence. It is important to note the relatively flat tops of the contours implying good generalization capability of the DIF model. Due to computational complexity, only two parameters were optimized simultaneously, as described in Section 2.4.6. The relatively high value of $\alpha_{\mathrm{L}} = 0.83$ shows that the subjective probability mainly follows the long-term memory. With the identified values for τ_1 and τ_2 we get $\beta_{\mathrm{L},1} = 40.3$ and $\beta_{\mathrm{L},192} = 11787$, which is further illustrated in **Figure A2** in the Appendix. With a short-term memory time constant of $\beta_{\mathrm{S}} = 1.82$ and a weight of $\alpha_{\mathrm{S}} = 0.12$ the influence of recent events to the subjective probability is captured. While the weight of the filter modeling alternation expectancy $\alpha_\Delta = 0.05$ appears to be small, **Figure 6C** clearly shows the importance of this contribution.

Table 4 | Comparison of the goodness-of-fit in terms of the mean squared error (MSE) and fraction of variance explained (FVE) of the fitted model predictions $\hat{Y}_\ell(n)$.

Model	MSE	FVE
MAR	6.0106	0.4909
SQU	5.7410	0.5138
DIF	5.6780	0.5191

Figure 7 shows timeline plots for one exemplary participant. **Figures 7A,B** show plots of the expectancy $E_{k=1}(n)$ and the subjective probability $P_{k=1}(n)$ of seeing stimulus $k = 1$ on trial n for the [0.5, 0.5] and [0.3, 0.7] probability category for all competing models. **Figures 7C,D** show plots of the expectancy $E_{k=s(n)}(n)$ and the subjective probability $P_{k=s(n)}(n)$ of seeing *the actually occurring stimulus* $k = s(n)$ on trial n for both probability categories and all models. The transition from **Figures 7A,B** to **Figures 7C,D** illustrates that the subjective probability is traced as a distribution for all possible events $k \in \{1, \ldots, K\}$ simultaneously over all trials and that at the moment of seeing a new stimulus $k = s(n)$ only the corresponding subjective probability of that event k is relevant for the surprise $I_\ell(n)$ and consequently for the model-based P300 estimate $\widehat{Y}_\ell(n)$. **Figures 7E,F** show plots of both the measured and the model-based trial-by-trial P300 estimates $Y_\ell(n)$ and $\widehat{Y}_\ell(n)$ for both probability categories, where the latter is calculated according to (17). It is visible that the DIF model estimates are smoother over trials than those of the SQU model but not as undynamic as the estimates of the MAR model, which loses its initial dynamic and becomes almost binary over increasing trial number n. This effect is especially prominent in **Figures 7E,F**. Similar consecutive trials n elicit a descent in the measured trial-by-trial P300 estimate. For small n all models show this behavior, but for increasing n the MAR model yields nearly constant estimates.

Furthermore the MAR model does not account appropriately for the well-documented sequence effects (Squires et al., 1976). **Figure 8** shows the tree diagrams of the measured $(Y_\ell(n))$ and model-based $(\widehat{Y}_\ell(n))$ P300 estimates as a function of the preceding stimuli sequences for the different probability categories. The DIF and SQU model are both capable of estimating the envelope and general tree structure quite well, but the SQU model fans out too much for higher order effects for the frequent stimulus in the [0.3, 0.7] and in general in the [0.5, 0.5] probability category, while for the MAR model higher-order effects are nearly non-existent.

As an additional measure of goodness-of-fit of the models **Table 4** shows the mean squared error (MSE) and the fraction of variance explained (FVE) of the fitted model predictions $\widehat{Y}_\ell(n)$. Although differences appear to be somewhat smaller, still the superiority of the DIF model is evident, supporting the log-Bayes factors presented in **Table 2**.

