32 Mind the Traps! Design Guidelines for Rigorous BCI Experiments

Camille Jeunet, Stefan Debener, Fabien Lotte, Jérémie Mattout, Reinhold Scherer, and Catharina Zich

CONTENTS
32.1 Introduction .......................................................................................................................... 614
32.2 Acquisition of the Signals .....................................................................................................615
  32.2.1 How to Choose the Appropriate Sensor Type, Location, and Number
        Depending on What We Want to Measure .................................................................... 615
  32.2.2 What Can We Infer from the Activation of One Electrode? Concept
        of Spatial Resolution ......................................................................................... 616
  32.2.3 How Muscular Activity Can Interfere with EEG Activity and Why It Should
        Be Controlled for to Avoid Confounds ..................................................................617
  32.2.4 What Kinds of Mental States Can Be Used? ............................................................618
32.3 Data Processing .................................................................................................................... 619
  32.3.1 Which Classifier Can We Use Depending on the Distribution of the Data? .......... 619
  32.3.2 Why Is It Important to Separate the Training Data Set from the Testing Data Set?... 620
  32.3.3 How to Determine the Chance Level for the Classified Data .................................. 620
  32.3.4 To Which Extent Can Commercial Algorithms Be Trusted? ...................................622
  32.3.5 How to Determine the Relevance of Neurophysiological Markers ..........................622
  32.3.6 Why Is It Important to Correct for Multiple Comparisons? ...................................623
32.4 Experimental Design and the User Component ...................................................................624
  32.4.1 What to Have in Mind When Designing a New BCI Experiment ........................... 624
  32.4.2 Why and How to Have a Good Control Group? .......................................................625
  32.4.3 How to Avoid Biases: Concepts of Counterbalancing, Sham Control, Double
        Blindness, and Randomization ..................................................................................626
  32.4.4 How to Select the Appropriate Statistical Tests .......................................................627
32.5 Summary and Conclusion .....................................................................................................629
Acknowledgments ..........................................................................................................................630
References ......................................................................................................................................630

Abstract

Designing brain–computer interface (BCI) experiments requires knowledge in many different disciplines: from neurosciences to signal processing and machine learning, through psychology. However, very few people have skills in all these disciplines. Yet, a lack of knowledge in a single aspect of BCIs is likely to result in flaws in the experimental design, statistical analyses, or the interpretation of the results. Moreover, because the BCI field is relatively
young, no widely accepted guidelines are available yet, while at the same time an exponentially increasing number of research teams contribute to this field and would benefit from such guidelines. Thus, the objective of this chapter is to propose, in a pedagogical way, step-by-step guidance to design a rigorous BCI experiment. We name potential pitfalls and explain how to avoid them using concrete examples. This chapter could be seen as a checklist of points that should be addressed when the aim is to design rigorous and scientifically valid BCI studies and experiments. It is structured into three categories: (1) the acquisition of brain signals, (2) data processing issues, and (3) the experimental design and consideration of the BCI user.

32.1 INTRODUCTION

Brain–computer interfaces (BCIs) enable users to control an application using their brain activity alone. Such control can be achieved following different paradigms. Mainly, BCIs can be active, reactive, or passive. Active BCIs rely on modifications of the amplitude of brain rhythms in different frequency bands while users perform an explicit task, such as mental imagery. The most common active BCI paradigm is based on motor imagery: users are asked to perform a motor imagery task (i.e., to imagine movements of their limbs), which lowers the amplitude of sensorimotor rhythms (SMRs) in the mu (8–12 Hz) and beta (12–30 Hz) frequency bands in the sensorimotor cortex. By detecting this spatio-spectral signal change, the system may be able to infer the mental task performed by the user and in return associate to it a command for the application. For instance, by performing left- and right-hand motor imagery tasks, one can make a wheelchair turn left or right, respectively (Millán et al. 2010). On the other hand, reactive BCIs rely on the detection of brain potentials generated in response to a stimulus. The most popular BCI application is probably the P300 speller (Farwell and Donchin 1988). The P300 speller was first designed to enable communication for the paralyzed. A matrix of letters is displayed on a screen. The rows and columns of this matrix flash sequentially in a random order. The user has to focus on the letter he or she wants to spell. When this letter flashes, a positive cortical potential is generated around 300 ms later. This way, after several repetitions, by intersecting the rows and columns, the flashing of which triggered a P300, the system will be able to infer the target letter. Finally, passive BCIs are systems that enable the user’s mental state to be measured in order to adapt an application/interface accordingly (Zander and Kothe 2011). Here, users do not follow specific instructions or voluntary interact with the BCI; in other words, they do not send conscious commands. Instead, their cognitive (e.g., workload), emotional (e.g., frustration), or motivational states are inferred from their ongoing electroencephalography (EEG) signals, which can also be related to other physiological and behavioral data to identify a particular mental state. A passive BCI application automatically adapts to fluctuations in user states.

