frontiers in Neuroscience

ORIGINAL RESEARCH
published: 27 December 2016
doi: 10.3389/fnins.2016.00573

Mind the Noise When Identifying Computational Models of Cognition from Brain Activity

Antonio Kolossa and Bruno Kopp *

Department of Neurology, Hannover Medical School, Hannover, Germany

The aim of this study was to analyze how measurement error affects the validity of modeling studies in computational neuroscience. A synthetic validity test was created using simulated P300 event-related potentials as an example. The model space comprised four computational models of single-trial P300 amplitude fluctuations which differed in terms of complexity and dependency. The single-trial fluctuation of simulated P300 amplitudes was computed on the basis of one of the models, at various levels of measurement error and at various numbers of data points. Bayesian model selection was performed based on exceedance probabilities. At very low numbers of data points, the least complex model generally outperformed the data-generating model. Invalid model identification also occurred at low levels of data quality and under low numbers of data points if the winning model's predictors were closely correlated with the predictors from the data-generating model. Given sufficient data quality and numbers of data points, the data-generating model could be correctly identified, even against models which were very similar to the data-generating model. Thus, a number of variables affects the validity of computational modeling studies, and data quality and numbers of data points are among the main factors relevant to the issue. Further, the nature of the model space (i.e., model complexity, model dependency) should not be neglected. This study provided quantitative results which show the importance of ensuring the validity of computational modeling via adequately prepared studies. The accomplishment of synthetic validity tests is recommended for future applications. Beyond that, we propose to render the demonstration of sufficient validity via adequate simulations mandatory to computational modeling studies.

Keywords: computational modeling, functional brain imaging, event-related potentials, signal-to-noise ratio, validity, model identifiability, design optimization

OPEN ACCESS

Edited by:
Pedro Antonio Valdes-Sosa,
Joint China Cuba Lab for Frontiers
Research in Translational
Neurotechnology, Cuba

Reviewed by:
Jean Daunizeau,
Brain and Spine Institute (ICM), France
Adeel Razi,
Wellcome Trust Centre for
Neuroimaging (WT) - UCL, UK

Correspondence:
Bruno Kopp
kopp.bruno@mh-hannover.de

Specialty section:
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: *27 June 2016*
Accepted: *28 November 2016*
Published: *27 December 2016*

Citation:
Kolossa A and Kopp B (2016) Mind
the Noise When Identifying
Computational Models of Cognition
from Brain Activity.
Front. Neurosci. 10:573.
doi: 10.3389/fnins.2016.00573

1. INTRODUCTION

Computational biology involves mathematical modeling techniques to study biological systems. For example, computational neuroscience is the study of brain function in terms of quantitative models of information processing in the nervous system (e.g., Sejnowski et al., 1988). Computational cognitive neuroscience (CCN) represents an emerging subfield of computational neuroscience which aims to identify computational models of cognition from measures of brain activity (e.g., Knill and Pouget, 2004; Friston, 2005; O'Reilly et al., 2012, 2013; Koechlin, 2014; Gerstner and Frémaux, 2015; Kira et al., 2015; Pecevski and Maass, 2016). The general framework of CCN rapidly gains popularity because of both its overwhelming advantages compared to non-computational methods and its potential clinical importance. The emergence of new fields, such as for example computational psychiatry (e.g., Montague et al., 2012; Corlett and Fletcher, 2014; Stephan and Mathys, 2014; Adams et al., 2016; Huys et al., 2016), may serve as an indicator of these developments. Still another reason why CCN rapidly gains popularity is that it represents a modality-general approach in that it deals with all functional brain imaging modalities such as, for example, functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and electroencephalography (EEG).

CCN includes many techniques for computational modeling, i.e., for analyzing relationships between observed data and latent variables. The common approach is *forward* modeling which expresses the observed data as functions of some predictors. Since forward models provide a model for the generation of the observed data, they are also referred to as generative models in the machine learning literature (Haufe et al., 2014). In contrast, *backward* models "extract" latent variables as functions of the observed data, i.e., they reverse the direction of the functional dependency between latent variables and data compared to forward models. They are typically used if there is no need to model the generation of the data, i.e., when one is only interested in transforming observed data into a (potentially low-dimensional) representation in which they exhibit certain desired characteristics (a familiar example is brain-computer interfaces; e.g., Nicolelis, 2001; Blankertz et al., 2007).

In this article, we present a simulation study evaluating methodological issues related to the validity of forward modeling studies. Validity implies in the context of CCN that a forward modeling study renders it possible to identify from observed brain activities the proper generative model which usually corresponds to a computational model of cognition. A major threat to the validity of forward modeling studies lies in the fact that observed brain activities are a combination of many sources of variation some of which are systematically related to the latent variables (i.e., the desired signal), but some of which represent measurement error and background noise. Our study is the first to handle the noise problem as a validity problem of identifying the proper data-generative model.

We recognized that the effects of the signal-to-noise ratio (SNR) on the validity of forward modeling studies are still not dealt with in the literature. In order to fill that gap, our study used synthetic EEG data of varying data quality (i.e., degrees of noise) to examine the circumstances under which the proper, rather than an improper, generative model can be identified, thereby providing formal insights into the relationship between variations in data quality and the validity of forward modeling studies.

We chose single-trial P300 (or P3, also P3b) amplitudes of the event-related brain potential (ERP) mainly for two reasons as an example. First, it is well-established that single-trial P300 amplitudes are sensitive to the degree to which eliciting stimuli are surprising (Donchin, 1981), and surprise is a well-defined information theoretic metric (Shannon and Weaver, 1948). Second, the relevant model space is comparatively limited: Squires et al. (1976) presented a model of P300 amplitude fluctuations, based on the concept of expectancy, see below. Basically following that lead, we suggested a computational model (Kolossa et al., 2013) which represents a refinement of Squires et al. (1976). An alternative to these two multifactorial computational models was proposed by Mars et al. (2008) who suggested a simple unifactorial computational model that keeps track of the relative frequency of stimuli.

We chose our own model (Kolossa et al., 2013) as the generative model of the synthetic EEG data, and we analyzed whether Bayesian model comparison techniques (Friston et al., 2007; Stephan et al., 2009; Penny, 2012)—which are commonly used for model selection of ERP data (Mars et al., 2008; Ostwald et al., 2012; Kolossa et al., 2013, 2015; Lieder et al., 2013)—were capable to identify the proper generative model under varying degrees of noise. For sake of simplicity, we consider single-channel EEG data (i.e., single-trial amplitude measures obtained from one single recording channel) rather than multichannel EEG data. Note that multivariate methods combine information from different channels and thus render it possible to cancel out some degree of noise (Makeig et al., 1996; Haufe et al., 2014). However, the application of multivariate methods does not solve, but merely ameliorates the noise problem.

2. MATERIALS AND METHODS

2.1. System Model

Single-trial ERPs can be extracted from the EEG and they are measurable traces of cognitive processes (Luck, 2014). Here, we used synthetic ERPs that have the advantage of providing well-defined observables, i.e., data = signal + noise. Thus, this study explores the framework of CCN by starting with known signals and by adding various levels of noise in order to see whether the signal-generating model can be re-established against a background of alternative models (see below for details). We are mainly interested to see how the validity of the model selection hinges upon (a) the SNR ratio, (b) the number of data, (c) the dependency in the model space, and (d) model complexity.