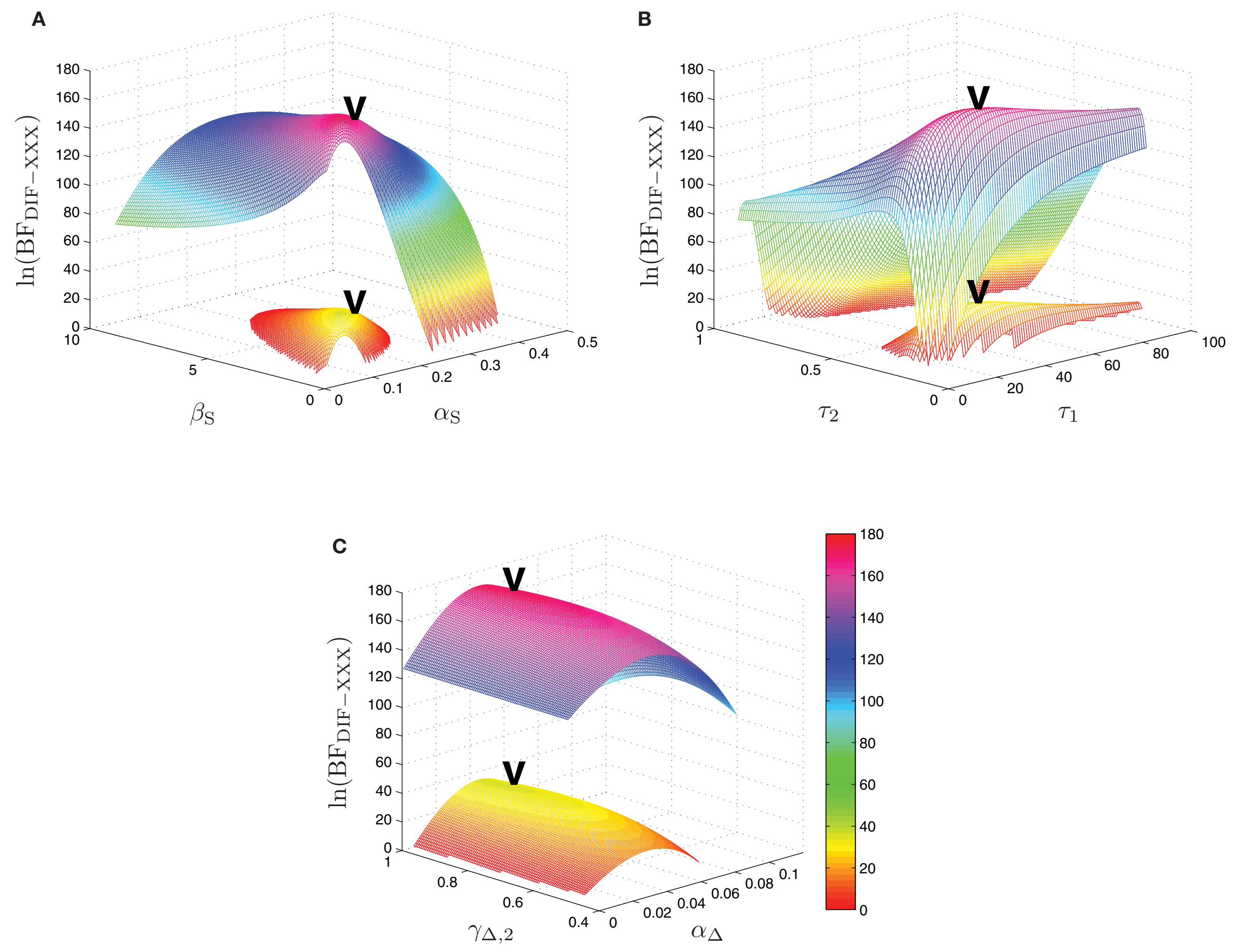


FIGURE 6 | Log-Bayes factors, ln(BF$_{\mathrm{DIF-XXX}}$), under variation of the model parameters. The upper contour always shows ln(BF$_{\mathrm{DIF-MAR}}$), the lower one ln(BF$_{\mathrm{DIF-SQU}}$). The "V" marks the parameter combination with the maximum log-Bayes factor, cf. **Table 3**. **(A)** Variation of the free short-term parameters β_S and α_S. **(B)** Variation of the free long-term parameters τ_1 and τ_2. In order to keep the lower contour visible, the log-Bayes factors for the 10 smallest values for τ_2 are not displayed. **(C)** Variation of the free alternation parameters $\gamma_{\Delta,2}$ and α_Δ.

4. DISCUSSION

We tested three computational models of trial-by-trial P300 amplitudes using Bayesian model selection (Kass and Raftery, 1995; Raftery, 1995). Trial-by-trial P300 amplitude estimates at Pz were obtained in a two-choice RT task. Behavioral data indicated that on average participants reacted slower and committed more errors when they responded to rarely occurring stimuli, consistent with many earlier reports (Miller, 1998). P300 amplitudes showed the expected relationships with stimulus probability. Further, they were influenced by the immediately preceding stimulus sequence, a finding which is also consistent with earlier reports. Thus, our data replicate two of the most ubiquitous P300 findings, notably probabilistic and sequential effects on P300 amplitudes.

The DIF model (9) possesses important advantages over previous models of P300 amplitude fluctuations. It relies completely on mathematical notations and definitions, unlike the notions of expectancy (Squires et al., 1976), global vs. local probability (Squires et al., 1976), temporal probability (Gonsalvez and Polich, 2002), or context updating (Donchin and Coles, 1988). It is a formal model, akin to the MAR model (Mars et al., 2008), but it offers a more parsimonious explanation of trial-by-trial P300 fluctuations. The competitive advantage of the DIF model over the MAR model stems, in large part, from the non-negligible contribution of the short-term traces to subjective estimates of event probabilities, as evidenced by the scarce sensitivity of the MAR model to the sequential effects on P300 amplitudes (**Figures 7** and **8**).

The SQU model of P300 amplitude fluctuations (Squires et al., 1976) can be considered as a precursor of our DIF model insofar as it comprised memory for event frequencies within the prior stimulus sequence (equivalent to the short-term trace) and event

