frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 27 February 2017
doi: 10.3389/fnhum.2017.00075

Variation in Event-Related Potentials by State Transitions

Hiroshi Higashi[1]*, *Tetsuto Minami*[1,2] and *Shigeki Nakauchi*[1]

[1] Department of Computer Science and Engineering, Toyohashi University of Technology, Aichi, Japan, [2] Electronics-Inspired Interdisciplinary Research Institute, Toyohashi University of Technology, Aichi, Japan

The probability of an event's occurrence affects event-related potentials (ERPs) on electroencephalograms. The relation between probability and potentials has been discussed by using a quantity called surprise that represents the self-information that humans receive from the event. Previous studies have estimated surprise based on the probability distribution in a stationary state. Our hypothesis is that state transitions also play an important role in the estimation of surprise. In this study, we compare the effects of surprise on the ERPs based on two models that generate an event sequence: a model of a stationary state and a model with state transitions. To compare these effects, we generate the event sequences with Markov chains to avoid a situation that the state transition probability converges with the stationary probability by the accumulation of the event observations. Our trial-by-trial model-based analysis showed that the stationary probability better explains the P3b component and the state transition probability better explains the P3a component. The effect on P3a suggests that the internal model, which is constantly and automatically generated by the human brain to estimate the probability distribution of the events, approximates the model with state transitions because Bayesian surprise, which represents the degree of updating of the internal model, is highly reflected in P3a. The global effect reflected in P3b, however, may not be related to the internal model because P3b depends on the stationary probability distribution. The results suggest that an internal model can represent state transitions and the global effect is generated by a different mechanism than the one for forming the internal model.

Keywords: Event-Related Potentials (ERPs), Electroencephalography (EEG), predictive surprise, model-based analysis, single-trial analysis

OPEN ACCESS

Edited by:
Juliana Yordanova, Bulgarian Academy of Sciences, Bulgaria

Reviewed by:
Rolf Verleger, University of Lübeck, Germany
Márk Molnár, Hungarian Academy of Sciences, Hungary

***Correspondence:**
Hiroshi Higashi
higashi@tut.jp

Received: *16 November 2016*
Accepted: *07 February 2017*
Published: *27 February 2017*

Citation:
Higashi H, Minami T and Nakauchi S (2017) Variation in Event-Related Potentials by State Transitions. Front. Hum. Neurosci. 11:75. doi: 10.3389/fnhum.2017.00075

1. INTRODUCTION

Humans make predictions by using prior information (Doya et al., 2007; Friston, 2008), and the prior information is derived from what humans have experienced. The prediction of an event's occurrence is equivalent to the estimation of the generative model for the event (Robert, 2007). However, the manner in which one utilizes that experience in estimating the generative model remains unclear.

One approach for revealing how experience affects human prediction is the observation of event-related potentials (ERPs). ERPs are the measured brain responses for a specific internal or external event (Clark, 2013). Some ERP components observed on electroencephalograms (EEGs) are affected by the probability of the occurrence of the event (Sutton et al., 1965; Squires et al., 1977; Picton, 1992). In particular, a slow variation observed about 300 ms after the event on

the EEG potentials is called P300, and its peak amplitude depends on the probability of the event's occurrence (Duncan-Johnson and Donchin, 1977). Such ERP components are considered to reflect the process for predicting events and are widely used as a medium for analyzing human cognition (Horovitz et al., 2002; Sanmiguel et al., 2013).

The relation between these ERP components and probability has been discussed by using a quantity called *the degree of surprise* or simply *surprise* (Ostwald et al., 2012). One concept of surprise is called *predictive surprise* that represents the subjective self-information, information content, or surprisal (Shannon, 1948) that an observer receives from an observed event (Donchin and Coles, 1988). Recently, Mars et al. (2008) and Kolossa et al. (2012) addressed the question of which factors in a preceding stimulus sequence affect predictive surprise to a present stimulus by investigating the relation between the stimulus sequence and P300 properties. To identify these factors, Mars et al. (2008) and Kolossa et al. (2012) used regression models in which the input was a stimulus history and the output was the P300 amplitude. Their approach with these regression models is called *the model-based approach*. Assuming that the observed brain activities reflect the prediction process, the model-based approach can confirm which factors human prediction depends on by finding a model that accurately estimates the amplitude of P300. Their results suggest that surprise estimated by the integration of three factors (long-term history, short-term history, and alternating expectations of the stimulus sequence) adequately predicts the P300 amplitude (Squires et al., 1976; Kolossa et al., 2012).

The other concept of surprise, proposed by Baldi and Itti (2010), is called *Bayesian surprise* that represents the degree of updating in the beliefs of an observer who experiences a new event. Recent studies (Kolossa et al., 2015; Seer et al., 2016) showed that predictive and Bayesian surprises affect different subcomponents of P300 called P3a and P3b (Polich, 2007). Predictive surprise better explains P3b, which has a long latency among the subcomponents. However, Bayesian surprise better explains P3a, which has a short latency. The results suggest that the subcomponents reflect distinct neural mechanisms for prediction.

To reveal the relation between ERPs and surprise, theoretical frameworks, such as the context-updating model (Donchin, 1981; Donchin and Coles, 1988; Polich, 2007), predictive coding (Friston, 2002; Spratling, 2010), and Bayesian brain hypothesis (Hampton et al., 2006; Kopp, 2006; Ostwald et al., 2012; Lieder et al., 2013), have been convincingly established. The frameworks explain human behavior or brain responses by positing the existence of an internal model that humans constantly and automatically generate about the external world (Donchin, 1981). Different processes of the response of the internal model to an external event lead to different brain activities, such as the P3a and P3b variations (Kolossa et al., 2015).

What state transition the internal model builds is discussed in this study. Previous studies, such as Mars et al. (2008) and Kolossa et al. (2012), assumed that the internal model is without state transitions. Accordingly, surprises were estimated based on a generative model in a stationary state (a stationary-state model). In contrast, the purpose of our study is to find evidence of a brain mechanism that codes state transitions. If the brain can generate an internal model with state transitions (the state transition model), then humans would not acquire a probability distribution of events but would acquire a model that describes how different states or situations of the world are connected to each other (Gläscher et al., 2010).

The possibility that the state transition models explain some effects in ERP components motivated this study. These properties of an event sequence, such as stationary-state models, alternation, and repetition (Matt et al., 1992; Rac-Lubashevsky and Kessler, 2016) that explain the variation in some ERP components can be generalized with a state transition model. Moreover, Gläscher et al. (2010) suggested that probability distributions with state transitions are coded in the brain during the performance of reinforcement learning tasks (Saito et al., 2015). Therefore, we hypothesized that, for prediction, a mechanism for coding state transitions exist; that is, surprise is modeled not only with the probability distribution of the stationary state but also with the probability distribution with state transitions.

