Behavioral/Systems/Cognitive

Trial-by-Trial Fluctuations in the Event-Related Electroencephalogram Reflect Dynamic Changes in the Degree of Surprise

Rogier B. Mars, Stefan Debener, Thomas E. Gladwin, Lee M. Harrison, Patrick Haggard, John C. Rothwell, and Sven Bestmann

1Sobell Department of Motor Neuroscience and Movement Disorders and 2Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London WC1N 3BG, United Kingdom, 3Institute of Cognitive Neuroscience, University College London, London WC1N 3AR, United Kingdom, 4Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom, 5Medical Research Council Institute of Hearing Research, Royal South Hants Hospital, Southampton SO14 0YG, United Kingdom, 6Rudolf Magnus Institute of Neuroscience, Department of Psychiatry, University Medical Center Utrecht, Utrecht 3508 GA, The Netherlands, 7Department of Psychiatry, Stuivenberg Hospital, Antwerpen B-2-60, Belgium, and 8Department of Psychology, University College London, London WC1E 6BT, United Kingdom

The P300 component of the human event-related brain potential has often been linked to the processing of rare, surprising events. However, the formal computational processes underlying the generation of the P300 are not well known. Here, we formulate a simple model of trial-by-trial learning of stimulus probabilities based on Information Theory. Specifically, we modeled the surprise associated with the occurrence of a visual stimulus to provide a formal quantification of the "subjective probability" associated with an event. Subjects performed a choice reaction time task, while we recorded their brain responses using electroencephalography (EEG). In each of 12 blocks, the probabilities of stimulus occurrence were changed, thereby creating sequences of trials with low, medium, and high predictability. Trial-by-trial variations in the P300 component were best explained by a model of stimulus-bound surprise. This model accounted for the data better than a categorical model that parametrically encoded the stimulus identity, or an alternative model of predictability. Trial-by-trial variations in P300 can be explained by predictions made by an ideal observer keeping track of the probabilities of possible events. This provides evidence for theories proposing a direct link between the P300 component and the processing of surprising events. Furthermore, this study demonstrates how model-based analyses can be used to explain significant proportions of the trial-by-trial changes in human event-related EEG responses.

Key words: P300; single-trial EEG; information theory; surprise; attention; independent component analysis

Introduction

Late positive components of the human event-related brain potential (ERP), in particular the P300, have traditionally been associated with the processing of unexpected events (Sutton et al., 1965) (for review, see Nieuwenhuis et al., 2005). The amplitude of the P300 appears to be determined at least partly by the probability and relevance of an event (Duncan-Johnson and Donchin, 1977). Functionally, the P300 has commonly been linked to the revision of a participant's expectation about the current task context (Donchin, 1981; Donchin and Coles, 1988; Barcelo et al., 2006), as well as the updating of task-relevant information in anticipation of subsequent events (Barcelo et al., 2008). The P300 has widely been suggested to be modulated at least in part by the surprise of a stimulus (Donchin, 1981) and some authors have used a terminology related to information theory to describe processes underlying generation of the P300 (Ruchkin and Sutton, 1978; Johnson, 1986; Barcelo et al., 2008).

However, we are not aware of any study that has quantified fluctuations in surprise on a trial-by-trial basis to study its impact on the P300. A number of recent computational models have been proposed that formally quantify the surprise conveyed by sensory stimuli. In these models, the surprise associated with an event relates to its improbability, given a prediction of the occurrence of all possible events (Strange et al., 2005). Computationally, it might be an efficient strategy to focus processing resources on such surprising events, because these provide the most information to an observer (Baldi, 2005). One apparent advantage of using a model-based approach to quantify the intuitive notion of surprising events is that competing models about the cognitive processes underlying observed neural data can be formally tested (Corrado and Doya, 2007). Using this approach, recent neuroimaging studies in humans have shown that activity in a wide-spread parietal-premotor network is associated with the surprise associated with the presentation of a visual stimulus (Strange et al., 2005).
Here, we asked whether trial-by-trial variations in the P300 can be explained by such a formal model of surprise and whether this provides a more parsimonious description of the data than alternative models. Healthy participants performed a choice reaction time (RT) task while their brain activity was measured using electroencephalography (EEG). We then quantified the surprise associated with the unique stimulus sequence given to every participant and investigated whether these quantifications could explain variations in P300 on a single-trial basis. Our findings show that trial-by-trial variabilities in the P300 component are not random noise. A substantial proportion of this variability can be explained by formal quantifications of surprise, providing a direct confirmation of previous heuristics about the computations underlying the P300 component.