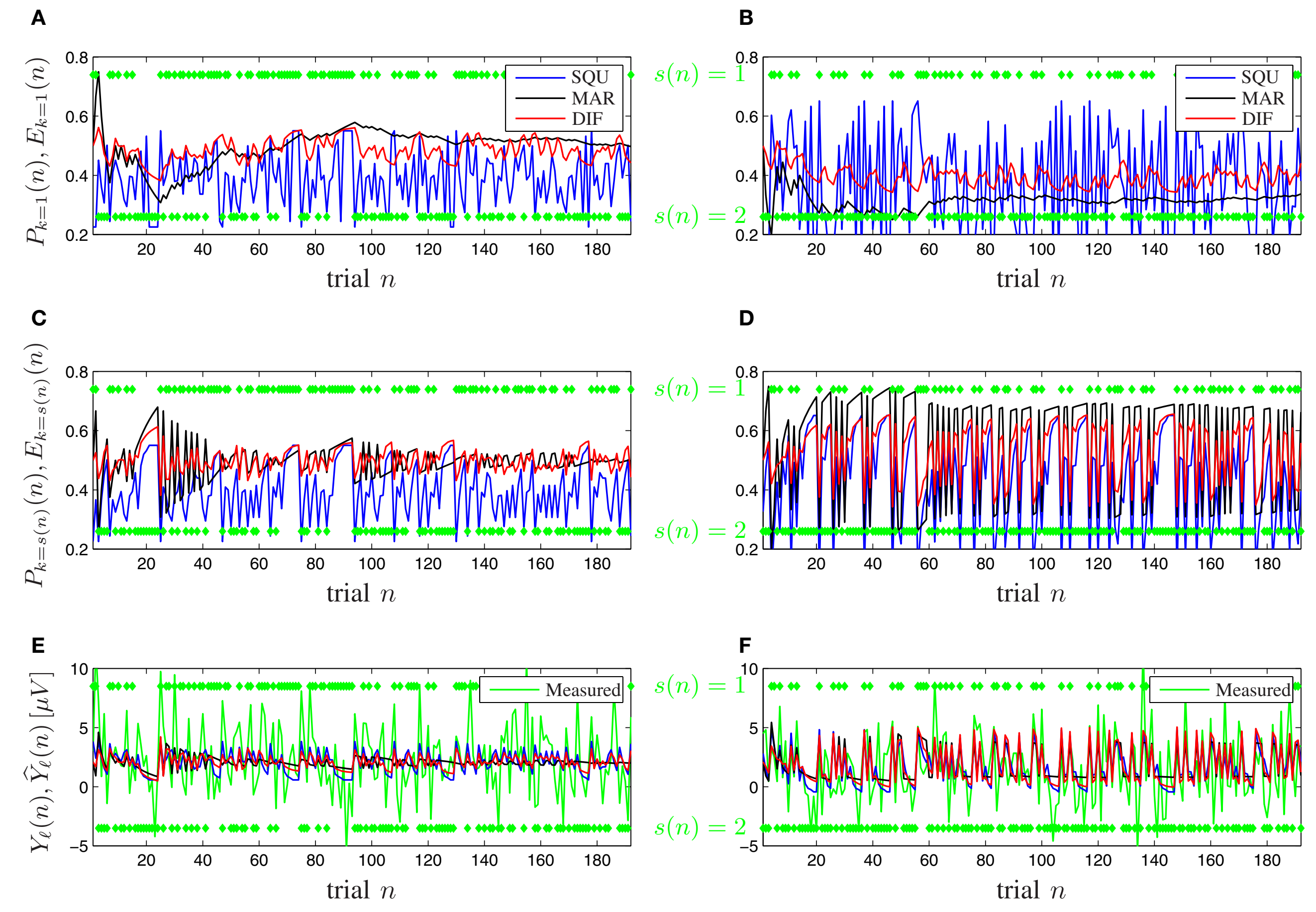


FIGURE 7 | Timeline plots for one exemplary participant. **(A,C,E)** Probability category [0.5, 0.5]. **(B,D,F)** Probability category [0.3, 0.7]. Green symbols denote the stimuli $s(n) \in \{1,2\}$ as they occurred. **(A,B)** Subjective probabilities $P_{k=1}(n)$ and expectancies $E_{k=1}(n)$ from (5), (7), and (9) for MAR, SQU, and DIF, respectively, for stimulus $s(n) = 1$. **(C,D)** Subjective probabilities $P_{k=s(n)}(n)$ and expectancies $E_{k=s(n)}(n)$ for the actually presented stimulus $k = s(n)$. **(E,F)** The measured P300 estimates $Y_\ell(n)$ and the model-based P300 estimates $\widehat{Y}_\ell(n)$ of the MAR, SQU, and DIF models, respectively.

probabilities (loosely related to the long-term trace). Yet, the long-term contribution to the DIF model does not incorporate global event probabilities, P_k, themselves, but rather subjective estimates, $c_{\mathrm{L},k}(n)$, of these probabilities, these being based on counting observed events in continuously larger samples. Thus, subjective probability estimates are constantly revised while evidence is accumulating, and the DIF model is a pure model of subjective statistical parameters, rather than a mixture of subjective and objective parameters such as the SQU model.

Given the hereby documented superiority of the DIF model over its competitors, we will shortly consider some of its cornerstones. To begin with, it is important to view the DIF model in the context of the processing of event frequencies (Sedlmeier and Betsch, 2002). In particular, the reliable encoding of the frequency with which events occur (Underwood, 1969; Hintzman, 1976) led to the claim that event frequency is automatically encoded in memory, placing only minimal demands on attentional resources (Hasher and Zacks, 1984; Zacks and Hasher, 2002). A variety of representation modes for memory for event frequency can be envisaged. According to multiple-trace views (Hintzman, 1976), a record of individual events is stored such that each attended occurrence of an event results in an independent memory trace. In contrast, according to strength views, each attended event occurrence produces an increment in the strength of a single memory trace or a frequency counter (Alba et al., 1980), supporting the event frequency counter assumption which is inherent in the DIF model. Both, the short-term, $c_{\mathrm{S},k}(n)$, and the long-term, $c_{\mathrm{L},k}(n)$, memory traces are frequency counters (11) and (13).

The retention functions describing the short-term and long-term traces in the DIF model are of exponential-decay nature (Lu et al., 2005), differing mainly with regard to their decay half-lives. The dual decay rate assumption is compatible with the fact that short-term and long-term memory functions depend on dissociable neuronal processes (Jonides et al., 2008). Further, recent functional brain imaging data suggest different distributions of cortical responses for short-term and long-term decay functions (Harrison et al., 2011).