Three related models of single-trial P300 amplitude fluctuations are taken from the literature, namely the SQU model proposed by Squires et al. (1976), the MAR model proposed by Mars et al. (2008), and the DIF model published in Kolossa et al. (2013). These models, along with the null model (NUL), constitute a model space $\mathcal{M} = \{\mathrm{NUL}, \mathrm{MAR}, \mathrm{SQU}, \mathrm{DIF}\}$. A series of N random events $k = \{1, ..., K\}$, here with $K=2$, is drawn to form observations $o(n) = k$, with trial index $n \in \{1, ..., N\}$. The DIF model (see below) is used to calculate the surprise $I_P(n)$ over the observation $o(n) = k$. An offset ϑ is added to $I_P(n)$ to yield the signal $s(n)$

$$s(n) = \vartheta + I_P(n). \tag{1}$$

Artificial noise $\epsilon(n)$ is then added to $s(n)$ to yield the synthetic ERP (sERP) $y(n)$ following

$$y(n) = s(n) + \epsilon(n). \tag{2}$$

This procedure is repeated for $L = 16$ virtual subjects $\ell = \{1, ..., L\}$. Random effects Bayesian model selection for group studies (Stephan et al., 2009) is then used to evaluate which of the models $m \in \mathcal{M}$ actually generated the ERP sequence $I_P(n)$. These analyses are repeated for different levels of noise and various numbers of data points N.

In the following, Bayesian model selection (BMS) is shortly introduced before the model space is formally defined. An introduction to SNR estimation for ERPs precedes the detailed description of the analysis framework. A short note on notation: small bold symbols refer to vectors, capital bold symbols to matrices, and $[]^T$ denotes the transpose. Thus, the vector $\mathbf{y} = [y(n = 1), ..., y(n = N)]^T$ captures the synthetic data over trials, while $\boldsymbol{\epsilon} = [\epsilon(n = 1), ..., \epsilon(n = N)]^T$ represents the corresponding noise. All simulations were performed using MATLAB 7.11.0 and the Statistical Parametric Mapping (SPM8) software.

2.2. Bayesian Model Selection

BMS methods are widely applied in many fields (Raftery, 1995; Hoeting et al., 1999; Penny and Roberts, 2002; Pitt and Myung, 2002; Beal and Ghahramani, 2003; Kemp et al., 2007; Hoijtink et al., 2008; Vyshemirsky and Girolami, 2008; Toni et al., 2009; Penny et al., 2010; Kolossa et al., 2016). We used a two-level hierarchical general linear model (GLM) with the Parametric Empirical Bayesian (PEB) scheme and random effects BMS for group studies as implemented in the SPM software (Friston et al., 2002, 2007; Stephan et al., 2009). The two-level hierarchical model equips a standard general linear model with a second level that places constraints on the parameter estimates of the first level. For each subject ℓ and model m of the model space $\mathcal{M}$, the log-evidence is approximated with a variational free energy bound $F_{m,\ell}$ which consists of an accuracy and a complexity term (Penny et al., 2004; Friston et al., 2007; Penny, 2012). Random effects (RFX) BMS for group studies computes exceedance probabilities φ_m for all models, which equals the probability that model m is more likely than all other models (Stephan et al., 2009).

The two-level GLM is of the form

$$\begin{aligned} \mathbf{y} &= \mathbf{X}^{(1)}\boldsymbol{\theta} + \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\theta} &= \boldsymbol{\epsilon}^{(2)}. \end{aligned} \tag{3}$$

The first level of the GLM contains two parameters $\boldsymbol{\theta} = [\theta_1 \ \theta_2]^T$ which model intercept and slope, respectively. The model-dependent compositions of the first-level design matrices $\mathbf{X}^{(1)}$ will be shown in the model-specific sections below. The second level of the GLM sets an unconstrained prior on the first-level parameters $\boldsymbol{\theta}$ and allows for single-level Bayesian inference (Ostwald et al., 2012).

All errors are assumed to be normally distributed with $\boldsymbol{\epsilon}(1) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon^{(1)})$ and $\boldsymbol{\epsilon}(2) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\epsilon^{(2)})$. The covariance is parameterized following $\boldsymbol{\Sigma}_\epsilon^{(1)} = \lambda^{(1)}\mathbf{I}_N$ and $\boldsymbol{\Sigma}_\epsilon^{(2)} = \lambda^{(2)}\mathbf{I}_2$, with $\mathbf{I}_N \in \mathbb{R}^{N\times N}$ being an identity matrix. The parameters $\boldsymbol{\theta}$ and the hyper-parameters $\lambda^{(1)}$ and $\lambda^{(2)}$ are estimated using an expectation maximization (EM) algorithm. After convergence of the EM, the conditional means of the first-level parameters $\boldsymbol{\mu}_{\theta|\mathbf{y}}$ are used as point estimates (Friston et al., 2002) for model fitting to yield the sERP estimates

$$\hat{\mathbf{s}} = \mathbf{X}^{(1)}\boldsymbol{\mu}_{\theta|\mathbf{y}} \tag{4}$$

before calculation of the Spearman correlation and the explained variance (see below).

2.3. Spearman Correlation

We use the Spearman correlation ρ as a measure of similarity between two models $m = 1$ and $m = 2$. It follows

$$\rho = 1 - \frac{6\sum_{n=1}^{N} d^2(n)}{N(N^2 - 1)}, \tag{5}$$

with $d(n)$ being the distance between the ranks of the sERP predictors from two models $\hat{s}_{m=1}(n)$ and $\hat{s}_{m=2}(n)$ on trial n.

2.4. Explained Variance

As an absolute measure of fit of the models to the data, we use the explained variance calculated as the squared correlation coefficient

$$R^2 = \left(\frac{\sum_{n=1}^{N}(\hat{s}(n) - \bar{\hat{s}})(y(n) - \bar{y})}{\sqrt{\sum_{n=1}^{N}(\hat{s}(n) - \bar{\hat{s}})^2 \sum_{n=1}^{N}(y(n) - \bar{y})^2}} \right)^2, \tag{6}$$

with $\bar{\hat{s}}$ and $\bar{y}$ as the means of $\hat{\mathbf{s}}$ and $\mathbf{y}$, respectively.

2.5. Model Space

This section details the four models which constitute the model space $\mathcal{M} = \{\text{NUL}, \text{MAR}, \text{SQU}, \text{DIF}\}$. It also specifies the respective first-level design matrices $\mathbf{X}^{(1)}$ which are input to the model estimation and selection framework. For all models except for the NUL model, the first-level design matrices consist of a constant term and the respective ERP predictor, as will be detailed below.

2.5.1. NUL Model

The NUL model represents the null hypothesis that the signal is constant and variation in the data is solely due to noise. Thus, the first level design matrix is an all one column vector

$$\mathbf{X}^{(1)} = \mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^N. \tag{7}$$

Notice that the GLM (3) for the NUL model consists only of an intercept θ_1, thus greatly reducing the complexity of this model.

2.5.2. MAR Model

The MAR model as proposed by Mars et al. (2008) uses predictive surprise $I_P(n)$ over observations to predict the sERP. This model keeps track of the observation probability $\mathrm{P}_k(n)$ according to

$$\mathrm{P}_k(n) = \frac{\tilde{c}_{\mathrm{L},k}(n) + 1}{(n-1) + K}, \tag{8}$$

with the long-term memory count function $\tilde{c}_{\mathrm{L},k}(n)$ counting the number of occurrences of event k until trial $n-1$. Please refer to Mars et al. (2008) or Kolossa et al. (2013) for further details on the count function. After an observation is made, the observation probability is transformed to predictive surprise following

$$I_P(n) = -\log_2(\mathrm{P}_{k\,=\,o(n)}(n)). \tag{9}$$

The first-level design matrix for the MAR model has the form

$$\mathbf{X}^{(1)} = \begin{bmatrix} 1 & I_P(n=1) \\ \vdots & \vdots \\ 1 & I_P(n=N) \end{bmatrix} \in \mathbb{R}^{N\times 2}, \tag{10}$$

thus modeling the sERP to be composed of an offset as in (1) and predictive surprise.