BCIs are very promising for a wide range of applications, from assistive technologies to communication devices. Beyond their obvious potential for patients with motor impairments, BCIs also offer new possibilities in different fields, ranging from sports to video games (Lécuyer et al. 2008) through education (Frey et al. 2014) and rehabilitation (Kübler et al. 2013) to name only a few examples. Thus, not only engineers and psychologists are interested in using such technologies, but also medical doctors, neuroscientists, teachers, or specialists in sports science, business, physiotherapy or linguistics, between many others. Yet, designing BCI experiments requires knowledge in many different disciplines: neurosciences to understand the properties of the brain signals used to control the BCI, signal processing to extract the relevant information from these brain signals, machine learning to make the system able to learn which brain patterns correspond to which mental command, psychology to understand the factors influencing users’ ability to control a BCI, and human–computer interaction to design usable and efficient BCI protocols. However, very few people have skills in all these disciplines. Yet, a lack of knowledge in a single aspect of BCIs is likely to result in flaws in the experimental design, statistical analyses, or the interpretation of the results. Moreover, because the BCI field is relatively young, no widely accepted guidelines are available yet, while at the same time an exponentially increasing number of research teams contribute to this field and thus would benefit from such guidelines.
The objective of this chapter is to propose, in a pedagogical way, step-by-step guidance to design a
rigorous BCI experiment. We do not propose a perfect experimental design, but rather name potential
pitfalls, and explain how to avoid them. This chapter could be seen as a checklist of points that should
be addressed when the aim is to design rigorous and scientifically valid BCI studies and experiments.
Therefore, this chapter targets more specifically the numerous researchers, professionals, passionate
people, or patients and their families who start using BCIs and want to avoid the current traps sur-
rounding BCIs. This chapter does not claim to be exhaustive but rather aims (1) at naming the impor-
tant elements to be considered while designing a BCI experiment and (2) at guiding the reader toward
relevant pieces of literature if they want to investigate further some specific elements.

The next part of this chapter is structured into three sections. The first section focuses on the
acquisition of brain signals. Here, we will focus on EEG-based BCIs, simply because this is the
most popular input modality for BCIs. The second section deals with data processing issues, while
the third one will introduce important points concerning the experimental design and the target
individuals, the BCI users.

To illustrate these different points, we propose to design a BCI experiment all along the chapter,
following the recommendations point by point. Let us say that our research question is the following:

We want to evaluate the relevance and reliability of new tasks to control a Mental Imagery–based BCI
(MI-BCI), namely, “remembering a positive vs. negative emotional souvenir.”

The object of providing this example is not to undertake a real research study or to answer a spe-
cific research question. Thus, no data or statistical analyses will be described. Rather, the aim is to
provide the reader with a clear application so that they know in which context the recommendations
could be applied. For specific applications of BCIs for therapeutic purposes, please refer to Guger

At the end of each subsection, we will discuss which tools/methods are the most relevant in the
context of this research question.