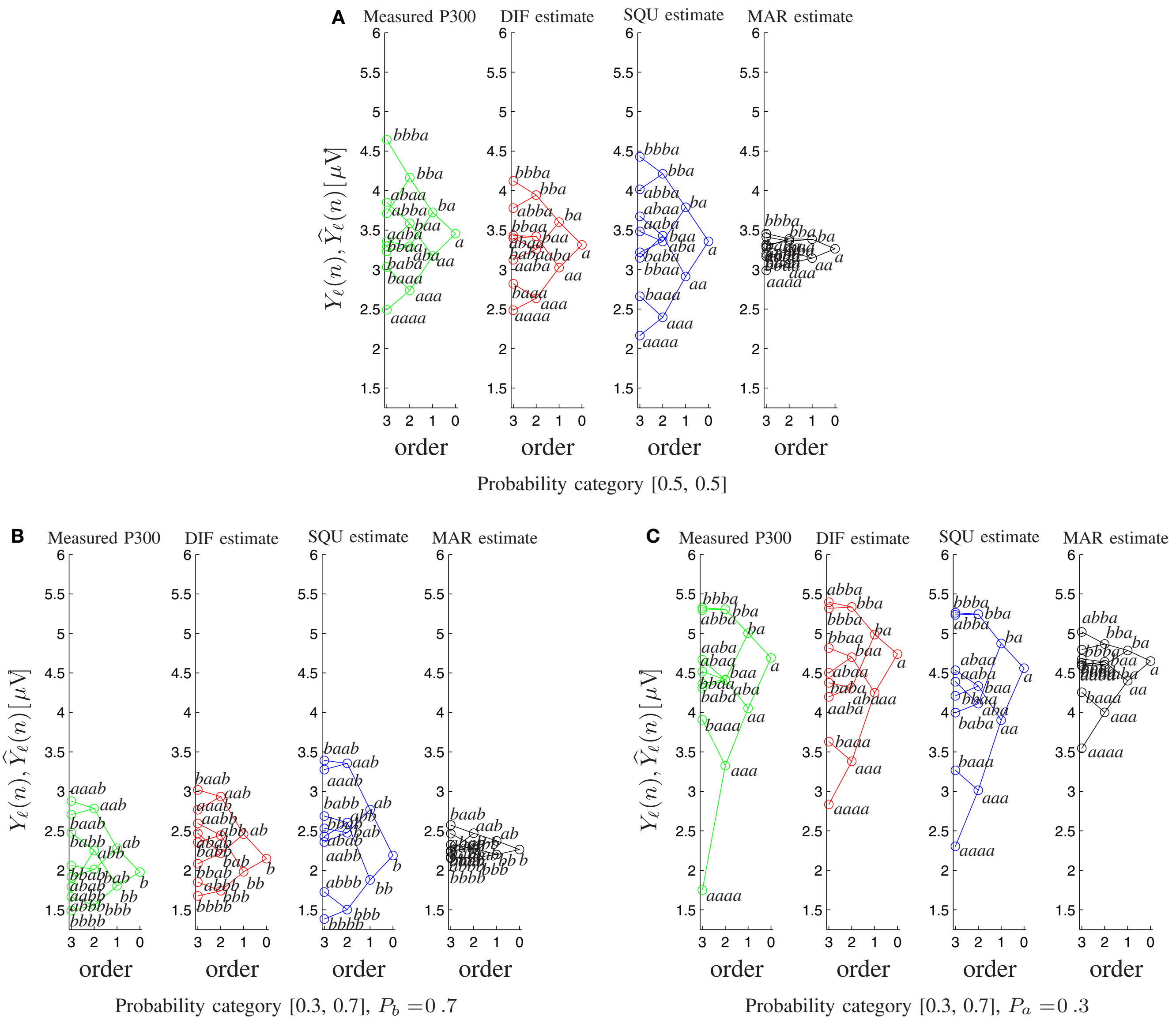


FIGURE 8 | Tree diagrams of measured P300 estimates $Y_\ell(n)$ and model-based P300 estimates $\widehat{Y}_\ell(n)$ as a function of the sequence of preceding stimuli. Within each order (0–3), the stimulus sequence is labeled, and related sequences are connected by lines. **(A)** For both stimuli on probability category [0.5, 0.5]. **(B)** For the frequently occurring stimulus *b* on probability category [0.3, 0.7]. **(C)** For the rarely occurring stimulus *a* on probability category [0.3, 0.7].

The optimal short-term time constant β_S approximated the value of two. Note that the DIF model does not allow for variations in β_S, implying that the short-term contribution to the subjective probability for the appearance of event k on trial n, $P_{k=s(n)}(n)$, occurs stable over the progression of n. In contrast, $\beta_{L,n}$ varies as a function of n (A12), with $\gamma_{L,n}$ gradually approximating the value of one. $\beta_{L,n}$ (and hence $\gamma_{L,n}$) are relatively low during early trials when compared to late trials within blocks of trials (**Figure A2** in Appendix). On the one hand, the long-term quality of the long-term contribution to $P_{k=s(n)}(n)$ is gradually increasing as a function of n, as revealed by the dynamics of the long-term low-pass filter (**Figure 4**). In other words, the decay half-life of the long-term trace gradually increases when the observer experiences more and more trials. On the other hand, the recursive formulation of the long-term contribution in the DIF model (14) reveals that the balance between the most recently experienced stimuli (which occurred on trial $n-1$, weighted by $(1-\gamma_{L,n-1})$) and the counted frequency (weighted by $\gamma_{L,n-1}$) is biased toward recent stimuli during early trials (when $\gamma_{L,n-1}$ is relatively low), but biased toward the counted frequency during late trials (when $\gamma_{L,n-1}$ is relatively high). Thus, the DIF model postulates that the decay half-life of the long-term trace evolves dynamically with the amount of experience. The observer is modeled to rely more and more on environmental experience rather than on prior assumptions, possibly reflecting the fact that the exploitation of statistical redundancy becomes gradually more reliable with progressing time (Barlow, 2001).

Visual working memory (vWM, Baddeley, 2003) can maintain representations of only around four objects at any given moment (Cowan, 2005). The surprisingly limited vWM capacity offers a rationale for the assumption of a capacity-limited alternation term in the DIF model (15), $c_{\Delta,k}(n)$. Its finite impulse response (FIR) characteristic resembles the alternation term in the SQU model (Squires et al., 1976). This FIR high-pass filter searches for alternation patterns over short sequences of trials (such as those in *abab* and in *baba* sequences). The discovery of such patterns leads one to expect the completion of the pattern in the upcoming trial, an expectation which will be confirmed in the *ababa* sequence, but will be disconfirmed in the *babaa* sequence. It is important that alternation expectation appears conditional upon the detection of alternation patterns in vWM, as revealed by the larger P300 amplitudes in response to *ba* sequences compared to *aa* sequences (**Figure 8**). Thus, the detection of alternation patterns in vWM entrains alternation expectation, as evidenced by our data (**Figure 5B**, see also Jentzsch and Sommer, 2001; Ford et al., 2010). Further, the effects of pattern completion (such as *baba* sequences) vs. pattern violation (such as *abaa* sequences) might underlie the first- by second- by third-order sequence interaction which we identified in the [0.5, 0.5] probability category.

While the effects of alternation expectation are non-negligibly measurable at Pz, visual inspection of our data at more anterior electrode sites suggested that these alternation expectancy effects might possess a more anterior, i.e., P3a-like scalp topography than the proper event frequency effects which showed the typical, P3b-like scalp topography (Polich, 2007; Duncan et al., 2009). While these initial observations ask for multi-channel data analyses, the present modeling work should be mainly considered as a model for P3b generation, since the task procedures (all stimuli required a button press) and the ERP waveforms [i.e., their scalp distribution (cf. **Figure 5**) and peak latencies] favor an interpretation in terms of predominant P3b-potentials.

The DIF model offers a digital filtering account of multiple memory systems in the brain (**Figures 1–4**). Specifically, the DIF model characterizes frequency memory as two digital first-order infinite impulse response (IIR) low-pass filters, one filter with an experience-invariant short-term exponential-decay function (**Figure 2**), and another filter with an experience-dependent long-term exponential-decay function, such that the low-pass characteristic becomes progressively apparent as the amount of experience increases (**Figure 3**). Moreover, vWM is conceptualized as an additional fourth-order finite impulse response (FIR) high-pass filter (**Figure 9**). The input signal $g_k(n)$ in (10) to all three filters is a binary representation of the stimulus sequence, with all samples prior to the first trial filled with the uniform initial prior.