2.5.3. SQU Model

The SQU model uses expectancy $\mathrm{E}_k(n)$ for event $k \in \{1,2\}$ on trial n as sERP predictor. While Squires et al. (1976) originally did not provide a complete analytical form of their model, Kolossa et al. (2013) present a thoroughly mathematical reformulation of their approach. The expectancy that event $k \in \{1,2\}$ will be observed on trial $n \in \{1,...,N\}$ consists of an exponentially decaying count function for short-term memory, $\check{c}_{\mathrm{S},k}(n)$, a count function for alternation expectancy, $\check{c}_{\mathrm{A},k}(n)$, along with the global event probability, P_k, which combine to

$$\mathrm{E}_k(n) = 0.235\cdot\check{c}_{\mathrm{S},k}(n) + 0.033\cdot\check{c}_{\mathrm{A},k}(n) + 0.505\cdot\mathrm{P}_k - 0.027. \tag{11}$$

The constants are empirically derived best-fitting parameters. The interested reader is referred to Kolossa et al. (2013) for a detailed derivation of the count functions $\check{c}_{\mathrm{S},k}(n)$ and $\check{c}_{\mathrm{A},k}(n)$. Analogously to the MAR model, the first-level design matrix for the SQU model is of the form

$$\mathbf{X}^{(1)} = \begin{bmatrix} 1 & \mathrm{E}_{k\,=\,o(n)}(n=1) \\ \vdots & \vdots \\ 1 & \mathrm{E}_{k\,=\,o(n)}(n=N) \end{bmatrix} \in \mathbb{R}^{N\times 2}. \tag{12}$$

2.5.4. DIF Model

The digital filtering (DIF) model predicts the sERP with predictive surprise akin to the MAR model. It keeps track of the observation probability $\mathrm{P}_k(n)$ but with an exponentially decaying short-term memory count function, $c_{\mathrm{S},k}(n)$, an alternation expectation contribution, $c_{\mathrm{A},k}(n)$, and exponentially decaying long-term memory count function, $c_{\mathrm{L},k}(n)$. It thus combines properties of the SQU and MAR model. The three contributions and an additive probability-normalizing constant $\frac{1}{C}$ combine to

$$\mathrm{P}_k(n) = 0.83\cdot c_{\mathrm{L},k}(n) + 0.12\cdot c_{\mathrm{S},k}(n) + 0.05\cdot\left[c_{\mathrm{A},k}(n) + \tfrac{1}{C}\right]. \tag{13}$$

The interested reader is referred to Kolossa et al. (2013) for details on the count functions and the empirical derivation of the constants. Once event k on trial n has been observed, the observation probability is transformed to predictive surprise (9), yielding the first-level design matrix

$$\mathbf{X}^{(1)} = \begin{bmatrix} 1 & I_P(n=1) \\ \vdots & \vdots \\ 1 & I_P(n=N) \end{bmatrix} \in \mathbb{R}^{N\times 2}, \tag{14}$$

akin to the MAR and SQU models.

2.6. Signal-to-Noise Ratio (SNR)

Though often neglected, the SNR (power) ratio of the EEG defines the boundary conditions in ERP research. It depends on the SNR how many trials are necessary for meaningful ERP estimates (Luck, 2004) and reliable BMS (Penny, 2012). Early methods for SNR estimation for ERPs go back to the times of the discovery of the P300 (Sutton et al., 1965; Schimmel, 1967). Only few approaches were presented in later years (Coppola et al., 1978; Başar, 1980; Raz et al., 1988; Puce et al., 1994). The one proposed by Möcks et al. (1988) is still used as the basis for current developments (beim Graben, 2001; Paukkunen et al., 2010).

2.6.1. Estimating the SNR

We follow the approach from Möcks et al. (1988) which we now briefly describe. Notice that the SNR is calculated for each event type k separately and averaged afterwards. So first, the sERP amplitudes $y(n)$ are separated according to their event type k, yielding $y_k(n)$, with $n \in \{1,...,N_k\}$ and N_k as the total number of trials in which event k was observed. The sERPs $y_k(n)$ are assumed to be composed of the signal s_k and stationary ergodic noise $\epsilon_k(n)$ with variance $\sigma^2_{\epsilon_k}$ (beim Graben, 2001), yielding

$$y_k(n) = s_k + \epsilon_k(n), \tag{15}$$

with s_k as constant over trials. These assumptions are not met for real ERP amplitudes, but they are nevertheless accepted as useful simplifications (Möcks et al., 1988). The SNR for event k is defined as the ratio of the power of the signal over the noise power (beim Graben, 2001)

$$\mathrm{SNR}_k = \frac{P_{s_k}}{P_{\epsilon_k}} = \frac{s_k^2}{\sigma^2_{\epsilon_k}}. \tag{16}$$

Möcks et al. (1988) propose the noise power estimate

$$\hat{P}_{\epsilon_k} = \frac{1}{N_k - 1}\sum_{n=1}^{N_k}\left(y_k(n) - \bar{y}_k\right)^2, \tag{17}$$

with

$$\bar{y}_k = \frac{1}{N_k}\sum_{n\,=\,1}^{N_k} y_k(n). \tag{18}$$

The power of the sERP follows (beim Graben, 2001)

$$P_{y_k} = \overline{y_k^2} = \frac{1}{N_k}\sum_{n\,=\,1}^{N_k} y_k^2(n) \tag{19}$$

and, assuming statistical independence between signal and noise, it is composed of the power of the signal P_{s_k} and the power of the noise P_{ϵ_k} according to

$$P_{y_k} = P_{s_k} + \frac{1}{N_k} P_{\epsilon_k}. \tag{20}$$

Notice that the noise left in P_{y_k} is attenuated by the factor N_k, therefore the scaling of the noise power P_{ϵ_k} by $\frac{1}{N_k}$ in (20) (Möcks et al., 1988; Paukkunen et al., 2010; Czanner et al., 2015). The signal power can now be estimated from (20)

$$\hat{P}_{s_k} = P_{y_k} - \frac{1}{N_k} \hat{P}_{\epsilon_k} \tag{21}$$

and the SNR estimate $\hat{\mathrm{SNR}}_k$ in [dB] follows

$$\hat{\mathrm{SNR}}_k \text{ [dB]} = 10 \log_{10} \frac{\hat{P}_{s_k}}{\hat{P}_{\epsilon_k}}. \tag{22}$$

2.6.2. Setting the SNR

We employ the SNR estimation methods described above for generating sERPs with a specific SNR. Notice that even in response to the same event type k, real ERPs are not constant over trials (Squires et al., 1976; Mars et al., 2008; Ostwald et al., 2012; Kolossa et al., 2013, 2015; Lieder et al., 2013). In order to make this work applicable to real ERPs we use a trial-variable signal $s_k(n)$ in (15) instead of a constant s_k. The signal power then follows

$$P_{s_k} = \frac{1}{N} \sum_{n=1}^{N} s_k^2(n) \tag{23}$$

and the noise power is known as

$$P_{\epsilon_k} = \sigma_{\epsilon_k}^2. \tag{24}$$

Inserting (24) in (22) and solving for $\sigma_{\epsilon_k}^2$ yields

$$\sigma_{\epsilon_k}^2 = \frac{P_{s_k}}{10^{\frac{\text{SNR [dB]}}{10}}} \tag{25}$$

which is the sought after error variance $\sigma_{\epsilon_k}^2$ for a desired SNR [dB], given $s_k(n)$.