In the present study, we investigated the relation between predictive surprise in a generative model that has state transitions and electrophysiological signals via a model-based analysis. To distinguish predictive surprises in the different state models, predictive surprise with state transitions is called *predictive transition surprise*, and predictive surprise in a stationary state is called *predictive stationary surprise*. Although the EEG signals were recorded with a two-choice response time task, the same one used by Mars et al. (2008) and Kolossa et al. (2012), the generative models for the event sequences were different. In previous studies, predictive transition surprise converged with predictive stationary surprise as the number of trials increased; therefore, that type of setting cannot isolate the effects of predictive transition surprise. To avoid this situation, we used state transition models for the generative models; we controlled generation of the event sequence with a simple Markov chain (Norris, 1998). In the model-based analysis, we used predictive stationary or transition surprise of the Markov chain as the explanatory variable. As the response variable, we used the EEG potentials, which were observed in various electrodes and latencies. This analysis enabled us to visualize the effects of these surprises on various ERP components, such as P3a and P3b. The results show different brain activities that seem to be associated with the stationary-state model and the state transition model.

2. MATERIALS AND METHODS

2.1. Measurement

2.1.1. Participants

Twelve individuals (10 male and 2 female) participated in the experiment. Their ages ranged from 21 to 27 years ($M = 23.6$; $SD = 1.7$). The participants had normal or corrected-to-normal visual acuity. All participants provided written informed consent. The experimental protocols were approved by the Committee for Human Research at the Toyohashi University of Technology,

Aichi, Japan, and the experiment was conducted in accordance with the committee's approved guidelines.

2.1.2. Experimental Design

The participants performed a two-choice response time (TCRT) task (**Figure 1**) without feedback about response accuracy (Mars et al., 2008; Kolossa et al., 2012): Two visual stimuli were presented about every 1.5 s, and the participants were required to respond to each stimulus with the previously associated button as quickly as possible. Visual stimuli were presented on an LCD display [VIEWPixx EEG (VPixx Technologies)] with Psychtoolbox-3 and MATLAB R2011b (The MathWorks, Inc.). The participants were seated in front of the display. They touched their left or right hand to the left or right button of a four-button trackball on the desk between the participant and the display.

A single trial consisted of the following procedures. First, the participant gazed at the fixation cross ($2.0^\circ \times 2.0^\circ$) presented at the center of the display for 400–600 ms. Then, the fixation cross vanished, and a circle or a square ($2.0^\circ \times 2.0^\circ$) was presented for 200 ms at the location where the fixation cross had originally been presented. The participant clicked the left or right button as quickly as possible. The participant's response was accepted from 200 ms to 1000 ms after the symbol was presented. If the participant did not respond during this period, then the trial was recorded as a no-response trial. The fixation cross was presented when the symbol vanished.

The symbol-response assignment and instructions were given to the participants before the experiment. For example, participants were instructed "to click the left (right) button when the circle (square) appears." Before the measurement blocks, the participants underwent a training blocks for the response task consisting of 50 trials; the training blocks was repeated until the participants' response accuracy reached 90%. During the training blocks, the stimulus sequences were generated randomly with a uniform probability distribution.

The symbol stimuli (circle and square) were generated with simple Markov chains. There were two conditions (*C1* and *C2*) with different Markov chains. Let Event *a* and Event *b* correspond to either the circle or the square, respectively, and let E_n be the present event and E_{n-1} be the preceding event. This assignment is denoted as the event-symbol assignment. Condition *C1* can be represented with the transition probabilities $P(E_n \mid E_{n-1})$ as $P(a \mid a) = P(b \mid b) = 0.3$ and $P(b \mid a) = P(a \mid b) = 0.7$. For Condition *C2*, $P(a \mid a) = 0.3$, $P(b \mid a) = 0.7$, and $P(a \mid b) = P(b \mid b) = 0.5$. The Markov chains for the two conditions are summarized in **Figure 2**.

A single block consisted of 300 trials, and the participants executed a total of four blocks. Two blocks were of Condition *C1*, and two were of Condition *C2*. The event sequence was the same for the two blocks that had the same condition. The participants took a break of at least two minutes between blocks. The participants were not told that there were two conditions for the generative model, and according to a question that we asked the participants after the experiment, none noticed that there were two conditions.

The symbol-response and event-symbol assignments and the block orders were decided randomly as follows. The symbol-response assignment was different for each participant. The event-symbol assignment was chosen randomly during the blocks. The order of the blocks with the two conditions was random.

2.1.3. EEG Acquisition

The EEG signals were recorded using a BioSemi ActiveTwo system. The EEG recording was performed at a sampling rate of 512 Hz with a 64-electrode cap, referenced to the common mode sense (CMS) active electrode. The 64 active electrodes were positioned to cover the whole head according to the extended International 10/10 system. The signals in the electrodes placed on the left and right earlobes, on the right side of the right eye (on the temple), and at the left, upper, and lower sides of the left eye were also measured. For preprocessing, the signals were re-referenced with the averaged potential of both earlobes. Moreover, a Butterworth bandpass filter (passband: 0.3–30 Hz, order: 4) was applied to the signals. Epochs were corrected using the −100 to 0 ms period as the baseline. The epochs in which the vertical electroculograms were more than $\pm 80\ \mu$V were removed.

2.2. Analysis of Behavior and EEG

In our analysis, the symbols *aa*, *ab*, *ba*, and *bb* represent the data for the present stimulus after the preceding stimulus. For example, *ab* represents the data for Event *b* after Event *a*.

For the behavioral data, the clicked buttons and the response times of the participants' responses for all trials were acquired. The trials in which the participants responded with the wrong button were removed from the analysis of the response time. We tested the averaged behavioral data for each participant with a two-way repeated-measures analysis of variance (ANOVA) (Cohen et al., 2003; Rac-Lubashevsky and

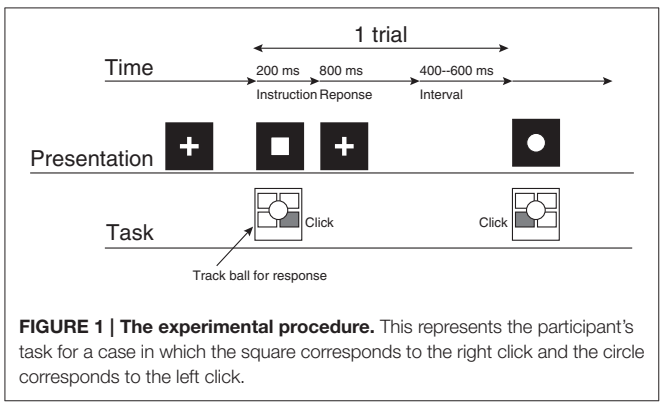

FIGURE 1 | The experimental procedure. This represents the participant's task for a case in which the square corresponds to the right click and the circle corresponds to the left click.

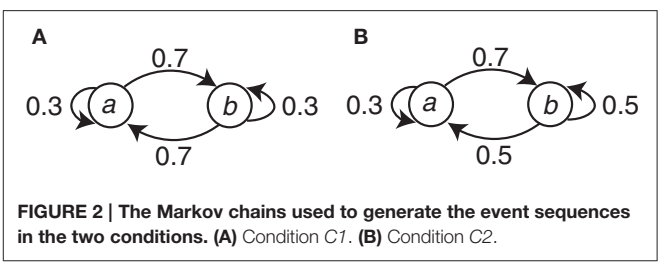

FIGURE 2 | The Markov chains used to generate the event sequences in the two conditions. (A) Condition *C1*. **(B)** Condition *C2*.