Materials and Methods

Participants, experimental design, and data acquisition. Twelve healthy participants (eight women, age range 18–29 years), all with normal or corrected-to-normal visual acuity participated in the experiment. All were recruited via the participants’ database of the Department of Psychology of University College London. Experimental procedures were approved by the local ethics committee and in accordance with the Declaration of Helsinki. Participants received £15 compensation for their time and travel.

Before the experiment, participants learned by trial-and-error the associations between four arbitrary visual stimuli (equated for surface area and brightness) and four button responses (using the index and middle fingers of both hands) for 60 trials. During this training, all stimuli were presented an equal number of times in random order. If participants did not perform the task without errors on the last 15 trials of the training block, it was repeated. During the main experiment participants performed 12 blocks of 60 trials of a choice reaction time task without feedback (see Fig. 1a). Visual stimuli were presented for 200 ms each, with a stimulus onset asynchrony of 2 s. Participants were required to respond to each stimulus with the previously associated button as quickly as possible, but not at the expense of accuracy. The probability of the occurrence of each event was manipulated between blocks such that the relative probabilities of events were either 0.25 for each event (low predictability), [0.4, 0.4, 0.1, 0.1] (medium predictability), or [0.7, 0.1, 0.1, 0.1] (high predictability). Participants were not informed about these probabilities. They were simply instructed to respond as quickly as possible to each presented stimulus and that the four different stimuli were randomly distributed across blocks. All stimuli occurred equally often over the course of the experiment and all stimuli had an equal behavioral relevance. Participants were given a break between blocks; they were free to initiate the subsequent block at their own pace.

The experiment was realized using the Cogent 2000 toolbox (University College London, http://www.vislab.ucl.ac.uk/Cogent2000/index.html) for Matlab (The Mathworks). EEG was recorded (bandpass filter: 0.05–100 Hz, 500 Hz sampling rate) using a Synamps2 amplifier (Neuropack) from the following electrode positions, using Ag/AgCl electrodes mounted in an elastic electrode cap: AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, T3, C3, CZ, C4, T8, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, PO3, PO4, Oz, and left and right mastoids. Horizontal and vertical eye movements were recorded using electrodes placed lateral to both eyes and above and below the left eye. Electrode Afz served as reference during recording and the electrode common was placed on the participants’ chin. Electrode impedances were kept at <10 kΩ.

Electrophysiological analyses. EEG data were analyzed using EEGLAB (Delorme and Makeig, 2004), implemented in Matlab 7.1. Each participant’s EEG data were bandpass filtered (0.3–30 Hz), down-sampled to 250 Hz, and re-referenced to average reference. Subsequently, epochs of −600 to 1400 ms around the presentation of the visual stimuli were extracted from each trial and linearly detrended. During the first step of artifact rejection, epochs containing unique, nonstereotyped artifacts (swallowing, head movements, etc) were rejected. In a second step, repeatedly occurring, stereotyped artifacts were removed using indepen-
assume initially that all stimuli are equally likely to occur. For the current setting, a prior distribution indicating the belief in all parameters before any observations is given by a prior Dirichlet distribution. A uniform Dirichlet prior over $p$ is parameterized by a vector $\alpha = [\alpha_1, \ldots, \alpha_K]$ and written as $P(\alpha) = \text{Dir}(\rho|\alpha)$. Choosing all elements of $\alpha$ equal to one represents the prior belief that the multinomial parameters are uniform. In the present case, this results in a belief that all four stimuli are likely to occur 25% of the time.