2.7. sERP Generation

The sERP $y(n)$ is generated following (2) in Section 2.1. Notice that due to the event type-dependent nature of the SNR estimation for ERPs, the signal $s(n)$ is first separated according to the event type k before zero-mean Gaussian noise $\epsilon_k \sim \mathcal{N}(0, \sigma_{\epsilon_k}^2)$ is added to the sub-signals $s_k(n)$ to yield $y_k(n)$

$$y_k(n) = s_k(n) + \epsilon_k(n). \tag{26}$$

The values of $\sigma_{\epsilon_k}^2$ are determined in dependence on the SNR condition according to (25). After the addition of noise to the sub-signals, the sERP $y(n)$ is derived by combining $y_k(n)$ for both event types in their original trial order. We created noise conditions of SNR [dB] $\in \{10, 8, 6, 4, 2, 0\}$ which are reasonable for ERP data (Kolossa, 2016). For each of the $L = 16$ subjects $\ell \in \mathcal{L} = \{1, ..., L\}$, the SNR is drawn from a normal distribution with 2 dB variance $\mathrm{SNR}_\ell \sim \mathcal{N}(\mathrm{SNR}, \sigma_{\mathrm{SNR}}^2 = 2\,\mathrm{dB})$ to model variability of SNRs over subjects. For each SNR condition, ten different numbers of data points $N \in \{50, 100, 150, 200, 250, 300, 350, 400, 450, 500\}$ are used, yielding a total of 60 scenarios with different combinations of SNR and N.

In each scenario, a sequence of N events is randomly drawn, with a probability for the frequent event of $\mathrm{P}_{k=1} = 0.7$ and for the rare event of $\mathrm{P}_{k=2} = 0.3$. The DIF model is used to calculate predictive surprise values which are then degraded by noise as described above in Section 2.1 to yield the sERP. All models m of the model space $\mathcal{M} = \{\mathrm{NUL, MAR, SQU, DIF}\}$ are then subjected to BMS (see Section 2.2). After fitting the models, the Spearman correlation (see Section 2.3) between the DIF model and the MAR model as well as between the DIF model and the SQU model plus the explained variance (see Section 2.4) of the MAR, SQU, and DIF model are calculated. When a scenario is completed for all L subjects, exceedance probabilities φ and the median Spearman correlation ρ and explained variance R^2 are calculated to obtain group-level results. Each scenario is simulated five hundred times with new sampling of stimuli and errors (Penny, 2012). Finally, the medians of exceedance probabilities, Spearman correlations and percentages of explained variance over all 500 repetitions are obtained. The following pseudocode summarizes the simulation procedure:

```
start
for SNR = 0,...,10 [dB]
   for num. data = 50,...,500
      for 500 simulations
         for 16 subjects
            SNR sampling
            stimulus sampling
            noise sampling
            Spearman correlation
            expl. variance
            model evidence
         end over subjects
         exceedance probabilities
         median Spearman correlation
         median expl. variance
      end over simulations
      median exceedance probabilities
      median Spearman correlation
      median expl. variance
   end over num. data
end over SNR
end
```

3. RESULTS

Figure 1 shows the median Spearman correlations ρ (5) between predictors from the data-generating DIF model and the SQU model (-○-) and the MAR model (··□··), respectively, as a

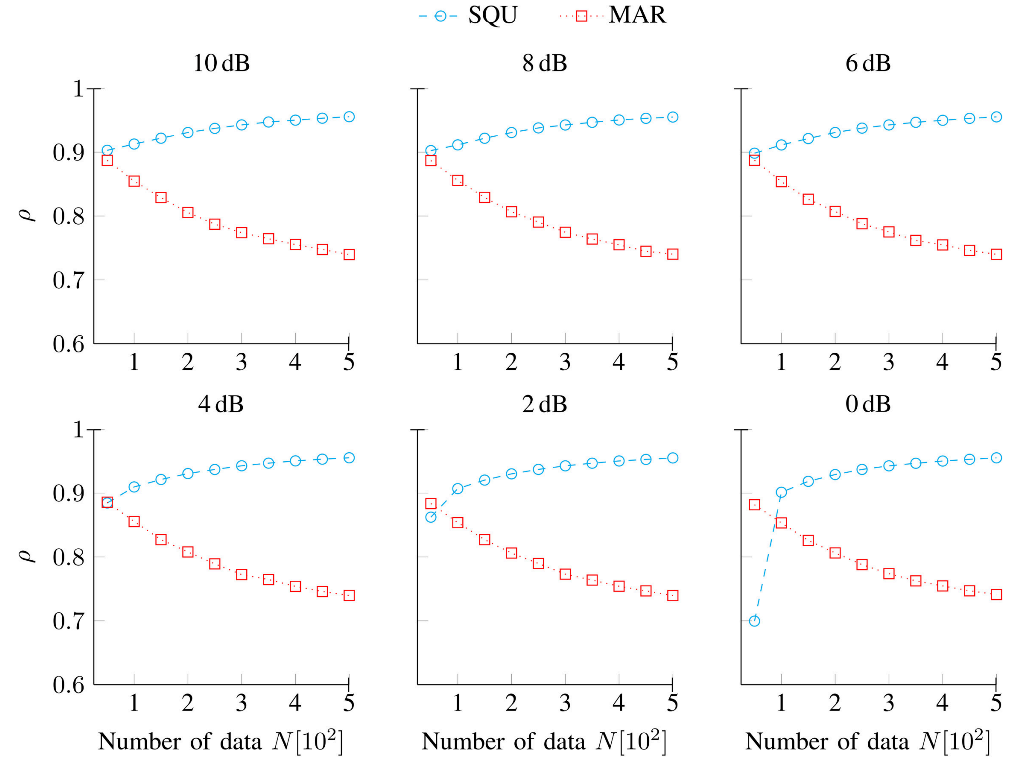

FIGURE 1 | Spearman correlations ρ (5) between predictors from the data-generating DIF model and predictors from the SQU and the MAR model, respectively. They are shown as a function of the number of data N and separately for noise conditions SNR [dB] $\in \{10, 8, 6, 4, 2, 0\}$.

function of the number of data N (per individual) and separately for noise conditions SNR [dB] $\in \{10, 8, 6, 4, 2, 0\}$. Overall, the values of these correlations are quite high, and they are largely independent from variations in data quality (SNR). It is clearly visible that the predictors from the DIF model and from the SQU model were generally much closer correlated than were the predictors from the DIF model and from the MAR model. As an exception from that rule, the DIF model and the SQU model were closer correlated than were the DIF model and the MAR model for the simulations under 2 dB and 0 dB for 50 trials. The DIF–MAR correlation under these circumstances may be attributable to the surprise metric (9) that both models incorporate, while the SQU does not make use of the surprise metric. Thus, for low data quality and low numbers of data points, the shared surprise metric seems to drive the dependency, whereas the model's parameter structure (multifactorial in case of the DIF and SQU models, unifactorial in case of the MAR model) is the stronger determinant of the dependency between the models under all other circumstances. Finally, the dissimilarity between the MAR and SQU model becomes more and more apparent with increasing numbers of data points N throughout all SNR conditions.

Figure 2 shows the explained variance R^2 (6) for the DIF (—×—), SQU (- ⊖ -), and MAR (··□··) models as a function of the number of data N for noise conditions SNR [dB] $\in \{10, 8, 6, 4, 2, 0\}$. As expected, the amount of explained variance decreased with decreasing SNRs (Penny, 2012). At highest levels of data quality, the maximum amount of explained variance approached around 25%, while at lowest levels of data quality, the maximum amount of explained variance approached less than 5%. Throughout the full range of SNRs and numbers of data points, the data-generating DIF model accounted for the maximum amount of variance. However, the DIF model's superiority in explaining variance decreased with decreasing SNRs. Finally, while the SQU model and the MAR model were clearly dissimilar with respect to their inter-correlations with the data-generating DIF model (see **Figure 1**), these two models were by-and-large indistinguishable in terms of the amount of variance that they accounted for.