Kessler, 2016) with the factors *Present* (the present stimulus with two levels: Event *a* and Event *b*) and *Preceding* (the preceding stimulus with two levels: Same and Different as the present stimulus). The combinations of the two factors resulted in the sequences *aa* for (Event *a*, Same), *ba* for (Event *a*, Difference), *bb* for (Event *b*, Same), and *ab* for (Event *b*, Difference).

For conventional analysis of ERPs, the trials in which the participants responded with the wrong button were removed from the analysis. We tested the averaged EEG potential at each channel and latency for each participant with an ANOVA in which the factors were the same as those for the behavioral data.

2.3. Model-Based Analysis

For finding the time periods and electrodes in the EEG signals that are well explained by trial-by-trial surprises, we used a model-based analysis of regression with a generalized linear model (GLM) (Bolker et al., 2009). Trial-by-trial surprises were estimated based on the preceding series of the stimulus. The two types of surprise were compared in their effects on the EEG potentials: surprise generated by stationary-state models and by state transition models. A model-based analysis evaluates the relation between two variables with an indicator representing how much one variable accurately explains and/or predicts the other. In this analysis, the variable used to explain the other variable is called the explanatory variable, and the variable to be explained is called the response variable (Haykin, 2005; Dobson and Barnett, 2011).

Surprise concerning the present event (called predictive surprise in Kolossa et al., 2012, 2015) was used as the explanatory variable. Predictive stationary surprise is defined as the logarithm of the overall probability given the preceding series of events. **Figure 3** shows the trial-by-trial change in predictive stationary surprise for each condition. Additionally, predictive transition surprise is defined as the logarithm of the probability transitioning from the preceding event to the present event given the preceding series of events. **Figure 4** shows the trial-by-trial change in predictive transition surprise for each condition. The detailed definitions of surprises are described in Section A.1.

The EEG potentials of each trial were used as the samples of the response variable for the model-based approach. Before the potentials were extracted, the EEG signals over the two blocks of the same condition in each trial were averaged. If either sample in the two blocks was missing because of a wrong task response, no response, or artifact rejection, those trials were removed from the analysis (Kolossa et al., 2012). The trial-by-trial ERPs were extracted by averaging the preprocessed EEG signals over a temporal window of ±50 ms around every 20 ms from 0 to 700 ms from the onset of the stimulus.

The samples of Conditions *C1* and *C2* were merged into a sample set. The merging reduced specific effects of the generative models, such as alternation expectation (Squires et al., 1977; Mars et al., 2008; Kolossa et al., 2012). The number of the samples for each channel and latency was 5578 (the participants' mean = 464.83; and $SD = 44.56$).

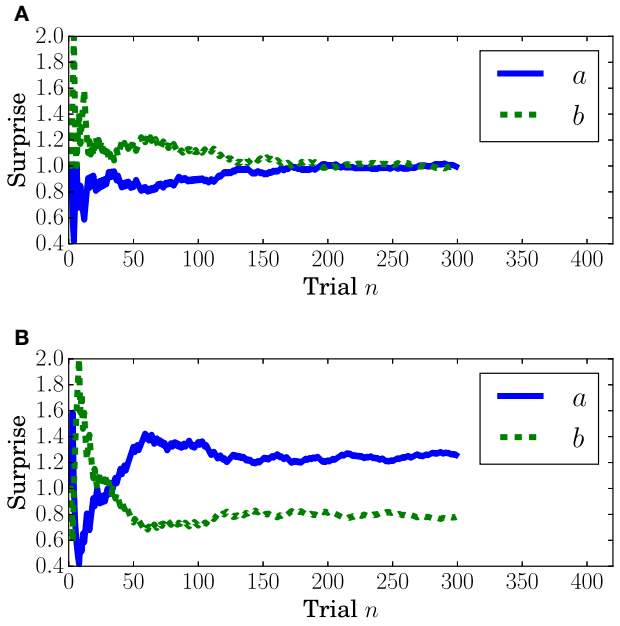

FIGURE 3 | Predictive stationary surprise I_S at the nth trial. (A) Condition *C1*. **(B)** Condition *C2*.

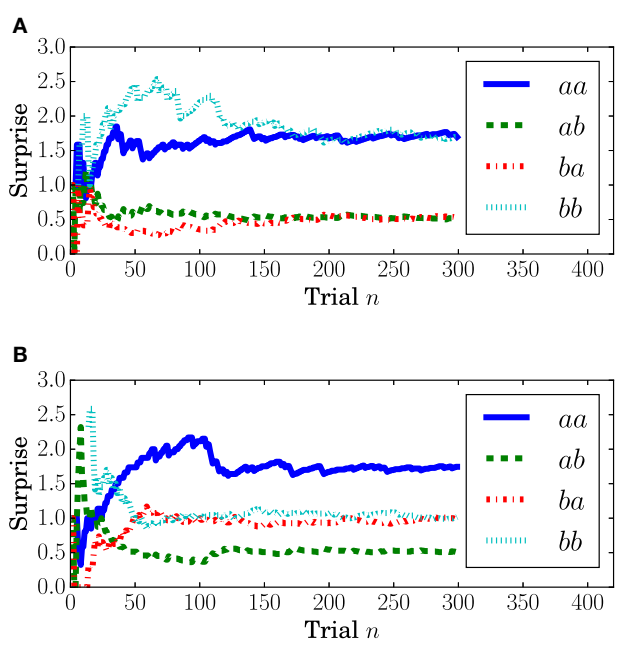

FIGURE 4 | Transition stationary surprise I_T at the nth trial. (A) Condition *C1*. **(B)** Condition *C2*.

A GLM (Bolker et al., 2009) was adopted for the regression of the explanatory and response variables. The model is summarized in **Figure 5**. The N observed samples of the set of explanatory and response variables ($N = 5578$) are represented

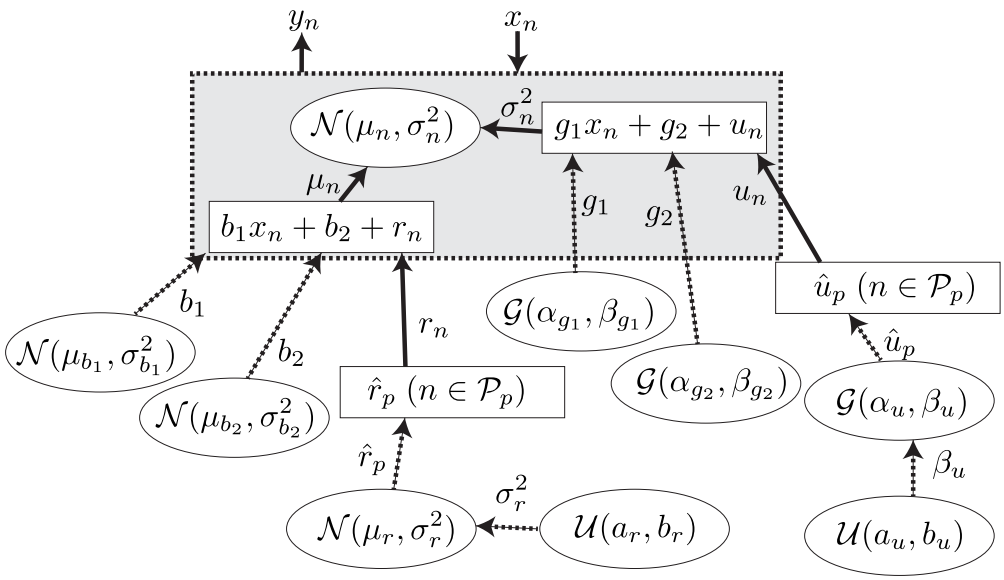

FIGURE 5 | The graphical model for the model-based analysis. The variables, the values of which are determined stochastically, are represented by the dashed lines. The deterministic variables are represented by the solid lines. The distributions surrounded by circles represent the non-informative prior probability distributions.