The degree of belief in the estimated probabilities $p$ will change when an event is observed. The posterior distribution representing the belief after $j$ trials, $X^j$, is given by

$$P(p|X^j, \alpha) = \text{Dir}(p|n_k + \alpha_k) = D^j,$$

where $n_k$ refers to the number of occurrences of outcome $k$ up until observation $j$. In words, this expression states that the estimated probability over the parameters $p$ is determined by the observations $X^j$ and a uniform prior (parameterized by $\alpha$). This is again a Dirichlet distribution, parameterized by the vector with elements equal to $n_k + \alpha_k$. Because the observer knows $n_k$ and $\alpha_k$ is fixed to be uniform, the posterior distribution can be computed easily and updated for each new observation. We abbreviate the estimated distribution following $j$ trials as $D^j$.

The posterior distribution after observing trial $j - 1$, i.e., $D^{j-1}$, can be used to predict the probability of each event occurring, i.e., the multinomial distribution, at the $j$th trial. The expression for this is

$$p(x_j = k|X^{j-1}, \alpha) = \frac{n_k^{j-1} + 1}{N^{j-1} + K} = \tilde{p}_k^j,$$

where the total number of observations up to the trial preceding $j$ is

$$N^{j-1} = \sum_{i=1}^{j} n_i^{j-1},$$

which is equal to $j - 1$. In words, the predicted probability of observing event (trial type) $k$ on the $j$th trial, given all preceding observations and a uniform prior is equal to $\tilde{p}_k^j$, where we have used the tilde to denote that it is a prediction. This quantity changes with each new observation and is the reason for including $j$ in the superscript. This can then be updated with each new event (trial) (cf. Strange et al., 2005).

Quantifying surprise. Following Strange et al. (2005), we can quantify the surprise, $I$, on each trial as follows (cf. Shannon, 1948):

$$I(x_j = k) = - \log_2 \tilde{p}_k^j.$$  (4)

This states that the surprise of observing event type $k$ at the $j$th trial is equal to the negative log of its predicted probability given all preceding trials. Accordingly, the amount of surprise conveyed by the occurrence of an event is high when an infrequent stimulus occurs in a stimulus sequence with high predictability. For example, in highly predictable blocks ([(0.7, 0.1, 0.1, 0.1)], the probability of one particular event is high, whereas the other three events occur only rarely. Given repeated samples of this distribution, these low frequency events are more surprising. An event is more surprising when occurring with 0.10 probability, compared with an event with a 0.70 probability of occurring (Fig. 1c). Note that in this experiment the generative distribution did not include dependencies between consecutive events. That is, the event at one time did not depend on earlier events. This is the same as in the study by Strange et al. (2005) and different to that investigated by Harrison et al. (2006), where the current event depended on the previous. Given the assumption that participants start each block anew, we refer to this model as blockwise surprise, $I_b$.

Alternative models. We compared the model of the previous section with a number of alternatives. The ideal observer described above assumed the generative model being stationary, i.e., unchanging within a block. This assumption is ideal in that it matches the true distribution used to generate trial types in the experiment. Furthermore, the model described above assumes participants start each block anew with the expectation that all events occur equally often, i.e., with a uniform (i.e., flat, uninformative) prior. Alternatively, one might expect that participants view each block merely as a continuation of the previous block, such that the experiment can be seen as one long session. We therefore also created a model based on an observer with no forgetting, here referred to as experiment-wise surprise, $I_e$ (Fig. 1d). This is suboptimal because contingencies did change from block to block.