Figure 3 shows the exceedance probabilities φ for the DIF (—×—), SQU (- ⊖ -), MAR (··□··), and NUL (- ◇ -) models as a function of the number of data N separately for noise conditions SNR [dB] $\in \{10, 8, 6, 4, 2, 0\}$. At the lowest numbers of data points N (between 50 and 200, depending on the SNR), the NUL model achieved maximum exceedance probabilities, while the MAR model never achieved maximum exceedance probabilities. At higher numbers of data points N, the data-generating DIF model rapidly achieved superiority for the highest levels of data quality (i.e., SNR condition 10 dB to 6 dB). At the lowest levels of data quality (i.e., SNR condition 4 dB to 0 dB), the SQU model transiently achieved higher exceedance probabilities than did the data-generating DIF model: At a level of data quality of 4 dB, this

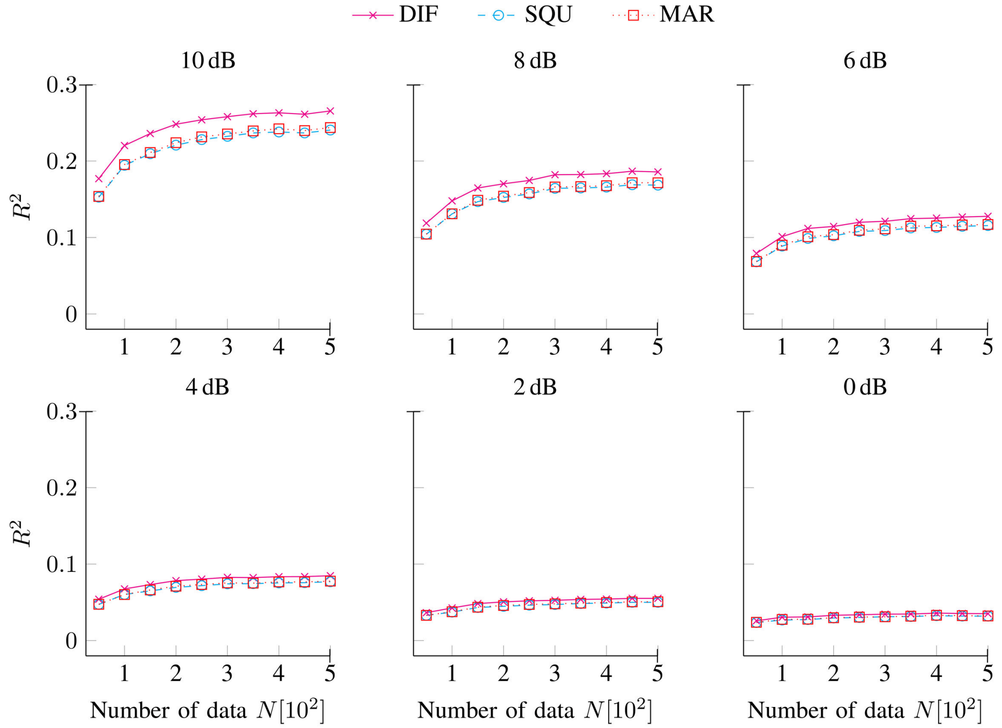

FIGURE 2 | Explained variance R^2 (6) for the DIF, SQU, and MAR models as a function of the number of data N separately for noise conditions SNR [dB] ∈ {10, 8, 6, 4, 2, 0}.

held true at $N = 150$ trials; at 2 dB, the range of SQU model superiority extended to $N = 200$ to 350 trials; and at 0 dB, the range of SQU model superiority extended to $N = 250$ to 500 trials.

Putatively, the SQU model's superiority across low SNRs and numbers of data points stems from two facts. First, the predictors from the data-generating DIF model and those from the SQU model were highly redundant (see **Figure 1**). Second, the SQU model incorporates event probabilities, while the data-generating DIF model estimates event probabilities via relative frequencies across trials, and this trial-by-trial variability contributes additional variance to the model's predictors. At relatively low data quality and at low numbers of data points, the PEB scheme probably misattributes this additional variability to noise rather than to the clean signal, rendering the SQU model superior to the data-generating DIF model under these circumstances.

4. DISCUSSION

The validity of forward modeling studies in CCN has, in the past, been culpably neglected in the literature albeit that topic is of utmost importance for our ability to identify proper computational models of cognition from studies of brain activity. Here, we showed in a synthetic EEG study that the validity of model selection varies with the data quality, with the numbers of data points, and with complexity and the dependency in the model space. **Figure 3** depicts the main findings of our simulation study in terms of exceedance probabilities, a main outcome measure of BMS.

To begin with, the least complex model (i.e., the NUL model) had a competitive advantage at very low numbers of data points throughout the full range of the SNRs that we examined. The data-generating (DIF) model could be easily identified even at relatively low numbers of data points (i.e., around 100 to 200 trials) when SNRs surmounted 7 dB. Below that point of data quality, the SQU model had a competitive advantage over the data-generating DIF model at intermediate (at 2 to 6 dB SNR) or even high (at 0 dB SNR) numbers of data points, such that the paradoxical advantage of the SQU model over the DIF model decreased with rising SNRs.

To summarize, we mis-identified a putative data-generating model at very low numbers of data points at all SNRs (i.e., the NUL model) and throughout intermediate, or even high, numbers of data points as a function of decreasing SNRs (i.e., the SQU model). On the other hand, it is important that we succeeded in identifying the data-generating (DIF) model, provided a sufficient SNRs and/or a sufficient number of data, while a more dissimilar model (i.e., the MAR model) remained less probable than the data-generating model throughout the full range of scenarios. The bottom line from our study is that simulation studies akin to our work should be made mandatory in designing and reporting CCN studies in order to substantiate, rather than merely to suppose, sufficient validity of any given forward modeling study, irrespective of its modality.

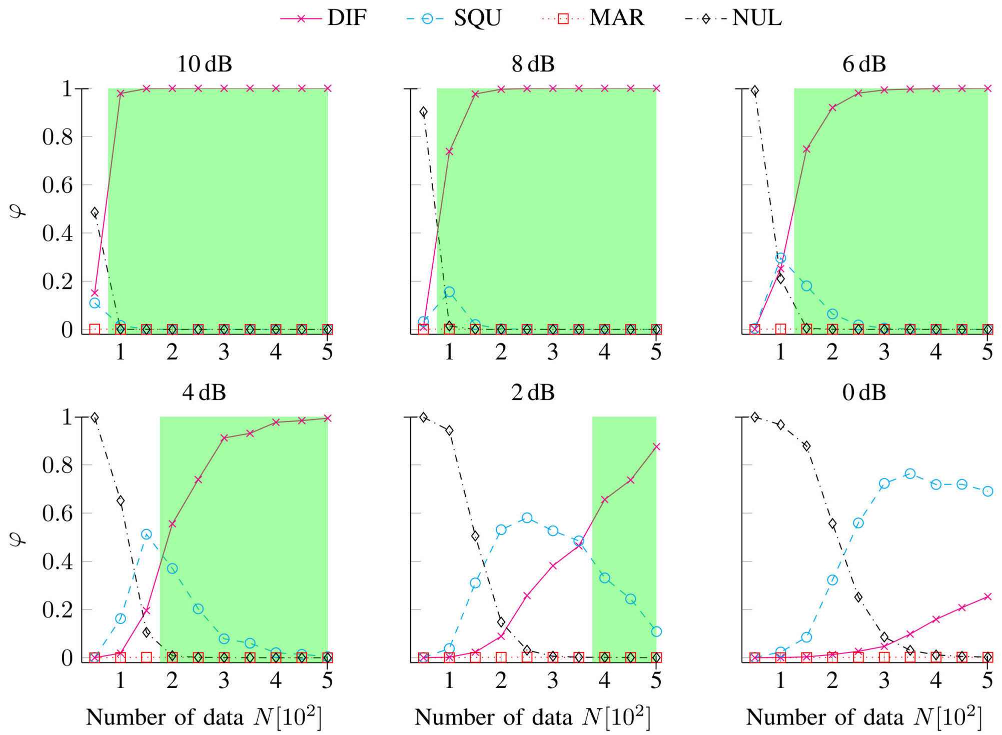

FIGURE 3 | Exceedance probabilities φ for the DIF, SQU, MAR, and NUL models as a function of the number of data N separately for noise conditions SNR [dB] $\in \{10, 8, 6, 4, 2, 0\}$. The green areas depict the range of valid inference (i.e., the maximum exceedance probability is assigned to the data-generating DIF model) separately for each SNR condition. It can be seen that the range of valid inference shrinks with decreasing SNRs such that no valid inference remains possible at the lowest level of data quality (i.e., at 0 dB) within the given numbers of data points.