as $\{x_n, y_n\}_{n=1}^N$, which corresponds to predictive surprise and the EEG potential at a channel and time period for the nth sample. The EEG potential is modeled by the linear model of estimated surprise with consideration of individual differences formulated as $\mu_n = b_1 x_n + b_2 + r_n$ and an additive Gaussian noise, $\mathcal{N}(0, \sigma_n^2)$, with the variance formulated as $\sigma_n^2 = g_1 x_n + g_2 + u_n$. The unknown parameters were found with the maximum log-likelihood method (Gelman et al., 2013). Details of the model and the fitting procedure are given in Section A.2.

The fitting accuracy of the regression model was evaluated using the log-Bayes factor of the estimated model. The model $\mathcal{M}_\mathrm{S}$ was estimated with predictive stationary surprise, and $\mathcal{M}_\mathrm{T}$ was estimated with predictive transition surprise. Moreover, a common reference model $\mathcal{M}_\mathrm{NULL}$ was also estimated with a set in which all samples for the response variables were 1 (Neyman and Pearson, 1933; Kolossa et al., 2015). The log-likelihood of $\mathcal{M}_M$ denoted by $\log L_M$ for M is S, T, or NULL. As an indicator for the fitting accuracy, the log-Bayes factor B_M with the common reference model (Kass and Raftery, 1995; Kolossa et al., 2015) was adopted:

$$B_M = \log L_M - \log L_\mathrm{NULL}, \tag{1}$$

for M is S or T.

The log-Bayes factor was evaluated with a likelihood-ratio test (Neyman and Pearson, 1933) that evaluates how more accurately the model fits than the common reference model. A parametric bootstrap method (Davison and Hinkley, 1997) (the number of sampling = 1000) was used for the test.

3. RESULTS

3.1. Behavioral Data

The response time for the button-clicking task was defined as the duration between the display of the stimulus and the clicking of the button. Mean values are displayed in **Figure 6**. The ANOVA showed a main effect of the factor *Present* in Condition *C2* [$F_{(1,\,11)} = 15.8042$, $p = 0.0022$].

The response accuracy for the button-clicking task was defined as whether or not the participant clicked the assigned button correctly. Mean values are displayed in **Figure 7**. The ANOVA showed a main effect of the factor *Preceding* in Condition *C1* [$F_{(1,\,11)} = 16.5023$, $p = 0.0019$]. In Condition *C2*, main effects were found for the factors *Present* [$F_{(1,\,11)} = 10.1730$, $p = 0.0086$] and *Preceding* [$F_{(1,\,11)} = 6.5105$, $p = 0.0269$]. An interaction of the two factors [$F_{(1,\,11)} = 6.2011$, $p = 0.03$] was also found in Condition *C2*. Simple effects for the interaction were found for *Present* at the level Same [$F_{(1,\,11)} = 14.8977$, $p = 0.0027$] and for *Preceding* at the level Event *a* [$F_{(1,\,11)} = 14.9957$, $p = 0.0026$].

The results of the statistical analysis suggest that the behavior (response time and accuracy) was affected by state transitions. The main effect of *Preceding* on response accuracy in Condition *C1* reflects the high transition probability for Sequences *ab* and *ba* in the generative model. The difference in the stationary probability between Events *a* and *b* in Condition *C2* can explain the main effects by *Present* for the response time and accuracy. In the response accuracy, the simple effect by *Present* at the level Same corresponds to the difference in the transition probabilities between Sequences *aa* (0.3) and *bb* (0.5) in Condition *C2*. The simple effect by *Preceding* at the level Event *a* corresponds to the difference between Sequences *aa* (0.3) and *ba* (0.5). The differences in the transition probabilities between Sequences *ab*

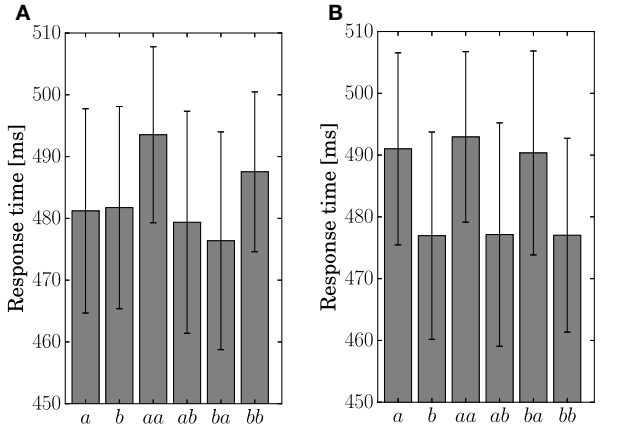

FIGURE 6 | Response time for the single stimulus or sequence. The error bars show the standard error. **(A)** Condition *C1*. **(B)** Condition *C2*.

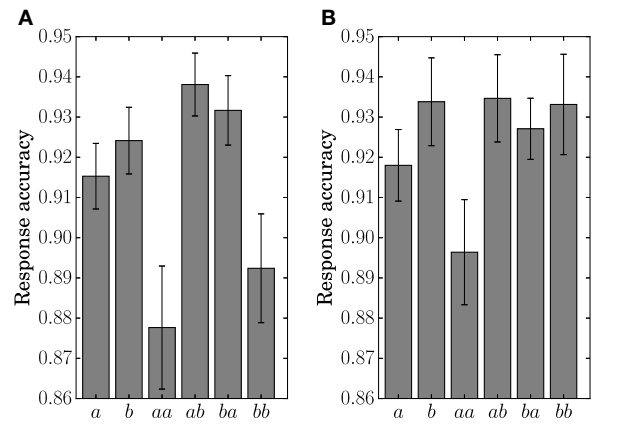

FIGURE 7 | Response accuracy for the single stimulus or sequence. The error bars show the standard error. **(A)** Condition *C1*. **(B)** Condition *C2*.

(0.7) and *bb* (0.5), and Sequences *ab* (0.7) and *ba* (0.5) in the generative model of Condition *C2*, however, did not appear in behavior.

3.2. Event-Related Potentials

Figure 8 depicts the grand-averaged ERP waveforms. The ANOVA showed an effect of *Preceding* in FCz (Conditions *C1* and *C2*) and CPz (Condition *C1*) at a latency around 340–400 ms. An effect of *Present* was found in CPz, Condition *C2* at a latency around 370–400 ms. In **Figure 8A**, the peak amplitude for the sequences that have a transition probability of 0.3 (*aa* and *bb*) is higher than that for those that have a transition probability of 0.7 (*ab* and *ba*). Sequence *aa* in Condition *C2*, which has a transition probability of 0.3, leads to the highest peak amplitude, as shown in **Figures 8B,D**.