32.2 ACQUISITION OF THE SIGNALS

32.2.1 HOW TO CHOOSE THE APPROPRIATE SENSOR TYPE, LOCATION,
AND NUMBER DEPENDING ON WHAT WE WANT TO MEASURE

Typical EEG signals refer to voltage fluctuations in the direct current to approximately 40 Hz fre-
cquency range, with amplitudes ranging from a few tens of microvolts to below 1 μV. To capture
these miniature signals, sensors are placed on the scalp and a conducting gel is applied to lower the
skin-electrode impedance to approximately <20 kΩ. However, wet sensors have the disadvantage
that individuals have to wash their hair after the EEG recording, to remove the conductive gel.
Therefore, new sensor types have been developed, such as active wet electrodes, active dry elec-
trodes, or miniature sensors requiring no or very small amounts of gel only (Debener et al. 2015).
Typical EEG recordings require the use of several electrodes (the minimum number is 3, common
are 32, 64, 128, or even 256 channel recordings), which are placed with a cap or net on the scalp.
Different sensor types have their advantages and disadvantages. The signal quality and wearing
comfort of dry electrodes, for instance, are typically inferior to conventional wet electrodes, but the
benefits are the faster setup time and that hair washing is not required after the end of the record-
ing. While alternative materials and electrodes may suffice for some EEG signals, we recommend
the use of sintered Ag/AgCl electrodes, which, when used as wet electrodes and applied correctly,
provide good signal characteristics. They do not generate voltage fluctuations on their own and do
not cause frequency distortion, which are two problems that should be avoided when measuring
biosignals with microvolt amplitude. The international 10–20 system is the standard for electrode
placement. Depending on the type of brain signals that should be recorded, appropriate electrode locations should be used.

Which sensor positions are important? The answer depends on the class of BCI that one wants to implement. However, as outlined in Section 32.2.2, a good spatial sampling is generally helpful even if only a few channels are used. For BCIs that detect sensory evoked responses, electrodes should be placed such that known topographical representations of sensory evoked potential are captured. This includes placing electrodes over posterior and occipital sites for capturing visual evoked responses and placing electrodes over fronto-central sites for capturing auditory evoked responses. However, the most discriminative information does not necessarily overlap spatially with locations giving the best signal-to-noise ratio. Moreover, to optimally classify signals, it is also helpful to place electrodes away from the signal of interest, in order to cancel out irrelevant activity, as explained by Blankertz et al. (2011). Accordingly, for BCIs that detect the neural correlates of motor imagery, electrodes should be located over sensorimotor areas, and, in order to disentangle sensorimotor mu from occipital alpha,* some sensors should be placed over posterior scalp sites as well.

In summary, multichannel EEG (32+ channels) acquisition from sensors covering wider parts of the head is beneficial, although for practical applications and for economic reasons, a limited setup is often used, under ideal circumstance, without much loss of performance.

And for our experiment? In our case, since we do not have a precise idea of the relevant brain areas for the tasks (remembering positive/negative emotional souvenirs), we will use at least 32 or 64 electrodes, placed all over the scalp. Also, because we want to maximize signal quality, we will use wet Ag/AgCl electrodes (i.e., with conductive gel).

32.2.2 WHAT CAN WE INFER FROM THE ACTIVATION OF ONE ELECTRODE?
CONCEPT OF SPATIAL RESOLUTION

It is important to understand the concept of differential amplification when voltage is measured. EEG signals reflect voltage fluctuations over time, and voltage fluctuations can be best captured as the difference between two locations, say one electrode placed on the top of the head (vertex) and another one behind the ear (mastoid). Even though one electrode is often defined as the reference electrode (mastoid), the recorded signal cannot be regarded as reflecting electrical activity from the patch of brain underneath the other electrode (vertex). Likewise, since the whole body conducts current fairly well (with different tissues having different conductivity properties), it is not required to place electrodes close to the heart to measure the electrocardiogram. The reason for that is that EEG measures the synchronized electrical activity of adjacent pyramidal cells aligned in parallel, which rise to electrical fields that are strong enough to be captured with electrode placed on the skin. Accordingly, all brain signals captured by EEG result from relatively large and highly synchronized patches of cortex. Each patch of cortex contributing to the EEG may be best regarded as an equivalent current dipole. If the aim is to capture the electrical activity of such a dipole, it is important to understand that dipole orientation is at least as important as dipole location. In fact, two electrodes placed very close to a dipolar generator may not record any activity if placed in the wrong orientation. On the other hand, two electrodes placed further away from the generator may capture its signature nicely.