One of the basic problems of CCN to date is that the number of data measured in typical forward modeling studies is usually planned without any formal consideration of data quality variations. Consequently, the effects of measurement error remain subject to variation which—as shown—affects the validity of the model selection. Several methods have been suggested for cleaning data (e.g., Turetsky et al., 1989; Effern et al., 2000; Quiroga, 2000; He et al., 2004; Gonzalez-Moreno et al., 2014; Ouyang et al., 2015). However, even though the average data quality can be improved, this does not compensate the insufficiency of the data in potentially many studies. Alternatively, it would also be possible to generally increase the number of data points toward high numbers. In practice, however, fatigue, for example, may affect the neuropsychological phenomena which are under scrutiny (Picton et al., 1995; Boksem et al., 2005; Muller-Gass et al., 2005; Thornton, 2008), and the risks of equipment-related errors also increase with time (Rahne et al., 2008).

Our study provided quantitative results to support the idea that the sufficiency of the number of data points can be better guaranteed by application of a synthetic validity test. For example, the number of data points ($N_\ell = 1152$) in the forward modeling study of Kolossa et al. (2013) was in fact sufficient for selecting between the SQU, MAR, and DIF models, given the empirical data quality (SNR $\approx$ 2 dB) since inspection of **Figure 3** reveals that such a model selection with an SNR = 2 dB falls within the range of valid inference if it is based on $N > 400$ data points per individual. In addition to data quality variations, the model complexity (see the comparison between the NUL vs. the DIF model) and the dependency of the model predictors (compare the comparison between the SQU vs. the DIF model and the comparison between the MAR vs. the DIF model) affect the sufficiency of the number of data points that are required in a valid forward modeling study.

The conductance of a synthetic validity test should incorporate four main variables, i.e., complexity of the models, dependency of the quantitative predictors from the model space, reasonable data quality variations, and feasible numbers of data points (trials). Synthetic validity tests answer the question whether a given number of data points is sufficient or not, given the particular model space under consideration, and given specific assumptions about data quality.

One of our reviewers raised a concern, namely that our explanation why the SQU model (i.e., not the model that was actually used to simulate the data) won the model comparison at very low SNR (and insufficient numbers of data points). We had argued that this can be explained by the way the model works, i.e., in terms of the way the SQU model incorporates event probabilities. An alternative reason why the SQU model might have won our BMS may be related to the hierarchical priors of the PEB approach (Friston et al., 2002; see e.g., Boos et al., 2016 for an example of hierarchical Bayesian modeling). In particular, the prior on the group variance (i.e., the variance of $\epsilon^{(2)}$ in 3) might induce more or less "shrinkage around the mean" on PEB estimates, eventually favoring the wrong model in low SNR situations. Thus, the specification of different hierarchical priors at the group level constitutes a variable, which was not explored systematically in our study, but which can be examined in appropriate follow-up studies. In these studies, one could also construct a factorial model space, where DIF, SQU, MAR and NUL would be one model dimension, and different prior variances would induce a orthogonal dimensions. One could then marginalize over prior variances to obtain family-wise exceedance probabilities, which do not depend upon the hierarchical priors (Penny et al., 2010). The same reviewer made the point that exceedance probabilities are but one of many summary statistics in BMS, including posterior estimates of model frequencies and protected exceedance probabilities (Rigoux et al., 2014) that are associated with different levels of statistical risk.

The reviewer also raised the concern that our approach misses a critical aspect of group studies, where data quality refers to the number of trials N and data quantity to the number of subjects L. Our study does not provide an answer to the question whether one should use, for example, two subjects with 300 trials each (maximizing data quality), or 20 subjects with 30 trials each (maximizing data quantitiy; e.g., Maus et al., 2011). A conceivable extension of our study to evaluate if BMS is more sensitive to data quality or data quantity would be varying (in a factorial way) the within-subject SNR (and/or number of trials per subject) and the group sample size, both chosen within typical ranges. Still another extension would be variations in group heterogeneity (e.g., a group could be composed of individuals best described by different models) which is of particular importance for random-effects BMS. A related issue is the clinical application of BMS, because clinical populations typically differ from normal control populations with regard to SNR (Sackett, 2001; Winterer and Weinberger, 2003), demanding additional strategies for paralleling the SNRs that can be obtained in clinical and normal samples.

We advocated here the idea of using numerical simulations to aid the interpretation of BMS. We showed that one should be cautious about the results of BMS, in case these simulations detect that some of the models may be confused with each other (as is the case for the DIF, SQU, and NUL models here). However, we have not formalized how one would (formally and/or practically) use this information to scaffold one's BMS. In other words: how should one integrate the results of a confusion analysis (derived from realistic numerical simulations) with one's BMS results (performed on experimental data)?

Our idea of conducting a confusion analysis during the design phase of an experiment can be extended to address this issue, as suggested by the reviewer (see Text S1 in Devaine et al., 2014 or Marković and Kiebel, 2016 for examples). To that end, one would derive the full quadratic confusion matrices $\mathbf{C} \in \mathbb{R}^{M\times M}$, with M denoting the number of models in the model space $\mathcal{M}$. This $M \times M$ confusion matrix yields the exceedence probabilities of having inferred each model, having simulated the data under each model (not just under one of the models, as in our simulation). In such a confusion matrix, the elements on the main diagonal represent the probability of inferring the true (data-generating) model, while the non-diagonal elements represent the probability of inferring a model that did not generate the data. Non-diagnonal elements in this confusion matrix signal potential confusions between the inferred model and the true (data-generating) model; hence, perfect model identifiability should exhibit no extra-diagonal non-zero element, i.e., an identity matrix.

There are many criteria conceivable, which may serve as minimum standards for acceptable levels of model identifiability. An exemplary cut-off criterion may be seen in the requirement that all diagonal elements > 0.50 (or alternatively, any other value above 0.50). In this case, however, the chosen cut-off value should strongly depend on the size of the model space. The determinant of the confusion matrix $|\mathbf{C}|$ may be considered as a more sophisticated approach to quantify model identifiability, because, e.g., $|\mathbf{C}| = 1$ for perfect identifiability, while $|\mathbf{C}| = 0$ if all models are equally probable. The dependency of the determinant of the confusion matrix on data quality and quantity should be analyzed during a-priori examination of model identifiability, because for any given level of data quantity $L = L'$ we get $|\mathbf{C}|\big|_{L = L', N\to\infty} = 1$, while the value of $0 \leq |\mathbf{C}|\big|_{L\to\infty, N = N'} \leq 1$ depends on the given level of data quality $N = N'$. Minimum determinants may be defined as cut-off criteria, with lower limits being associated with less confidence in the conclusions that can be drawn from a forward modeling study. While one may leave the choice of a particular cut-off criterion for acceptable levels of model identifiability to the discretion of the authors, CCN would certainly profit from such an explicit treatment of a-priori model identifiability.

But what if the most plausible model for one's experimental data is easily confused with another model? As far as we know, there are no existing solutions to this issue once the experiment has already been carried out. However, the a-priori calculation of confusion matrices renders it possible to quantify the overall risk of model confusion, which decreases with increasing data quantity and quality. These calculations enable one to conduct a feasible experiment, while controlling for the overall risk of model confusion, as discussed above. However, it may simply not be feasible to collect data with sufficient levels of quality and quantity to surmount the pre-defined cut-off criteria. In this case, the model space may be partitioned into model families, and reasonable amounts of data may suffice for an acceptable confusion risk within the respective model families.