The peaks of P300 are at around 400 ms, which are 100 ms later than the peak latencies reported by Kolossa et al. (2012), who employed a color discrimination task. This difference could be caused by the difference in the stimulus features that an observer should detect for the TCRT tasks (Smid et al., 1999). Mars et al. (2008), who employed a shape discrimination task, reported a similar peak latency as in this analysis, around 400 ms.

3.3. Model-Based Analysis

Figure 9 displays the log-Bayes factors. We found high log-Bayes factors (shown in red), which indicate high fitting accuracy, in some channels and latencies.

For the stationary-state model $\mathcal{M}_S$, the likelihood-ratio test showed that the models that reached a log-Bayes factor ≥ 1487 fitted significantly more accurately than the common reference model ($p < 0.05$). The latencies at the channels FCz and CPz in which the statistically significant differences were found are shown as the red bars with the ERP waveforms in **Figure 8**. At FCz, effects are found within 300–360 and 500–580 ms. At CPz, effects are found within 380–400 and 480–600 ms.

For the state transition model $\mathcal{M}_T$, the likelihood-ratio test showed that the models that reached a log-Bayes factor ≥ 2241 were statistically significant ($p < 0.05$). The latencies at the channels FCz and CPz in which the statistically significant differences were found are shown as the green bars with the ERP waveforms in **Figure 8**. At FCz, an effect is found within 340–480 ms. At CPz, an effect is found within 340–440 ms.

In **Figure 10A**, which shows the change in the log-Bayes factors according to the stage of the trials, an increase at the middle stage and a decrease at the last stage in the log-Bayes factors for $\mathcal{M}_T$ ($t = 360$ ms) were observed at FCz and Cz. **Figure 10B** shows that the log-Bayes factor for $\mathcal{M}_S$ at Cz and CPz increases as the trials accumulate.

4. DISCUSSION

This study investigated the effects of predictive stationary surprise and predictive transition surprise on EEG potentials under the assumption that the internal model is formed with state transitions and predictive surprise is based not only on a stationary-state model but also on a state transition model. For this, we applied Markov chains to generate event sequences in order to isolate the effects of stationary and transition surprises. The results show that predictive stationary surprise better explains P3b and predictive transition surprise better explains P3a. This suggests two distinct mechanisms in human prediction. The effect of predictive transition surprise on P3a suggests that a mechanism for estimating the generative model exists and that the internal model forms a state transition model. The result also indicates a mechanism for processing a stationary-state model as observed by the variability of P3b. The dependencies on time (the number of observed events) of these effects could reflect the process to form the observer's prediction.

We adopted a simple procedure in which predictive surprise was estimated as the self-information of the present event. The self-information for the event was estimated from the preceding event sequences. This procedure is equivalent to the procedure proposed by Mars et al. (2008). Kolossa et al. (2012) improved the model as the digital filtering (DIF) model. However, the optimization problem for the parameters in the DIF model is very complex, and the optimization needs to use an empirical

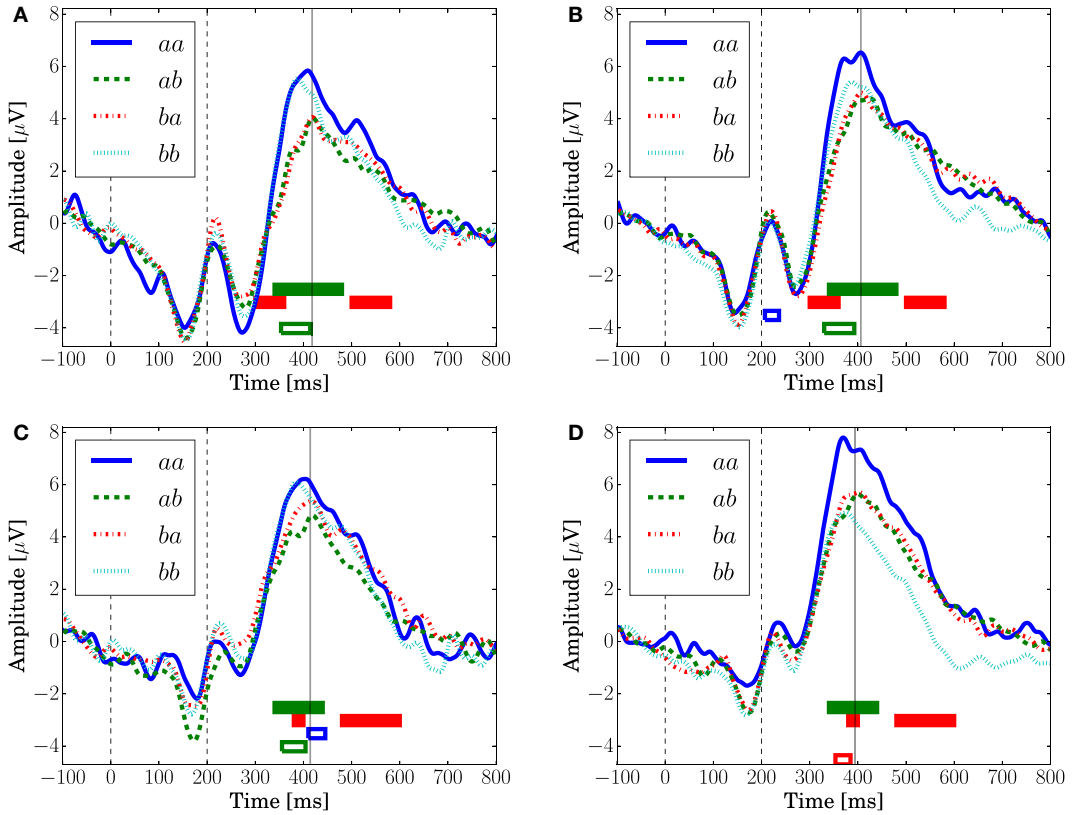

FIGURE 8 | The grand-averaged EEG potentials observed at the channels FCz and CPz for each sequence with two successive stimuli. The vertical dashed lines located at 0 and 200 ms represent the onset and offset of the symbol stimulus. The vertical solid lines represents the latency when the maximum amplitude was observed. The time periods in which the main effects are found in the ANOVA at each channel are shown by the red (*Present*), green (*Preceding*), and blue (interaction) unfilled bars ($p < 0.05$). The time periods in which the main effects are found in the log-Bayes factors at each channel are shown by red (the stationary-state model $\mathcal{M}_S$) and green (the state transition model $\mathcal{M}_T$) filled bars ($p < 0.05$). **(A)** FCz (Condition *C1*). **(B)** FCz (Condition *C2*). **(C)** CPz (Condition *C1*). **(D)** CPz (Condition *C2*).