Our theory of variation in trial-by-trial P300 amplitudes bears implications on the nature of cortical processing. It is in agreement with the predictive coding approach (Friston, 2002, 2005; Spratling, 2010). Viewed from this perspective, predictive surprise – and hence trial-by-trial P300 amplitude – is proportional to the residual error between top-down priors and bottom-up sensory evidence. Predictive coding theory is a successful guiding model for functional neuroimaging and electrophysiological studies of sensory cortical processing (Summerfield et al., 2006; Garrido et al., 2009; Summerfield and Egner, 2009; Egner et al., 2010; Rauss et al., 2011; Winkler and Czigler, 2011). Further, the DIF model is a Bayesian model of cortical processing (Knill and Pouget, 2004; Friston, 2005). It conceives performance on our two-choice RT task as sequential Bayesian learning (MacKay, 2003), with initial prior knowledge being conceptualized as a uniform prior probability distribution (10), consistent with Laplace's Principle of Indifference.

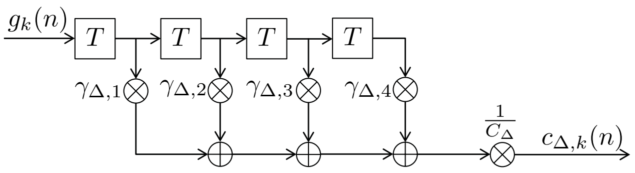


FIGURE 9 | Block diagram of the fourth-order finite impulse response (FIR) filter $H_\Delta(f)$. Elements $\boxed{T}$ represent a delay of one trial, the multipliers $\gamma_{\Delta,i}$ compose the filter coefficients and C_Δ constitutes a normalizing constant.

It is important to note that we do not claim that the observed P300 modulations were exclusively related to predictive surprise *over sensory input*, since in the present task design the probabilities of sensory events were mirrored on probabilities of motor responses, as each stimulus was mapped onto a distinct motor response. Thus, particular stimuli also called for particular motor programs, and it could be that the observed P300 modulations are related to predictive surprise *over motor responses*. We deliberately leave it open whether the observed P300 modulations were due to surprise conveyed by the visual stimuli, or whether they were related to surprise associated with the selection of a motor response, given a visual stimulus on each trial (Barceló et al., 2008; O'Connell et al., 2012).

The DIF model specifies how predictive surprise determines trial-by-trial P300 amplitudes, seemingly representing barely more than a re-iteration of Donchin's (1981) surprise hypothesis of P300 amplitude fluctuations. However, one should not confuse predictive surprise, as defined by the DIF model, with Bayesian surprise (Ostwald et al., 2012). Bayesian surprise numeralizes the divergence between $P_1(n), \ldots, P_K(n)$ and $P_1(n+1), \ldots, P_K(n+1)$, i.e., the divergence between probability distributions across successive trials which can be computed using the Kullback–Leibler metric (Baldi and Itti, 2010). Bayesian surprise thus quantifies the revision of the internal model of the world, given stimulus $s(n)$, whereas predictive surprise $I(n)$ in (2) refers to the unpredictability of $s(n)$, given the internal model immediately before observing $s(n)$. To conclude, we propose a formal computational model of predictive surprise, along with a strategy for testing the model's ability to predict trial-by-trial P300 amplitudes.

ACKNOWLEDGMENTS

Thanks are due Sandra Mann for her help in data collection and analysis. We also thank Herbert Hoijtink, University of Utrecht, NL, for his valuable suggestions and advice concerning Bayesian model selection.

REFERENCES

Alba, J. W., Chromiak, W., Hasher, L., and Attig, M. S. (1980). Automatic encoding of category size information. *J. Exp. Psychol. Hum. Learn.* 6, 370–378.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* 4, 829–839.

Baldi, P., and Itti, L. (2010). Of bits and wows: a Bayesian theory of surprise with applications to attention. *Neural Netw.* 23, 649–666.

Barceló, F., Periáñez, J. A., and Nyhus, E. (2008). An information theoretical approach to task-switching: evidence from cognitive brain potentials in humans. *Front. Hum. Neurosci.* 1:13. doi:10.3389/neuro.09/013.2007

Barlow, H. (2001). Redundancy reduction revisited. *Network* 12, 241–253.

Blankertz, B., Curio, G., and Müller, K. R. (2002). Classifying single trial EEG: towards brain computer interfacing. *Adv. Neural Inf. Process. Syst.* 1, 157–164.

Claasen, T. A. C. M., and Mecklenbräuker, W. F. G. (1982). On stationary linear time-varying systems. *IEEE Trans. Circuits Syst.* 29, 169–184.

Cowan, N. (2005). *Working Memory Capacity. Essays in Cognitive Psychology*. New York, NY: Psychology Press.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., von Cramon, D. Y., and Engel, A. K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *J. Neurosci.* 25, 11730–11737.

Donchin, E. (1981). Surprise! Surprise? *Psychophysiology* 18, 493–513.

Donchin, E., and Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behav. Brain Sci.* 11, 357–427.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT Press.

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., et al. (2009). Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clin. Neurophysiol.* 120, 1877–1908.

Duncan-Johnson, C. C., and Donchin, E. (1977). On quantifying surprise: the variation of event-related potentials with subjective probability. *Psychophysiology* 14, 456–467.

Egner, T., Monti, J. M., and Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *J. Neurosci.* 30, 16601–16608.

Ford, J. M., Roach, B. J., Miller, R. M., Duncan, C. C., Hoffman, R. E., and Mathalon, D. H. (2010). When it's time for a change: failures to track context in schizophrenia. *Int. J. Psychophysiol.* 78, 3–13.

Friston, K. J. (2002). Beyond phrenology: what can neuroimaging tell us about distributed circuitry? *Annu. Rev. Neurosci.* 25, 221–250.

Friston, K. J. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836.

Friston, K. J., Mattout, J., Trujillo-Bareto, N., Ashburner, J., and Penny, W. D. (2007). Variational free energy and the Laplace approximation. *Neuroimage* 34, 220–234.

Friston, K. J., Penny, W. D., Phillips, C., Kiebel, S. J., Hinton, G., and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: theory. *Neuroimage* 16, 465–483.

Garrido, M. I., Kilner, J. M., Kiebel, S. J., and Friston, K. J. (2009). Dynamic causal modeling of the response to frequency deviants. *J. Neurophysiol.* 101, 2620–2631.

Gonsalvez, C. J., and Polich, J. (2002). P300 amplitude is determined by target-to-target interval. *Psychophysiology* 39, 388–396.

Gratton, G., Coles, M. G. H., and Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalogr. Clin. Neurophysiol.* 55, 468–484.

Harrison, L. M., Bestmann, S., Rosa, M. J., Penny, W., and Green, G. G. R. (2011). Time scales of representation in the human brain: weighing past information to predict future events. *Front. Hum. Neurosci.* 5:37. doi:10.3389/fnhum.2011.00037

Hasher, L., and Zacks, R. T. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. *Am. Psychol.* 39, 1372–1388.