Another solution to this problem would be to strengthen the informativeness of the experimental design, e.g., by applying the technique of adaptive design optimization as proposed in cognitive science (Myung and Pitt, 2009; Cavagnaro et al., 2010, 2011; Myung et al., 2013; Kim et al., 2014). In the context of model identifiability in BMS, adaptive design optimization implies maximizing the determinants of the $M \times M$ confusion matrices under fixed levels of data quality and quantity. This goal can be achieved through the employment of the experimental design that yields the best possible discrimination between model outputs. Engineering solutions for system identification rely on pre-experimental optimization as well, e.g., pseudo random input sequences in non-linear system identification (Billings and Fakhouri, 1980; Vincent et al., 2010) or so-called perfect sequences in acoustic system identification (Ipatov, 1979; Lüke and Schotten, 1995). In the context of CCN, these techniques may be employed to guarantee maximum orthogonality between model outputs, which would enable BMS to better discriminate between the models. Given that these two proposals complement each other, one should first optimize the experimental design, and subsequently analyze minimum levels of data quality and quantity that are necessary for acceptable levels of model identifiability. Those are some of the routes for more systematic efforts toward a-priori calculation of confusion matrices, which may eventually lead to novel solutions to the problem of model identifiability that begins to fan out.

Despite the discussed shortcomings of our work, we recommend for researchers who plan to conduct a forward modeling CCN study, to run an unsolicited a-priori synthetic validity test in order to guarantee sufficiency of the to be gathered data. We further propose that this kind of synthetic validity tests should be made mandatory to all forward modeling studies in the future with the goal to improve the validity of these CCN studies.

AUTHOR CONTRIBUTIONS

AK and BK conceptualized the study. AK programmed and performed the simulations. AK and BK wrote the manuscript.

REFERENCES

Adams, R. A., Huys, Q. J., and Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psychiatry* 87, 53–63. doi: 10.1136/jnnp-2015-310737

Başar, E. (1980). *EEG-brain Dynamics: Relation Between EEG and Brain Evoked Potentials.* Amsterdam, NL: Elsevier-North-Holland Biomedical Press.

Beal, M. J., and Ghahramani, Z. (2003). "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics 7*, Vol. 7. eds J. M. Bernado, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, M. West (Oxford: Oxford University Press), 453–464.

beim Graben, P. (2001). Estimating and improving the signal-to-noise ratio of time series by symbolic dynamics. *Phys. Rev. E* 64:051104. doi: 10.1103/PhysRevE.64.051104

Billings, S. A., and Fakhouri, S. Y. (1980). Identification of non-linear systems using correlation analysis and pseudorandom inputs. *Int. J. Syst. Sci.* 11, 261–279. doi: 10.1080/00207728008967012

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., and Curio, G. (2007). The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage* 37, 539–550. doi: 10.1016/j.neuroimage.2007.01.051

Boksem, M. A., Meijman, T. F., and Lorist, M. M. (2005). Effects of mental fatigue on attention: an ERP study. *Cogn. Brain Res.* 25, 107–116. doi: 10.1016/j.cogbrainres.2005.04.011

Boos, M., Seer, C., Lange, F., and Kopp, B. (2016). Probabilistic inference: task dependency and individual differences of probability weighting revealed by hierarchical Bayesian modeling. *Front. Psychol.* 7:755. doi: 10.3389/fpsyg.2016.00755

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., and Kujala, J. V. (2010). Adaptive design optimization: a mutual information-based approach to model discrimination in cognitive science. *Neural Comput.* 22, 887–905. doi: 10.1162/neco.2009.02-09-959

Cavagnaro, D. R., Pitt, M. A., and Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychon. Bull. Rev.* 18, 204–210. doi: 10.3758/s13423-010-0030-4

Coppola, R., Tabor, R., and Buchsbaum, M. S. (1978). Signal to noise ratio and response variability measurements in single trial evoked potentials. *Electroencephalogr. Clin. Neurophysiol.* 44, 214–222. doi: 10.1016/0013-4694(78)90267-5

Corlett, P. R., and Fletcher, P. C. (2014). Computational psychiatry: a Rosetta Stone linking the brain to mental illness. *Lancet Psychiatry* 1, 399–402. doi: 10.1016/S2215-0366(14)70298-6

Czanner, G., Sarma, S. V., Ba, D., Eden, U. T., Wu, W., Eskandar, E., et al. (2015). Measuring the signal-to-noise ratio of a neuron. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7141–7146. doi: 10.1073/pnas.1505545112

Devaine, M., Hollard, G., and Daunizeau, J. (2014). The social Bayesian brain: does mentalizing make a difference when we learn? *PLoS Comput. Biol.* 10:e1003992. doi: 10.1371/journal.pcbi.1003992

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

Effern, A., Lehnertz, K., Fernández, G., Grunwald, T., David, P., and Elger, C. (2000). Single trial analysis of event related potentials: non-linear de-noising with wavelets. *Clin. Neurophysiol.* 111, 2255–2263. doi: 10.1016/S1388-2457(00)00463-6

Friston, K. J. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

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. doi: 10.1016/j.neuroimage.2006.08.035

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. doi: 10.1006/nimg.2002.1090

Gerstner, W., and Frémaux, N. (2015). Neuromodulated spike-timing-dependent plasticity and theory of three-factor learning rules. *Front. Neural Circuits* 9:85. doi: 10.3389/fncir.2015.00085

Gonzalez-Moreno, A., Aurtenetxe, S., Lopez-Garcia, M.-E., del Pozo, F., Maestu, F., and Nevado, A. (2014). Signal-to-noise ratio of the MEG signal after preprocessing. *J. Neurosci. Methods* 222, 56–61. doi: 10.1016/j.jneumeth.2013.10.019

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* 87, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

He, P., Wilson, G., and Russell, C. (2004). Removal of ocular artifacts from electro-encephalogram by adaptive filtering. *Med. Biol. Eng. Comput.* 42, 407–412. doi: 10.1007/BF02344717

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Stat. Sci.* 14, 382–401.

Hoijtink, H., Klugkist, I., and Boelen, P. A. (2008). *Bayesian Evaluation of Informative Hypotheses*. New York, NY: Springer.

Huys, Q. J. M., Maia, T. V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* 19, 404–413. doi: 10.1038/nn.4238

Ipatov, V. P. (1979). Ternary sequences with ideal periodic autocorrelation properties. *Radio Eng. Electr. Phys.* 24, 75–79.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Dev. Sci.* 10, 307–321. doi: 10.1111/j.1467-7687.2007.00585.x

Kim, W., Pitt, M. A., Lu, Z. L., Steyvers, M., and Myung, J. I. (2014). A hierarchical adaptive approach to optimal experimental design. *Neural Comput.* 26, 2465–2492. doi: 10.1162/NECO_a_00654

Kira, S., Yang, T., and Shadlen, M. N. (2015). A neural implementation of Wald's sequential probability ratio test. *Neuron* 85, 861–873. doi: 10.1016/j.neuron.2015.01.007

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. doi: 10.1016/j.tins.2004.10.007

Koechlin, E. (2014). An evolutionary computational theory of prefrontal executive function in decision-making. *Philos. Trans. R. Soc. B* 369:20130474. doi: 10.1098/rstb.2013.0474

Kolossa, A. (2016). *Computational Modeling of Neural Activities for Statistical Inference.* Cham: Springer.

Kolossa, A., Abel, J., and Fingscheidt, T. (2016). "Comparing instrumental measures of speech quality using Bayesian model selection: correlations can be misleading!," in *Procedings of ICASSP 2016* (Shanghai), 634–638.

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

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

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

Luck, S. J. (2004). "Ten simple rules for designing and interpreting ERP experiments," in *Event-Related Potentials: A Methods Handbook*, ed T. C. Handy (Cambridge, MA: MIT Press), 17–32.