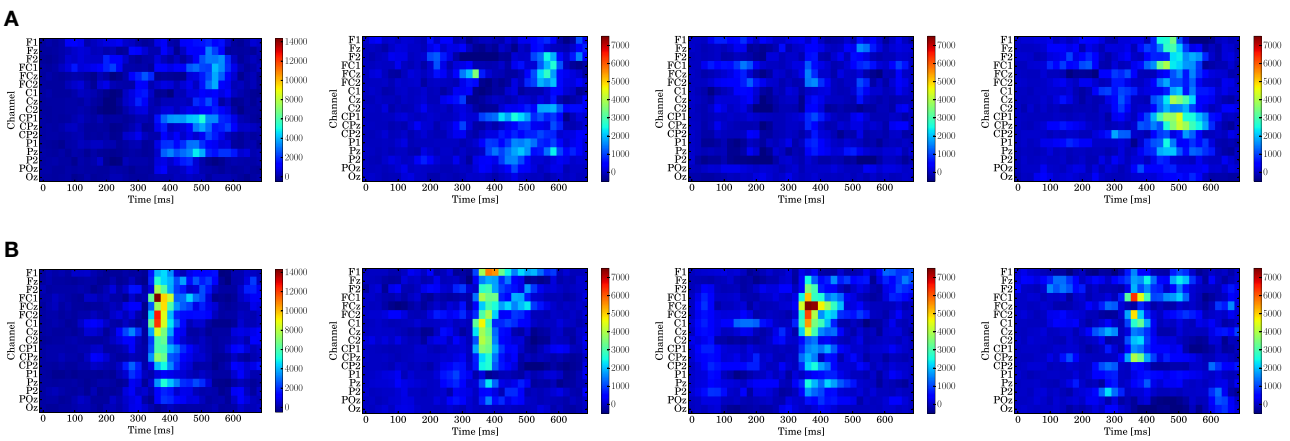

FIGURE 9 | Log-Bayes factors B_M for each stage of the trials (left to right: whole trials (1st–300th) and early (1st–100th), middle (101st–200th), and last (201st–300th) stages). The factor decreases as the color turns from red to blue (see the color bar beside each figure). **(A)** Model based on predictive stationary surprise $\mathcal{M}_S$. **(B)** Model based on predictive transition surprise $\mathcal{M}_T$.

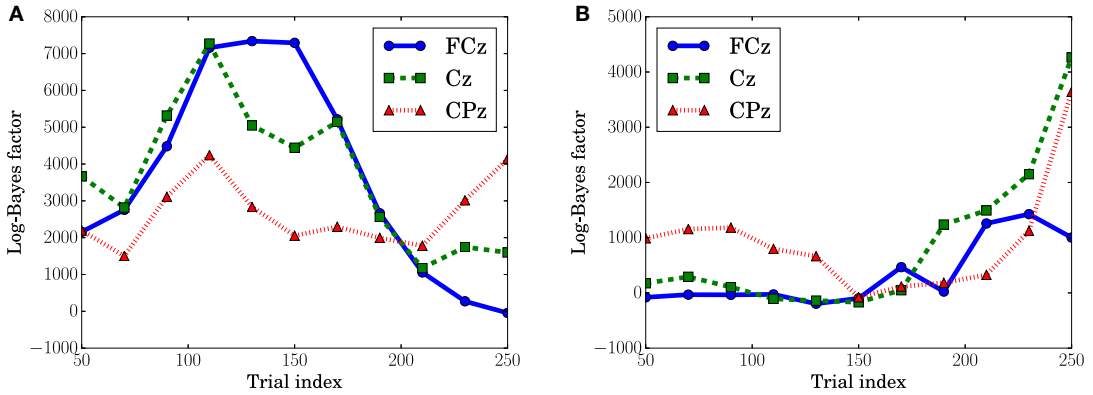

FIGURE 10 | Log-Bayes factors B_M at the channels, FPz and CPz. The factors were derived using the ±50th trials from the index of the *x*-axis (e.g., for the trial index 150, the 100th to 200th trials were used). **(A)** Predictive transition surprise, $t = 360$ ms. **(B)** Predictive stationary surprise, $t = 480$ ms.

procedure, which does not have the guarantee of a global optimum. Since we focused on the effects on the brain activity that differs between the stationary-state and state transition probabilities, we adopted a fairly simple model, one that does not have any parameters that need to be optimized. Moreover, the DIF model does not accurately produce surprise associated with the state transition model because it is based on a linear combination of the three factors.

The results show that the behavioral data (response accuracy and response time), ERP waveforms, and log-Bayes factors depend on predictive stationary surprise. The behavioral data results correspond to those of Miller (1998) and Kolossa et al. (2012). In the ERP results, as Polich (2007) suggested, the ERP at 390 ms in the centro-parietal region dependent on the stationary probability can be observed. From the model-based analysis, the high fitting accuracy with a centro-parietal focus within 480–600 ms can be considered to be a result of a variation in the P3b component. This speculation is supported by Kolossa et al. (2015), who suggested that P3b is more strongly associated with predictive stationary surprise than P3a.

The effects of predictive transition surprise can be seen in the present results. The behavioral results suggest that response accuracy and response time depend on the preceding event even if the present event is the same: The difficulty of the response depends on the transition probability distribution. The feature observed in the ERP waveforms (**Figure 8**), that high transition surprise leads to a high peak, is similar to ERP responses to stationary surprise. We suggest here that the effect of the transition probability on behavior and ERPs has not been revealed clearly. In the model-based analysis, the high fitting accuracy with a central focus within 340–480 ms can be considered to be caused by a variation in P3a because similar features in its area (Kopp and Lange, 2013) and latency (Kolossa et al., 2015) have been reported.

The dependence of the P3a component on predictive transition surprise suggests that the participants estimated state transition models as the generative model. This can be explained by introducing Bayesian surprise. Kolossa et al. (2015) showed that Bayesian surprise yields a superior model for explaining the variation in P3a distributed in the fronto-central region. This suggests that the variation in P3a occurs via the updating of the internal model. Because the P3a, and thus the update, can be modeled better by predictive transition surprise than by predictive stationary surprise in this experimental setting, it appears that the internal model is associated more strongly with a model with state transitions than with a stationary model. Namely, the internal model approximates a Markov chain. This speculation is supported by the decrease in the fitting accuracy for P3a in $\mathcal{M}_{\mathrm{T}}$ at the last stage (**Figure 10A**) because the internal model converges by accumulating the event observations and Bayesian surprise is slight in the last stage. This convergence corresponds to the convergence in predictive transition surprise shown in **Figure 4**.

The effect of predictive stationary surprise shows that human prediction has a mechanism different from that of the generative model. Although the internal model is built based on the state transition model, the P3b components depend on the stationary probability distribution. This result suggests that P3b is not affected by the state transitions and mainly reflects the stationary state. The effect of stationary-state models on P3b has been confirmed by El Karoui et al. (2015) and Bekinschtein et al. (2009) as *the global effect*. The increase of the fitting accuracy for P3b at the last stage is consistent with a feature of the global effect that is related to the accumulation of an event on longer time scales (d'Acremont et al., 2013; El Karoui et al., 2015).