An alternative formulation of surprise has been suggested by Baldi et al. (Baldi, 2002; Itti and Baldi, 2006), based on the Kullback–Leibler (KL) divergence (Kullback, 1959; Clover and Thomas, 1999). The KL divergence is a scalar quantity that summarizes the difference between two probability distributions. In our case, it is used to measure the change in
belief about the stimulus probabilities, \( P(x) \), after an event (i.e., visual stimulus). If this change is large then the event has a high degree of “surprise”, compared with one that has little or no effect. For the current experimental setting, the Kullback–Leibler divergence (KL surprise) at trial \( j \) is a function of the current, prior, and posterior distributions, \( D^1 - D^j \) (cf. Baldi, 2002):

\[
\text{KLsurprise} = KL(D^1, D^j).
\] (5)

In words, the (blockwise) KL surprise is the “distance” between the distributions before and after observing the \( j \)th trial. Intuitively this means that events can be quantified in terms of how much they change posterior beliefs.

The difference between the KL divergence measure of surprise (see Fig. 1e) and surprise as defined in Equation 4 is that the former is an average quantity, i.e., summed over all probabilities in the distribution. In contrast, the latter is a function of the predicted probability of an observed event, i.e., trial type presented to the subject. In other words, the KL divergence is a distance measure between the current, prior, and posterior distributions, whereas \( I_p \) is a function of the predicted probability of the observed event, i.e., just one event and not an average over all possible events. The KL surprise measure relates to those proposed by Ruchkin and Sutton (1978) and Kopp (2007) to account for variations in P300.

Last, we included a conventional explanation using a categorical model of events parametrically modulated by the probability of occurrence. In this model, each trial had one of the values \([0.10, 0.25, 0.40, 0.70]\). This regressor models variance related to stimulus probability within a block and does not take into account any learning; hence it is similar to the traditional method of averaging ERP data over a priori probabilities. Note that this model is similar to the model used by Duncan-Johnson and Donchin (1977), who used a linear regression analysis of single-trial P300 amplitudes and a priori event probabilities.

Model estimation and comparison. To test the hypothesis that surprise can predict event-related P300 responses we used a hierarchical general linear model (GLM), in which the parameters were optimized using empirical Bayes (Friston et al., 2007). Data from all \( S \) subjects were concatenated in a vector \( Y \) of length \( T \times S \), where \( T \) is the number of trials per subject. These data were fitted using a three-level hierarchical model of the following structure:

\[
Y = Z_1 w_1 + e_1
\]

where \( Z_1 \) is a design matrix. Hyperparameters \( \Lambda_1, \Lambda_2, \Lambda_3 \) control the precision (inverse variance) of noise at each level, given by \( \{e_1, e_2, e_3\} \); these correspond to within-subject error, between-subject error and shrinkage priors on the group-parameters, \( w_1, I \) is an identity matrix. The first level design matrix, \( Z_1 \) was block-diagonal, with dimensions \( TS \times PS \), with \( P \) regressors per subject. These regressors are the explanatory variables provided by our different models of the task sequence (see above). Additional regressors indicated the identity of trials on which participants responded erroneously, trials that were rejected during the preprocessing of the EEG data, and a constant term. By modeling incorrect responses explicitly, we accounted for the known effects of correct or incorrect responding on reaction times and P300 (Krigolson and Holroyd, 2007). The second design matrix, \( Z_2 = I_S \otimes I_P \), represented between-subject differences in the parameter weights, where \( I_S \) is a column of ones of length \( S \). We computed the posterior densities over model parameters and hyperparameters using standard techniques (Friston et al., 2007), where a posterior density represents the degree of belief in a parameter given data, i.e., single-trial P300 estimate.

The model evidence \( p(y|M_m) \), is the probability of the data given the \( m \)th model, which was approximated using the marginal likelihood (Penny et al., 2004; Friston et al., 2007). It is important to note that this quantity is computed by integrating out all model [hyper]-parameters and so it includes a complexity term as well as an accuracy term (expected likelihood). This evidence was used to compare competing models defined in terms of the explanatory variables in \( Z_i \).