Luck, S. J. (2014). *An Introduction to the Event-related Potential Technique.* Cambridge, MA: MIT Press.

Lüke, H. D., and Schotten, H. D. (1995). Odd-perfect, almost binary correlation sequences. *IEEE Trans. Aerospace Electron. Syst.* 31, 495–498. doi: 10.1109/7.366335

Makeig, S., Bell, A. J., Jung, T.-P., Sejnowski, T. J., et al. (1996). Independent component analysis of electroencephalographic data. *Adv. Neural Inform. Proc. Syst.* 8, 145–151.

Marković, D., and Kiebel, S. J. (2016). Comparative analysis of behavioral models for adaptive learning in changing environments. *Front. Comput. Neurosci.* 10:33. doi: 10.3389/fncom.2016.00033

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

Maus, B., Van Breukelen, G. J. P., Goebel, R., and Berger, M. P. F. (2011). Optimal design of multi-subject blocked fMRI experiments. *NeuroImage* 56, 1338–1352. doi: 10.1016/j.neuroimage.2011.03.019

Möcks, J., Gasser, T., and Köhler, W. (1988). Basic statistical parameters of event-related potentials. *J. Psychophysiol.* 2, 61–70.

Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80. doi: 10.1016/j.tics.2011.11.018

Muller-Gass, A., Stelmack, R. M., and Campbell, K. B. (2005). "...and were instructed to read a self-selected book while ignoring the auditory stimuli": the effects of task demands on the mismatch negativity. *Clin. Neurophysiol.* 116, 2142–2152. doi: 10.1016/j.clinph.2005.05.012

Myung, J. I., Cavagnaro, D. R., and Pitt, M. A. (2013). A tutorial on adaptive design optimization. *J. Math. Psychol.* 57, 53–67. doi: 10.1016/j.jmp.2013.05.005

Myung, J. I., and Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychol. Rev.* 116, 499–518. doi: 10.1037/a0016104

Nicolelis, M. A. L. (2001). Actions from thoughts. *Nature* 409, 403–407. doi: 10.1038/35053191

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., and Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proc. Natl. Acad. Sci. U.S.A.* 110, E3660–E3669. doi: 10.1073/pnas.1305373110

O'Reilly, R. C., Munakata, Y., Frank, M., and Hazy, T. (2012). *Computational Cognitive Neuroscience. Wiki Book, 1st Edn*. Available online at: http://ccnbook.colorado.edu

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

Ouyang, G., Sommer, W., and Zhou, C. (2015). A toolbox for residue iteration decomposition (RIDE) — a method for the decomposition, reconstruction, and single trial analysis of event related potentials. *J. Neurosci. Methods* 250, 7–21. doi: 10.1016/j.jneumeth.2014.10.009

Paukkunen, A. K. O., Leminen, M. M., and Sepponen, R. (2010). Development of a method to compensate for signal quality variations in repeated auditory event-related potential recordings. *Front. Neuroengineering* 3:2. doi: 10.3389/fneng.2010.00002

Pecevski, D., and Maass, W. (2016). Learning probabilistic inference through STDP. *eNeuro* 3, 1–35. doi: 10.1523/ENEURO.0048-15.2016

Penny, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *NeuroImage* 59, 319–330. doi: 10.1016/j.neuroimage.2011.07.039

Penny, W. D., and Roberts, S. J. (2002). Bayesian multivariate autoregressive models with structured priors. *IEE Proc. Vis. Image Signal Process.* 149, 33–41. doi: 10.1049/ip-vis:20020149

Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., et al. (2010). Comparing families of dynamic causal models. *PLoS Comput. Biol.* 6:e1000709. doi: 10.1371/journal.pcbi.1000709

Penny, W. D., Stephan, K. E., Mechelli, A., and Friston, K. J. (2004). Comparing dynamic causal models. *NeuroImage* 22, 1157–1172. doi: 10.1016/j.neuroimage.2004.03.026

Picton, T. W., Lins, O. G., and Scherg, M. (1995). The recording and analysis of event-related potentials. *Handbook Neuropsychol.* 10, 3–73.

Pitt, M. A., and Myung, I. J. (2002). When a good fit can be bad. *Trends Cogn. Sci.* 6, 421–425. doi: 10.1016/S1364-6613(02)01964-2

Puce, A., Berkovic, S. F., Cadusch, P. J., and Bladin, P. F. (1994). P3 latency jitter assessed using 2 techniques. I. Simulated data and surface recordings in normal subjects. *Electroencephalogr. Clin. Neurophysiol.* 92, 352–364. doi: 10.1016/0168-5597(94)90103-1

Quiroga, R. Q. (2000). Obtaining single stimulus evoked potentials with wavelet denoising. *Physica D* 145, 278–292. doi: 10.1016/S0167-2789(00)00116-0

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–164. doi: 10.2307/271063

Rahne, T., von Specht, H., and Mühler, R. (2008). Sorted averaging-application to auditory event-related responses. *J. Neurosci. Methods* 172, 74–78. doi: 10.1016/j.jneumeth.2008.04.006

Raz, J., Turetsky, B., and Fein, G. (1988). Confidence intervals for the signal-to-noise ratio when a signal embedded in noise is observed over repeated trials. *IEEE Trans. Biomed. Eng.* 35, 646–649. doi: 10.1109/10.4598

Rigoux, L., Stephan, K. E., Friston, K. J., and Daunizeau, J. (2014). Bayesian model selection for group studies–revisited. *NeuroImage* 84, 971–985. doi: 10.1016/j.neuroimage.2013.08.065

Sackett, D. L. (2001). Why randomized controlled trials fail but needn't: 2. failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). *Can. Med. Assoc. J.* 165, 1226–1237.

Schimmel, H. (1967). The (±) reference: accuracy of estimated mean components in average response studies. *Science* 157, 92–94. doi: 10.1126/science.157.3784.92

Sejnowski, T. J., Koch, C., and Churchland, P. S. (1988). Computational neuroscience. *Science* 241, 1299–1306. doi: 10.1126/science.3045969

Shannon, C. E., and Weaver, W. (1948). The mathematical theory of communication. *Commun. Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

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. doi: 10.1126/science.959831

Stephan, K. E., and Mathys, C. (2014). Computational approaches to psychiatry. *Curr. Opin. Neurobiol.* 25, 85–92. doi: 10.1016/j.conb.2013.12.007

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage* 46, 1004–1017. doi: 10.1016/j.neuroimage.2009.03.025

Sutton, S., Braren, M., Zubin, J., and John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science* 150, 1187–1188. doi: 10.1126/science.150.3700.1187

Thornton, A. R. (2008). Evaluation of a technique to measure latency jitter in event-related potentials. *J. Neurosci. Methods* 168, 248–255. doi: 10.1016/j.jneumeth.2007.09.031

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interf.* 6, 187–202. doi: 10.1098/rsif.2008.0172

Turetsky, B. I., Raz, J., and Fein, G. (1989). Estimation of trial-to-trial variation in evoked potential signals by smoothing across trials. *Psychophysiology* 26, 700–712. doi: 10.1111/j.1469-8986.1989.tb03176.x

Vincent, T. L., Novara, C., Hsu, K., and Poolla, K. (2010). Input design for structured nonlinear system identification. *Automatica* 46, 990–998. doi: 10.1016/j.automatica.2010.02.029

Vyshemirsky, V., and Girolami, M. A. (2008). Bayesian ranking of biochemical system models. *Bioinformatics* 24, 833–839. doi: 10.1093/bioinformatics/btm607

Winterer, G., and Weinberger, D. R. (2003). Cortical signal-to-noise ratio: insight into the pathophysiology and genetics of schizophrenia. *Clin. Neurosci. Res.* 3, 55–66. doi: 10.1016/S1566-2772(03)00019-7

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.