Although Kolossa et al. (2015) showed that P3b is distributed in the centro-parietal region, the effect of the stationary-state model is observed also in fronto-central region. This effect could be caused by a P300 latency shift that novelty detection (Courchesne et al., 1975; Knight, 1984) and attention (Kahneman, 1973) affect. This hypothesis is supported by the effect observed in both time periods of the ascending and descending flanks of P300 as shown in **Figures 8A,B**.

The electrophysiological effects of the two generative models we tested in this study, the stationary-state and state transition models, support theoretical frameworks regarding ERPs, such as

the context-updating model, predictive coding, and the Bayesian brain hypothesis. Moreover, the effects suggest the following hypotheses about what kind of models the brain adopts as the internal model. (1) If the external generative model has state transitions, then the internal model can represent the state transitions. (2) If the external generative model does not change over time, then updating of internal model ceases at a certain point. (3) P3b considered to be affected by prediction errors (Spratling, 2010; Kolossa et al., 2012) does not directly reflect the errors between a present event and a prediction generated by the internal model—the P3b variability is caused by the prediction errors for a stationary-state model translated from the internal model or is led by a different process from the generating of the internal model.

As pointed out in Mars et al. (2008) and Kolossa et al. (2012), the TCRT task requires motor responses. Therefore, it is still an open problem whether the cause of the variation in the ERP components is surprise conveyed by the stimulus or surprise associated with the motor responses.

We conclude that our approach using Markov chains provides observation of the different effects on ERPs produced by surprises on the stationary-state and state transition models. The differences in the effects suggest that an internal model in the brain can form a probability model with state transitions. The effects of a stationary-state model suggest the existence of a different brain mechanism from that for forming the internal model. Moreover, a change in these effects by the accumulation of events was observed. This shows the part of neural responses that reflects a brain mechanism by which humans gain predictions from their experiences.

AUTHOR CONTRIBUTIONS

HH, TM, and SN designed the work. HH and TM collected data. HH analyzed data. HH drafted the manuscript. TM and SN revised the manuscript.

ACKNOWLEDGMENTS

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [Grant numbers 15K21079, 26240043].

REFERENCES

Baldi, P., and Itti, L. (2010). Of bits and wows: a Bayesian theory of surprise with applications to attention. *Neural Netw.* 23, 649–666. doi: 10.1016/j.neunet.2009.12.007

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., and Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1672–1677. doi: 10.1073/pnas.0809667106

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., et al. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* 24, 127–135. doi: 10.1016/j.tree.2008.10.008

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Science, 3rd Edn.* Mahwah, NJ: Lawrence Erlbaum Associates.

Courchesne, E., Hillyard, S. A., and Galambos, R. (1975). Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephal. Clin. Neurophysiol.* 39, 131–143.

d'Acremont, M., Schultz, W., and Bossaerts, P. (2013). The human brain encodes event frequencies while forming subjective beliefs. *J. Neurosci.* 33, 10887–10897. doi: 10.1523/JNEUROSCI.5829-12.2013

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Dobson, A. J., and Barnett, A. (2011). *An Introduction to Generalized Linear Models*. Boca Raton, FL: CRC Press.

Donchin, E. (1981). Surprise!… Surprise? *Psychophysiology* 18, 493–513. doi: 10.1111/j.1469-8986.1981.tb01815.x

Donchin, E., and Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behav. Brain Sci.* 11, 357–374. doi: 10.1017/S0140525X00058027

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

Duncan-Johnson, C. C., and Donchin, E. (1977). On quantifying surprise: the variation of event-related potentials with subjective probability. *Psychophysiology* 14, 456–467. doi: 10.1111/j.1469-8986.1977.tb01312.x

El Karoui, I., King, J.-R., Sitt, J., Meyniel, F., Van Gaal, S., Hasboun, D., et al. (2015). Event-related potential, time-frequency, and functional connectivity facets of local and global auditory novelty processing: an intracranial study in humans. *Cereb. Cortex* 25, 4203–4212. doi: 10.1093/cercor/bhu143

Friston, K. (2002). Beyond phrenology: what can neuroimaging tell us about distributed circuitry? *Ann. Rev. Neurosci.* 25, 221–250. doi: 10.1146/annurev.neuro.25.112701.142846

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput. Biol.* 4:e1000211. doi: 10.1371/journal.pcbi.1000211

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, 3rd Edn*. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton, FL: CRC Press.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585–595. doi: 10.1016/j.neuron.2010.04.016

Hampton, A. N., Bossaerts, P., and O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* 26, 8360–8367. doi: 10.1523/JNEUROSCI.1010-06.2006

Haykin, S. (2005). *Adaptive Filter Theory, 4th Edn*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Horovitz, S. G., Skudlarski, P., and Gore, J. C. (2002). Correlations and dissociations between BOLD signal and P300 amplitude in an auditory oddball task: a parametric approach to combining fMRI and ERP. *Mag. Reson. Imaging* 20, 319–325. doi: 10.1016/S0730-725X(02)00496-4

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi: 10.1080/01621459.1995.10476572

Knight, R. T. (1984). Decreased response to novel stimuli after prefrontal lesions in man. *Electroencephalogr. Clin. Neurophysiol.* 59, 9–20.

Kolossa, A., Fingscheidt, T., Wessel, K., and Kopp, B. (2012). A model-based approach to trial-by-trial P300 amplitude fluctuations. *Front. Hum. Neurosci.* 6:359. doi: 10.3389/fnhum.2012.00359

Kolossa, A., Kopp, B., and Fingscheidt, T. (2015). A computational analysis of the neural bases of Bayesian inference. *NeuroImage* 106, 222–237. doi: 10.1016/j.neuroimage.2014.11.007

Kopp, B. (2006). The P300 component of the event-related brain potential and Bayes' theorem. *Cogn. Sci.* 2, 113–125. doi: 10.13140/2.1.4049.4402

Kopp, B., and Lange, F. (2013). Electrophysiological indicators of surprise and entropy in dynamic task-switching environments. *Front. Hum. Neurosci.* 7:300. doi: 10.3389/fnhum.2013.00300

Lieder, F., Daunizeau, J., Garrido, M. I., Friston, K. J., and Stephan, K. E. (2013). Modelling trial-by-trial changes in the mismatch negativity. *PLoS Comput. Biol.* 9:e1002911. doi: 10.1371/journal.pcbi.1002911

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. doi: 10.1523/JNEUROSCI.2925-08.2008

Matt, J., Leuthold, H., and Sommer, W. (1992). Differential effects of voluntary expectancies on reaction times and event-related potentials: evidence for automatic and controlled expectancies. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 810–822. doi: 10.1037/0278-7393.18.4.810

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. doi: 10.1037/0096-1523.24.5.1521

Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Char.* 231, 289–337. doi: 10.1098/rsta.1933.0009

Norris, J. R. (1998). *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

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. doi: 10.1016/j.neuroimage.2012.04.050

Patil, A., Huard, D., and Fonnesbeck, C. J. (2010). PyMC: Bayesian stochastic modelling in Python. *J. Stat. Softw.* 35, 1–81. doi: 10.18637/jss.v035.i04

Picton, T. W. (1992). The P300 wave of the human event-related potential. *J. Clin. Neurophysiol.* 9, 456–479. doi: 10.1097/00004691-199210000-00002

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Rac-Lubashevsky, R., and Kessler, Y. (2016). Dissociating working memory updating and automatic updating: the reference-back paradigm. *J. Exp. Psychol. Learn. Mem. Cogn.* 42, 951–969. doi: 10.1037/xlm0000219

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York, NY: Springer.