Hintzman, D. L. (1976). "Repetition and memory," in *The Psychology of Learning and Motivation*, Vol. 10, ed. G. H. Bower (New York, NY: Academic Press), 47–87.

Jentzsch, I., and Sommer, W. (2001). Sequence-sensitive subcomponents of P300: topographical analyses and dipole source localization. *Psychophysiology* 38, 607–621.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., and Moore, K. S. (2008). The mind and brain of short-term memory. *Annu. Rev. Psychol.* 59, 193–224.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55, 271–304.

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation for perception and action. *Trends Neurosci.* 27, 712–719.

Kopp, B. (2008). "The P300 component of the event-related brain potential and Bayes' theorem," in *Cognitive Sciences at the Leading Edge*, ed. M. K. Sun (New York, NY: Nova Science Publishers), 87–96.

Leuthold, H., and Sommer, W. (1993). Stimulus presentation rate dissociates sequential effects in event-related potentials and reaction times. *Psychophysiology* 30, 510–517.

Lu, Z. L., Neuse, J., Madigan, S., and Dosher, B. A. (2005). Fast decay of iconic memory in observers with mild cognitive impairments. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1797–1802.

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., et al. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *J. Neurosci.* 28, 12539–12545.

Miller, J. (1998). Effects of stimulus-response probability on choice reaction time: evidence from the lateralized readiness potential. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 1521–1534.

O'Connell, R. G., Dockree, P. M., and Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nat. Neurosci.* 15, 1729–1735.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113.

Ostwald, D., Spitzer, B., Guggenmos, M., Schmidt, T. T., Kiebel, S. J., and Blankenburg, F. (2012). Evidence for neural encoding of Bayesian surprise in human somatosensation. *Neuroimage* 62, 177–188.

Penny, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage* 59, 319–330.

Penny, W. D., Stephan, K. E., Mechelli, A., and Friston, K. J. (2004). Comparing dynamic causal models. *Neuroimage* 22, 1157–1172.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R. Jr., et al. (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology* 37, 127–152.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148.

Prater, J. S., and Loeffler, C. M. (1992). Analysis and design of periodically time-varying IIR filters, with applications to transmultiplexing. *IEEE Trans. Signal Process.* 40, 2715–2725.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–164.

Rauss, K., Schwartz, S., and Pourtois, G. (2011). Top-down effects on early visual processing in humans: a predictive coding framework. *Neurosci. Biobehav. Rev.* 35, 1237–1253.

Ritter, W., and Vaughan, H. G. (1969). Averaged evoked responses in vigilance and discrimination: a reassessment. *Science* 164, 326–328.

Sedlmeier, P. E., and Betsch, T. E. (2002). *Etc. Frequency Processing and Cognition*. New York, NY: Oxford University Press.

Shannon, C. E., and Weaver, W. (1948). The mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423.

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area V1. *J. Neurosci.* 30, 3531–3543.

Squires, K. C., Wickens, C., Squires, N. K., and Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science* 193, 1142–1146.

Strange, B. A., Duggins, A., Penny, W. D., Dolan, R. J., and Friston, K. J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Netw.* 18, 225–230.

Summerfield, C., and Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends Cogn. Sci. (Regul. Ed.)* 13, 403–409.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., and Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science* 314, 1311–1314.

Sutton, S., Braren, M., Zubin, J., and John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science* 150, 1187–1188.

Underwood, B. J. (1969). Attributes of memory. *Psychol. Rev.* 76, 559–573.

Winkler, I., and Czigler, I. (2011). Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. *Int. J. Psychophysiol.* 83, 132–143.

Zacks, R. T., and Hasher, L. (2002). "Frequency processing: a twenty-five year perspective," in *Etc. Frequency Processing and Cognition*, eds P. E. Sedlmeier and T. E. Betsch (New York, NY: Oxford University Press), 21–36.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 October 2012; accepted: 27 December 2012; published online: 08 February 2013.

Citation: Kolossa A, Fingscheidt T, Wessel K and Kopp B (2013) A model-based approach to trial-by-trial P300 amplitude fluctuations. Front. Hum. Neurosci. ***6****:359. doi: 10.3389/fnhum.2012.00359*

APPENDIX

ALTERNATION EXPECTATION OF THE SQU MODEL

This Appendix gives computational details on the expectation w.r.t. alternating stimuli $\check{c}_{\Delta,k}(n)$, as required in (7) and described in words in Squires et al. (1976). The expectation w.r.t. alternating stimuli, and how this expectancy is met by the present stimulus $s(n)$ is given as:

$$\begin{aligned}\check{c}_{\Delta,k}(n) &= u_k(n) \cdot (\min\{N_{\text{alt}} - 2, 0\} + N_{\text{alt}}) \\ &\in \left\{-\left(N_{\text{depth}} - 2\right), \ldots, -3, -2, 0, +2, +3, \ldots, \left(N_{\text{depth}} - 2\right)\right\}\end{aligned} \tag{A1}$$

with $N_{\text{depth}} = 5$. The use of the minimum function ensures that an expectation for alternation requires at least $N_{\text{alt}} = 2$ consecutive previous stimulus alternations. Following Squires et al. (1976), its sign

$$u_k(n) = 2 \cdot |d_k(n) - d_k(n-1)| - 1 \in \{-1, +1\} \tag{A2}$$

is negative ($u_k(n) = -1$) if stimulus $s(n)$ violates the alternation expectation [i.e., $s(n)$ and $s(n-1)$ are identical]. On the other hand, if the alternation expectation is met [i.e., $s(n)$ and $s(n-1)$ differ from each other], the sign is positive: $u_k(n) = +1$. The number of *previous* alternations *in a row* constitutes the amplitude of the expectation, which is given by

$$N_{\text{alt}} = \arg \max_{N'_{\text{alt}} \in \mathcal{N}} \prod_{\nu=1}^{N'_{\text{alt}}} \left[2\, |d_k(n-\nu) - d_k(n-\nu-1)|\right] \tag{A3}$$

with[1] $\mathcal{N} = \{1, 2, \ldots, (N_{\text{depth}} - 2)\}$.