We compared models using the ratio of the evidence for two competing models known as the Bayes Factor (Kass and Raftery, 1995). This can be formulated as a difference in approximate log model evidence for two models \( m \) and \( n \) (\( F_m \) and \( F_n \)) as follows:

\[
F_m - F_n = \ln \left( \frac{p(y|M_m)}{p(y|M_n)} \right) \approx \exp(F_m - F_n) = \frac{p(y|M_m)}{p(y|M_n)}. \] (7)

Here, a difference of +3 corresponds approximately to 20:1 odds, i.e., \( \exp(3) \approx 20 \), in favor of model \( m \) over \( n \) (Harrison et al., 2006; Bestmann et al., 2008). In the present case, positive values reflect stronger evidence in favor of the model containing surprise \( I_p \), whereas negative values would indicate stronger evidence for the alternative models tested.

Results
Behavioral results
Inspection of average reaction times on correct trials showed that participants’ reaction times were affected by changes in probabilistic context. A repeated-measures ANOVA with factor “probability” (4 levels: 0.1, 0.25, 0.40, and 0.70, indicating the overall a priori probability of a stimulus within a block) showed that participants responded slower to stimuli with a lower (0.10; 573 ± 21 ms; RT ≥ SEM) probability of occurrence, than stimuli with a higher probability of occurrence (0.70; 427 ± 18 ms): \( F_{(3,9)} = 107.904, p < 0.001 \). Participants responded incorrectly on 4.2% (SEM ± 0.69) of trials, making more errors in response to less frequent stimuli \( F_{(3,9)} = 15.01, p = 0.001 \).

Event-related potentials: trial-averaged results
Figure 2 shows the scalp topographies and grand average ERP over all trials, showing the traditional distribution of the P300. Our statistical analyses focused on the single-trial estimates of P300. Central latency of the time window used for single-trial P300 estimates was on average 531 (SEM ± 24, range 392–660 ms) after stimulus onset. To verify that our averaged single-trial P300 estimates showed the same scalp topology and ordering by stimulus probability as commonly reported for average P300s, averaged single-trial estimates were entered into a repeated measures ANOVA with factors electrode (4 levels: Fz, Cz, Pz, and O2) and probability (4 levels: 0.10, 0.25, 0.40, and 0.70). This analysis showed that average single-trial estimates differed reliably between electrodes \( F_{(3,9)} = 19.484, p < 0.001 \) and probability \( F_{(3,9)} = 14.936, p < 0.001 \). The difference in average P300 for each probability was most pronounced at electrode Pz (electrode × probability interaction: \( F_{(9,33)} = 7.659, p < 0.001 \), as is well established for the P300 (Duncan-Johnson and Donchin, 1977; Debener et al., 2005a) (Fig. 2c). This ordering of single-trial estimates is not due simply to a potential confounding relationship between trials rejected by the artifact correction and a priori stimulus probability, as there as no systematic relationship between the two \( F_{(3,33)} = 0.008, \) not significant).

Event-related potentials: model-based single-trial analyses
Having replicated the traditional P300 effects in choice reaction time tasks, we subsequently focused on the model-based analyses of the single-trial P300 estimates, following the procedure advocated by MacKay (1992). First, each model was fitted to the data using the procedure described above. Second, the model evidence
was calculated for each model and the models were compared using the Bayes factor. This analysis showed that the blockwise surprise $I_b$ model provided a more parsimonious account of the data when compared to a categorical model of a priori stimulus probabilities that was used by Duncan-Johnson and Donchin (1977). Moreover, the surprise $I_b$ model was favored over two alternative models of surprise, the KL surprise and surprise without forgetting $I_f$. The direct comparison of surprise $I_b$ with all other candidate models is presented in Figure 3a. A log-evidence ratio $> 3$ indicates 20:1 odds in favor of the surprise $I_b$ model.