Saito, H., Takiyama, K., and Okada, M. (2015). Estimation of state transition probabilities: a neural network model. *J. Phys. Soc. Jpn.* 84:5. doi: 10.7566/JPSJ.84.124801

Sanmiguel, I., Saupe, K., and Schröger, E. (2013). I know what is missing here: electrophysiological prediction error signals elicited by omissions of predicted "what" but not "when". *Front. Hum. Neurosci.* 7:407. doi: 10.3389/fnhum.2013.00407

Seer, C., Lange, F., Boos, M., Dengler, R., and Kopp, B. (2016). Prior probabilities modulate cortical surprise responses: a study of event-related potentials. *Brain Cogn.* 106, 78–89. doi: 10.1016/j.bandc.2016.04.011

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

Smid, H., Jakob, A., and Heinze, H.-J. (1999). An event-related brain potential study of visual selective attention to conjunctions of color and shape. *Psychophysiology* 36, 264–279.

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area V1. *J. Neurosci.* 30, 3531–3543. doi: 10.1523/JNEUROSCI.4911-09.2010

Squires, K., Petuchowski, S., Wickens, C., and Donchin, E. (1977). The effects of stimulus sequence on event related potentials: a comparison of visual and auditory sequences. *Percept. Psychophys.* 22, 31–40.

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.

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

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.

A. DETAILS OF THE MODEL-BASED ANALYSIS

A.1. Predictive Surprise

Given an event sequence composed of n stimuli $E_0, E_1, \ldots, E_{n-1}$ where $E_{n'} \in \{a, b\}$ for $n' = 0, \ldots, n-1$, surprise concerning the nth trial that is based on the stationary or transition probability is defined as follows.

The stationary probability that E_n is X is denoted by $P_n(X)$, defined by

$$\begin{aligned} P_n(X) &= P_n(X \mid \{E_{n'}\}_{n'=0}^{n-1}) \\ &= \frac{\left|\{E_{n'} \mid E_{n'} = X, n' = 0, \ldots, n-1\}\right|}{n}, \end{aligned} \tag{A1}$$

where $X \in \{a, b\}$. Then, predictive stationary surprise $P_n(X)$ is defined by

$$I_n^{(\mathrm{S})} = \log_2 P_n(E_n). \tag{A2}$$

The transition probability that E_n is X is denoted by $P_n(X \mid E_{n-1})$, defined as

$$\begin{aligned} &P_n(X \mid E_{n-1}) \\ =&P_n(X \mid E_{n-1}, \{E_{n'}\}_{n'=0}^{n'-2}) \\ =&\frac{\left|\{E_{n'} \mid E_{n'} = X, E_{n'-1} = E_{n-1}, n' = 0, \ldots, n-1\}\right|}{|\{E_{n'} \mid E_{n'} = X, n' = 0, \ldots, n-1\}|}. \end{aligned} \tag{A3}$$

Then predictive transition surprise $P_n(X \mid E_{n-1})$ is defined as

$$I_n^{(\mathrm{T})} = \log_2 P_n(E_n | E_{n-1}). \tag{A4}$$

A.2. Regression Model

In this section, we describe the regression model summarized in **Figure 5**. We assume that the samples of the response variable are generated with a Gaussian distribution by

$$y_n \sim \mathcal{N}(\mu_n, \sigma_n^2), \tag{A5}$$

where $\mathcal{N}(\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2. The parameters of the distribution are assumed to be

$$\mu_n = b_1 x_n + b_2 + r_n, \tag{A6}$$

and

$$\sigma_n^2 = g_1 x_n + g_2 + u_n, \tag{A7}$$

where b_1 and g_1 are the slopes, b_2 and g_2 are the intercepts of the model, and r_n and u_n are the individual differences of the participants. Let $\mathcal{P}_p$ be the set of the indexes of the samples obtained from the participant p, where $\mathcal{P}_i \cap \mathcal{P}_j = \emptyset$ $(i, j = 1, \ldots, N_P,\ i \neq j)$, $\mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_{N_P} = \{1, \ldots, N\}$, and N_P is the number of participants. Then, r_n and u_n are defined as

$$r_n = \hat{r}_p, \quad n \in \mathcal{P}_p \tag{A8}$$

and

$$u_n = \hat{u}_p, \quad n \in \mathcal{P}_p. \tag{A9}$$

Therefore, the unknown parameters for the individual difference in the model are $\{\hat{r}_p, \hat{u}_p\}_{p=1}^{N_P}$, not $\{r_n, u_n\}_{n=1}^{N}$. Moreover, we assume the priors for b_1, b_2, g_1, g_2, $\{\hat{r}_p\}_{p=1}^{N_P}$, and $\{\hat{s}_p\}_{p=1}^{N_P}$ to be

$$b_1 \sim \mathcal{N}(\mu_{b_1}, \sigma_{b_1}^2), \tag{A10}$$

$$b_2 \sim \mathcal{N}(\mu_{b_2}, \sigma_{b_2}^2), \tag{A11}$$

$$g_1 \sim \mathcal{G}(\alpha_{g_1}, \beta_{g_1}), \tag{A12}$$

$$g_2 \sim \mathcal{G}(\alpha_{g_2}, \beta_{g_2}), \tag{A13}$$

$$\hat{r}_p \sim \mathcal{N}(\mu_r, \sigma_r^2),\ p = 1, \ldots, N_P, \tag{A14}$$

and

$$\hat{u}_p \sim \mathcal{G}(\alpha_u, \beta_u),\ p = 1, \ldots, N_P, \tag{A15}$$

where $\mathcal{G}(\alpha, \beta)$ is a gamma distribution with shape α and scale β. Furthermore, we define the hyper priors for $\sigma_{r_p}^2$ and β_{u_p} as

$$\sigma_r^2 \sim \mathcal{U}(a_r, b_r), \tag{A16}$$

and

$$\sigma_u^2 \sim \mathcal{U}(a_u, b_u), \tag{A17}$$

where $\mathcal{U}(a, b)$ is the uniform distribution over the interval $[a, b]$. The undefined parameters for the model were assumed to be constants. The parameters shown in **Table A1** were set to give a non-informative prior distribution for the parameters.

We found the unknown parameters for the hierarchical model by sampling with the Markov chain Monte Carlo (MCMC) method (Gelman et al., 2013). In particular, we used the Metropolis–Hastings algorithm implemented in PyMC 2.3.6 (Patil et al., 2010) for the sampling. The number of sampling was 100,000 (burn-in: 10,000).

TABLE A1 | The parameters that we consider to be constants in the model (Figure 5) and their values.

Parameter	Value	Parameter	Value
μ_{b_1}	0	σ_{b_1}	100
μ_{b_2}	0	σ_{b_2}	100
α_{g_1}	1	β_{g_1}	100
α_{g_2}	1	β_{g_2}	100
μ_r	0	α_u	1
a_r	0	b_r	100
a_u	0.1	b_u	100