Given that an unknown number of brain (and non-brain, e.g., electrocardiogram) generators contribute to the scalp EEG, one can safely assume that the number of generators contributing to the mixed recorded signal is much higher than the number of channels used to record the signals, even if high-density EEG acquisition is performed. The inverse problem means that one cannot determine the brain source of any particular EEG signal for sure. Inferring the number and locations of active brain regions that caused the recorded EEG signals on scalp is sometimes compared to

* The 8–12 Hz frequency range in the literature is referred to as mu rhythm when related to sensorimotor activity; otherwise, it is referred to as alpha rhythm.
the problem of guessing what 3D shape created an observed shadow on a wall. It is a so-called ill-posed problem that has no unique solution. However, good guesses are possible, and EEG source localization and spatial filtering procedures—which combine the signals from multiple electrodes; see Blankertz (2008) and Chapter 18 (“Gentle Introduction to Signal Processing and Classification for Single-Trial EEG Analysis”)—can be used to confirm ideas about brain sources contributing to the EEG, with reasonable spatial resolution and precision. To this end, making use of multiple EEG channels can help. The rich spatial detail of multichannel recordings has two key advantages. First, it is much easier to disentangle brain from non-brain contributions to the measured EEG signal, and second, it is much easier to identify, and disentangle, different brain signals from each other.

And for our experiment? To be able to disentangle the brain area(s) involved in each of the tasks, we will apply, offline, source localization procedures (i.e., source reconstruction algorithms).

### 32.2.3 How Muscular Activity Can Interfere with EEG Activity and Why It Should Be Controlled for to Avoid Confounds

As stated in Section 32.2.2, EEG sensors measure all electrical currents at the sensor location, not only those coming from the brain. In particular, muscle tensions and eye movements generate electrical currents, known respectively as electromyography (EMG) and electrooculography (EOG), that are of much larger magnitude than currents of cortical origin, that is, signals originating from the brain (Fatourechi et al. 2007) (see Figure 32.1). EOG signals may dominate the EEG at frontal scalp sites (near the eyes), while EMG signals are common for electrodes placed near muscles (Goncharova et al. 2003), contributing broad band and in particular high-frequency activity (Whitham et al. 2007).

EOG/EMG artifacts can corrupt EEG signals and result in poor BCI performance. However, EMG and EOG can also contribute to BCI control and lead to high classification accuracy if their presence or amplitude happens to be correlated with that of the mental states decoded by the BCI. In other words, EOG and EMG are typical confounding factors in BCI experiments. Thus, one may conclude that a given mental state can be decoded from EEG when it is actually decoded from EMG/EOG.

Such a risk of EMG/EOG confound can, for instance, be found when decoding emotions from EEG signals (Mühl et al. 2014b). Indeed, a given emotion often appears with a given facial expression, for example, by frowning when experiencing anger. This would lead to specific facial muscle contractions and thus to specific EMG signatures that would be picked up by EEG sensors. Thus, a BCI classifier based on these EEG signals may actually recognize facial expressions from EMG.
signals rather than actual emotions from cortical brain signals. As Mühl and colleagues pointed out, if one wants to ensure that the classification is truly based on cortical signals, it is necessary to control for EMG or EOG activity when classifying emotions from EEG recordings (Mühl et al. 2014b).

This is only one example, and ideally all BCIs, relying solely on brain activity, should be controlled for EOG/EMG confounds. While there is no perfect solution, minimizing such confounds can be done by

- Designing protocols limiting eye and motor/facial movements
- Using manual or automatic EOG/EMG cleaning techniques (Islam et al. 2016) (note that existing techniques are still improvable)
- Studying and reporting the spatial and spectral EEG topographies for each condition, to ensure they differ from that of EMG/EOG (Goncharova et al. 2003)
- Classifying directly EOG/EMG to see if they contain class-related information (if so, there is a risk of confound)
- Performing classification on high-frequency EEG signals, likely to contain EMG, to estimate EMG impact (see, e.g., Mühl et al. 2014a)
- Reanalyzing BCI data, offline, after removal of artifactual signals

And for our experiment? Here, we will definitely measure EOG and EMG of the face during the experiment as we use emotional tasks that are most likely related to facial expressions. Then, we will apply EOG/EMG cleaning techniques online. Afterward, we will classify EOG/EMG (offline) to ensure that they do not contain class-related information and also classify artifact-corrected EEG data (offline) to verify that BCI performance is above chance (and as high as possible).

### 32.2.4 What Kinds of Mental States Can Be Used?

The question of what kinds of mental states can be used to operate a BCI goes back to the question of what kinds of mental states can be detected with the neuroimaging technique at hand, which, in turn, is simply a matter of the signal-to-noise-ratio. In other words, if the neural signature of a mental state is larger than the background noise, the mental state can be detected and used to operate a BCI. Although some mental states can be detected on a single-trial level, generally the signal-to-noise ratio, and thus the reliability of the BCI output, can be increased with more repetitions. And although more repetitions typically increase the single-trial classification accuracy, repetitions may come at the expense of a lower information transfer rate. Consequently, while theoretically every mental state that can be detected with the neuroimaging technique used can be employed to operate a BCI, the actual range is narrowed down by practical considerations such as the reliability of the neural signature, the speed and accuracy of the BCI system, as well as the BCI application.