ALTERNATION EXPECTATION OF THE DIF MODEL

This Appendix gives details on the coefficients $\gamma_{\Delta,n-\nu}$ and the probability normalizing constants C and C_Δ. **Figure 9** shows the block diagram of this fourth order finite impulse response filter. In order to reduce the complexity of the model by keeping the number of independent parameters small, only the coefficient $\gamma_{\Delta,2}$ is chosen freely within the range $0 \leq \gamma_{\Delta,2} \leq \gamma_{\Delta,\max}$. The effect of the multiplicative constant C_Δ and additive constant C (as shown in **Figure 1**) is merely to ensure $0 \leq c_{\Delta,k}(n) + 1/C \leq 1$. Thus, $\gamma_{\Delta,\max}$ can be set to an arbitrary value if some constraints are met: The filter coefficients are set following $\gamma_{\Delta,1} = -\gamma_{\Delta,2}$, $\gamma_{\Delta,3} = -\gamma_{\Delta,4}$, and $\gamma_{\Delta,4} = \gamma_{\Delta,\max} - \gamma_{\Delta,2}$. The normalizing constants have to be set according to: $C_\Delta = \gamma_{\Delta,\max} + \gamma_{\Delta,2} + \gamma_{\Delta,4} = 2\gamma_{\Delta,\max}$ and $C = \frac{\gamma_{\Delta,\max} + \gamma_{\Delta,2} + \gamma_{\Delta,4}}{\gamma_{\Delta,\max}} = 2$. We chose $\gamma_{\Delta,\max} = 1$.

ON COUNT FUNCTIONS AND DIGITAL FILTERS (DIF MODEL)

In this Appendix we present some mathematical details of the DIF model, its constants, and its relation to the count functions. We will show that the count functions (11) for $c_{S,k}(n)$ and (13) for $c_{L,k}(n)$ are equivalent to their recursive formulations (12) and (14), respectively. To this end we will transform the block diagram of **Figure 3** to obtain the count functions from the new resulting block diagrams.

Figure A1A shows the block diagram of **Figure 3** according to (14) with some notation omitted. A delay of one trial is represented by $\boxed{T}$, the input signal is defined as $g_k(n)$, the output signal as $c_{L,k}(n)$, $(1 - \gamma_{L,n-1})$ and $\gamma_{L,n-1}$ are time-dependent values. The multiplier $(1 - \gamma_{L,n-1})$ can be moved to the right yielding the block diagram shown in **Figure A1B**. **Figure A1C** shows the block diagram after moving $(1 - \gamma_{L,n-1})$ even further to the right. Finally, when moving the multiplier $\gamma_{L,n-1}/(1 - \gamma_{L,n-1})$ to the right of the lower branch delay unit, the time dependency has to be accounted for as is shown in **Figure A1D**. For the long-term filter we obtain an effective filter coefficient

$$\begin{aligned}\gamma'_{L,n} &= \gamma_{L,n} \frac{1 - \gamma_{L,n-1}}{1 - \gamma_{L,n}} \\ &= e^{-\frac{1}{\beta_{L,n}}} \frac{1 - e^{-\frac{1}{\beta_{L,n-1}}}}{1 - e^{-\frac{1}{\beta_{L,n}}}},\end{aligned} \tag{A4}$$

while for the short-term filter this simplifies to

$$\gamma'_{S,n} = \gamma_S = e^{-\frac{1}{\beta_S}}. \tag{A5}$$

The time-varying discrete-time impulse response of the long-term filter depicted in **Figure A1D** is given as

$$h_{L,\nu}(n) = \epsilon(n - \nu - 1) \frac{1 - \gamma_{L,n-1}}{\gamma'_{L,n}} \prod_{\upsilon=\nu+1}^{n} \gamma'_{L,\upsilon} \tag{A6}$$

with $n (> \nu)$ being the currently observed trial, and ν being the trial when the stimulus initiated the impulse response. We make use of the step function

$$\epsilon(n - \nu - 1) = \begin{cases} 0, & n - \nu - 1 < 0 \\ 1, & n - \nu - 1 \geq 0. \end{cases} \tag{A7}$$

Substituting (A4) in (A6) yields

$$\begin{aligned}h_{L,\nu}(n) &= \epsilon(n - \nu - 1) \frac{1 - \gamma_{L,n}}{\gamma_{L,n}} \prod_{\upsilon=\nu+1}^{n} \gamma'_{L,\upsilon} \\ &= \epsilon(n - \nu - 1) \frac{1}{C_{L,n}} \prod_{\upsilon=\nu+1}^{n} \gamma'_{L,\upsilon}\end{aligned} \tag{A8}$$

with the dynamic normalizing value $C_{L,n} = \frac{\gamma_{L,n}}{1 - \gamma_{L,n}}$. The input-output relation of this time-varying linear system is given by (c.f. Claasen and Mecklenbräuker, 1982; Prater and Loeffler, 1992):

$$c_{L,k}(n) = \sum_{\nu=-\infty}^{\infty} h_{L,\nu}(n)\, g_k(\nu). \tag{A9}$$

[1] As an example, Squires et al. (1976) have given the values of $\check{c}_{\Delta,k}(n)$ for $K = 2$, $N_{\text{depth}} = 5$, and $s(n) = k = 1$ (which shall be denoted as a): $(s(n-4)s(n-3)s(n-2)s(n-1)s(n)) = bbaba$: $\check{c}_{\Delta,k}(n) = +2$, $ababa$: $\check{c}_{\Delta,k}(n) = +3$, $babaa$: $\check{c}_{\Delta,k}(n) = -3$, $aabaa$: $\check{c}_{\Delta,k}(n) = -2$; all other 12 fourth order sequences $xxxxa$ terminating with a yield $\check{c}_{\Delta,k}(n) = 0$ according to the minimum function in (A1).