Having established that the surprise $I_b$ model provided the most parsimonious explanation of the data, the group posterior density over the model parameter indicates the contribution of surprise to the data, i.e., the single-trial P300 estimate (cf. the $\beta$ in a standard regression analysis). This analysis showed that variations in P300 could be explained by surprise, with more surprising events leading to an increased P300 ($5.9 \mu V/bit$) (Fig. 3b). This finding was consistent across all participants (Fig. 3c).

**Discussion**

We investigated whether single-trial P300 estimates in a choice reaction time task could be explained by a formal model of the surprise conveyed by events experienced by participants. Behavioral data indicated that on average participants responded slower to less frequently occurring, i.e., more surprising, events. Consistent with earlier reports on the P300, we found that averaged P300s over central-parietal electrode sites interacted with the relative probability of event occurrence. Importantly, a model of the surprise within a block of trials provided a more parsimonious explanation of single-trial P300 changes than alternative models, including a categorical model of stimulus frequency, an alternative model of surprise based on the KL divergence, and surprise without forgetting. This novel model-based approach applied to single-trial EEG data allows for a formal quantification of the psychological variable “surprise” and its relationship to the psychophysiological marker P300.

Previous studies on the P300 have introduced the term “subjective probability” to denote that it is participants’ estimation of the environment that is crucial in predicting modulations in P300 (Donchin and Coles, 1988). This has led to the suggestion that P300 reflects the updating of information in anticipation of subsequent information processing (Sutton et al., 1965; Nieuwenhuis et al., 2005; Verleger et al., 2005; Barcelo et al., 2008). The P300 has previously been linked with information theoretic concepts (Ruchkin and Sutton, 1978; Johnson, 1986; Barcelo et al., 2006, 2008; Barcelo and Knight, 2007) or Bayes’ theory (Kopp, 2007). Here, we draw on information theoretic concepts to investigate the trial-by-trial influence of stimulus-bound surprise on P300 variation. Characterizing the subjective estimate of task probabilities has only recently become a major focus of research in cognitive and neurosciences (Oaksford and Chater, 2007). In the present case, we used a model of how the “subjective probability” is represented and updated over time, rather than how it changes on average.

To achieve this, the present approach combines two novel methodologies that, to our knowledge, have not been combined earlier in studies of event-related potentials. First, the model-based approach provides models about the trial-by-trial variability, an alternative model of surprise based on the KL divergence, and surprise without forgetting. This novel model-based approach applied to single-trial EEG data allows for a formal quantification of the psychological variable “surprise” and its relationship to the psychophysiological marker P300.
the present results agree with these authors’ suggestion that trial-

tion” that has been suggested to underlay the generation of the 
latter measure can be taken as a formal description of “equivoca-

otic model in which our ideal observers had suboptimal (i.e., no) 
from different distributions. We therefore included an alterna-

participants were privy to different blocks of events being sampled 
in corticospinal excitability (Bestmann et al., 2008). Here, we 
assumed that events were stationary and unchanging within a 
block, matching the true generative distribution from which 
events were sampled. Therefore, all previous blocks and events 
were forgotten in an optimal way and trials within the current 
block were weighted equally. Note that this assumption is ideal in 
relation to the actual experimental paradigm but assumes partic-

ipants were privy to different blocks of events being sampled 
from different distributions. We therefore included an alterna-
tive model in which our ideal observers had suboptimal (i.e., no) 
forgetting with respect to the actual experimental paradigm. In 
the present experiment, a model of an ideal observer beginning 
each block with flat priors, was superior to a model without 
forgetting.

Moreover, we also compared our model to an alternative mea-

ure of surprise based on the Kullback–Leibler divergence. This 
latter measure can be taken as a formal description of “equivoca-
tion” that has been suggested to underlay the generation of the 
P300 (Ruchkin and Sutton, 1978; see also Kopp, 2007). Although 
the present results agree with these authors’ suggestion that trial-
by-trial estimates of surprise based on each participant’s unique 
trial history is important in predicting fluctuations in P300, we 
show that surprise I_s based on only the estimated probability of 
the stimulus presented on a given trial rather than the full distri-
bution of trials, provides a better explanation to characterize 
changes in P300.