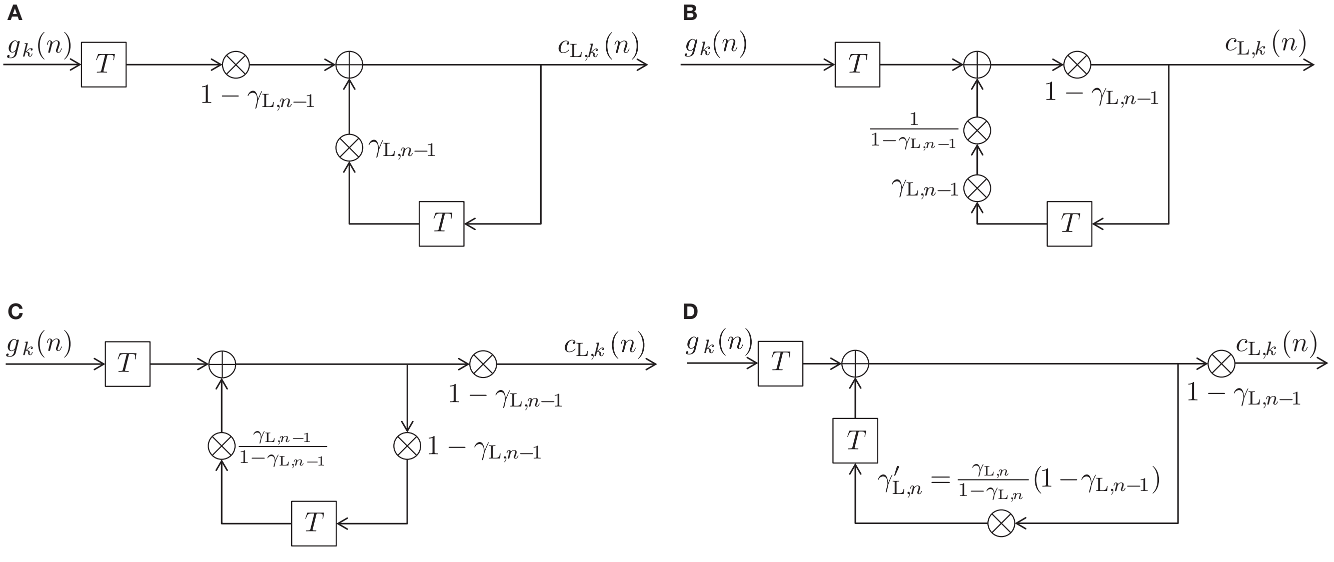


FIGURE A1 | Illustration of the equivalence of the digital filter (14) (Figure 3) to the count function as in (13). (A) Block diagram of the DIF model's long-term memory filter of **Figure 3** and (14) with input signal $g_k(n)$ and output signal $c_{L,k}(n)$. **(B)** Block diagram of a filter equivalent to **(A)**, where the multiplier $(1-\gamma_{L,n-1})$ has been moved to the right. **(C)** Block diagram of a filter equivalent to **(B)**, where the multiplier $(1-\gamma_{L,n-1})$ has been moved even further to the right. **(D)** Block diagram of a filter equivalent to **(C)**, where the multiplier $\gamma_{L,n-1}/(1-\gamma_{L,n-1})$ has been moved to the right of the delay unit in the lower branch.

Due to the step function (A7) in the impulse response (A6) only stimuli at trials $\nu \leq n-1$ contribute to the output at trial n and we obtain

$$\begin{aligned} c_{L,k}(n) &= \sum_{\nu=-\infty}^{n-1} h_{L,\nu}(n)\, g_k(\nu) \\ &= \frac{1}{C_{L,n}} \sum_{\nu=-\infty}^{n-1} \left(\prod_{\upsilon=\nu+1}^{n} \gamma'_{L,\upsilon} \right) g_k(\nu) \qquad \text{(A10)} \\ &= \frac{1}{C_{L,n}} \sum_{\nu=-\infty}^{n-1} \gamma_{L,n}(\nu)\, g_k(\nu), \end{aligned}$$

which is exactly the long-term count function (13) with $\gamma_{L,n}(\nu) = \prod_{\upsilon=\nu+1}^{n} \gamma'_{L,\upsilon} = \prod_{\upsilon=\nu+1}^{n} \gamma_{L,\upsilon} \frac{1-\gamma_{L,\upsilon-1}}{1-\gamma_{L,\upsilon}}$, and $g_k(\nu)$, as before. Note that for the short-term filter we obtain $\prod_{\upsilon=\nu+1}^{n} \gamma_S = \gamma_S^{n-\nu}$ and $C_S = \gamma_S/(1-\gamma_S)$, which simplifies (A10) to the discrete-time convolution

$$c_{S,k}(n) = \tfrac{1}{C_S} \sum_{\nu=-\infty}^{n-1} \gamma_S^{n-\nu} g_k(\nu), \qquad \text{(A11)}$$

which is the short-term count function (11). Thus, equivalence has been shown between the count functions (11) and (13), and their simple recursive formulations (12) and (14), respectively. The behavior of the dynamic long-term time value $\beta_{L,n}$ follows:

$$\beta_{L,n} = \begin{cases} e^{-\left(\frac{1}{\tau_1}\cdot 1+\frac{1}{\tau_2}\right)}, & \text{if } n \leq 0 \\ e^{-\left(\frac{1}{\tau_1}\cdot n+\frac{1}{\tau_2}\right)}, & \text{otherwise,} \end{cases} \qquad \text{(A12)}$$

with normalized time constants τ_1, τ_2, controlling the speed of transition from reliance on prior assumptions to experience. The time value $0 \leq \beta_{L,n} < \infty$ holds $\beta_{L,n} > \beta_S$. The effect of this dynamic formulation of $\beta_{L,n}$ is further illustrated in **Figure A2** which shows the values of $\beta_{L,n}$ and the corresponding forgetting factor $\gamma_{L,n} = e^{-\frac{1}{\beta_{L,n}}}$ for trials $n \in \{1, \ldots N\}$ with $N = 192$.

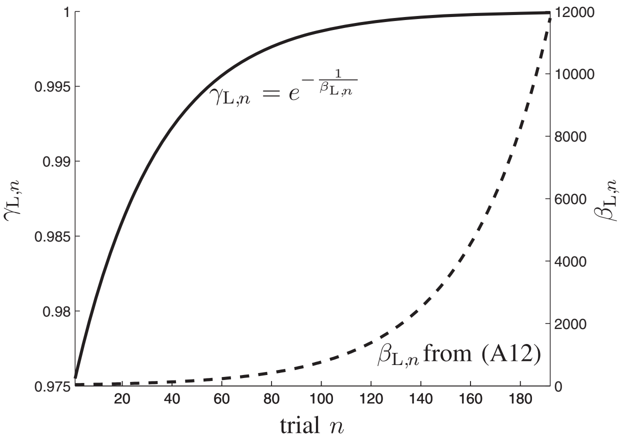


FIGURE A2 | The dynamics of the coefficients $\gamma_{\mathrm{L},n}$ (solid) and $\beta_{\mathrm{L},n}$ (dashed) over trials $n = 1, \ldots, N$, with $N = 192$.