A remaining question is how the present modeling approach 
of single-trial P300 links with recent neurophysiological models 
of P300 generation. Nieuwenhuis et al. (2005) proposed that the 
P300 reflects the arrival of a phasic norepinephrine (NE) signal in 
cortical areas, which serves to increase signal transmission in the 
cortex. This proposal is based on a number of considerations, 
such as the similarities between the anteceditions for phasic 
increases in NE and the generation of the P300 and between the 
target areas of NE projections and known P300 generators, and 
pharmacological studies that seem broadly consistent with this 
proposal (for review, see Nieuwenhuis et al., 2005). In this re-

spect, it is interesting to note that recent advances in computa-
tional neuroscience point to a role of NE in the processing of 
contextual uncertainty. Specifically, Dayan and Yu (2006) pro-
posed that phasic NE signals unexpected changes in the world 
within the context of a task. The hope of the approach taken in 
the current study is to use such computationally informed models to 
investigate the link between phasic NE to single-trial P300 data.

In the present task, each visual stimulus was linked to a dis-
tinct motor response and other factors that might influence P300, 
such as stimulus salience and task relevance (Johnson, 1986; De 
Brujin et al., 2004), were kept constant. Therefore, we cannot 
determine whether the P300 modulation was purely due to the 
surprise conveyed by the visual stimuli, or whether it was related 
to the response selection on each trial (cf. Koechlin and Summer-
field, 2007). Previous studies indeed show that P300 modulation 
can be explained in terms of the probabilistic updating of the 
corresponding motor response (Barcelo and Knight, 2007; 
Barcelo et al., 2008).

We have referred to the centroparietal component we found as 
the P300. Other studies have made a further distinction be-
tween the so-called P3a and P3b subcomponents (Polich, 2007). 
The P3b is the component commonly referred to as “P300,” and 
is commonly evoked by target stimuli at around 300–600 ms, 
similar to the component observed in the present study. In con-
trast, the P3a is linked to infrequent, task-novel events, and has a 
frontocentral maximum occurring at ~250–400 ms (Courchesne et al., 1975; Friedman et al., 2001). In addition, the 
P3a component habituates fast, possibly following the pattern 
predicted by the KL surprise, rather than the surprise I_s that 
predicts P3b. This may be tested directly in experiments specifi-
cally designed for eliciting P3a responses (Debener et al., 2005a), 
using the modeling framework presented here. The present study 
did not focus on the difference between the novelty and 
attention-related P3a and the target and response-related P3b 
component. Moreover, our focus on the amplitude of P300 did 
not focus on potential information conveyed by P300 latency 
(Donchin, 1981; Donchin and Coles, 1988). Nevertheless, the 
amplitude contains sufficient structure that can be explained by 
a formal definition of surprise. By taking into account P3b versus 
P3a effects and latency information, it may be possible to con-
sider surprise in the context of other mental states contributing 
to goal-oriented behavior.

To conclude, model-based single-trial analyses can be used for 
testing hypotheses of event-related EEG fluctuations. This ap-
proach provides a bridge between cognitive theories and more 
formal neurophysiological models of the P300 ERP. The focus on 
single-trial EEG data provides a more direct link to behavior and 
nervous processing than averaged EEG activity (Debener et al., 
2006). This is supported by our observation that P300 trial-by-
trial amplitude fluctuations are not random noise, and can be 
explained by a formal model of surprise experienced in the con-
text of a behavioral task. Our findings provide direct evidence for 
thories linking the P300 component and the processing of sur-
prising events.

References


Bald P (2005) Surprise: A shortcut for attention. In: Neurobiology of atten-
Academic.

Barcelo F, Knight RT (2007) An information-theoretical approach to context-
tual processing in the human brain: Evidence from prefrontal lesions. 
Cereb Cortex 17:515–516.

Barcelo F, Escera C, Corral MJ, Periáñez JA (2006) Task switching and nov-