The variety of types of mental states that can be used within BCIs has led to classify BCIs from active over reactive up to passive (Mühl et al. 2009; Zander and Kothe 2011). As raised in Section 32.1, active BCIs require direct and conscious modulation of brain activity, whereby external stimulations serve at most as cues. Motor imagery, the mental imagination of movements, is a prominent active BCI paradigm (Pfurtscheller et al. 1997). Contrariwise, reactive BCIs rely on the indirect modulation of brain activity as a reaction to an external stimulation. Well-known examples for reactive BCIs are the P300 speller (De Vos et al. 2014; Farwell and Donchin 1988) and BCIs that are based on steady-state visual/somatosensory evoked potentials (Lalor et al. 2005; Müller-Putz et al. 2005, 2006). Finally, passive BCIs use brain activity arising without the users’ conscious modulation or without external stimulation, such as in the detection of error potentials (Zander and Kothe 2011). Additionally, different kinds of BCIs can be combined together, to make what is called a hybrid BCI (see Pfurtscheller et al. 2010 and Chapter 27 [“Hybrid Brain–Computer Interfaces and Their Applications”]). Given a BCI application, it is advisable to use the mental state that optimally balances accuracy and speed for the target application.
And for our experiment? Our experiment aims at investigating whether or not remembering positive/negative emotional souvenirs is reliable enough to control a BCI: thus, we are focusing on active BCIs. In other words, we want to investigate the feasibility of discriminating both these tasks on a single-trial basis. Importantly, the length of the trial will be limited: the system has to be able to discriminate the tasks in a few seconds. In other words, the information transfer rate must be at least as good as with standard motor imagery tasks for these new tasks to be useful for BCI control.

32.3 DATA PROCESSING

32.3.1 Which Classifier Can We Use Depending on the Distribution of the Data?

The choice of a machine learning algorithm significantly affects the BCI decoding accuracy. To obtain optimal performance, the algorithm capabilities and the data properties have to match. Statistical classifiers are mainly used in BCI (Lotte et al. 2007). In order to discriminate EEG signals into different classes (commands), such classifiers rely on the EEG feature data distribution, that is, on their probability density function. There are different ways to estimate such density and to infer classifiers from them. Most importantly, the classifier type should be selected based on this data distribution. In general, linear methods such as linear discriminant analysis (LDA) or support vector machines (SVMs) are used for classification of EEG signals. Such methods use linear hyperplanes to subdivide the feature space into regions belonging to the different classes/commands. The position and orientation of the hyperplanes are typically computed using the mean and covariance of the EEG features. During subsequent online BCI use, new and unseen data, that is, independent data, are then assigned to the label of their area, as defined by the hyperplanes. Linear methods are successful when the feature density follows a normal distribution. If it does not, the features can be transformed to be so. For example, applying a logarithm transformation converts band power estimates extracted from EEG (the power is the square of the amplitude and hence not a linear measure) into a normal-like distribution. A linear hyperplane between two normal distributions can therefore potentially lead to reasonable performance. An advantage of linear methods is that they do not need a lot of training data compared to nonlinear methods and thus a shorter calibration time when used online. Since the lack of sufficient training data is a common issue in BCI (Lotte 2015), linear methods are leading. As far as LDA is concerned, when little data are available, regularized LDA (notably shrinkage LDA) should be used preferably to LDA, as they were shown to lead to higher accuracy (Blankertz et al. 2011; Lotte 2015). Nonlinear methods can separate data using curves instead of “just” (hyper)planes and thus can better capture the shape of the feature density, which may lead to good generalization. Generalization is the ability of a classifier to achieve good results also with independent data. Generalizing is different from memorizing the data (a.k.a. overfitting): more details on this point are provided in Section 32.3.2. Recent nonlinear methods that are somewhat robust against overfitting are Random Forest (Steyrl et al. 2016) and neural networks such as Restricted Boltzmann Machines (Kobler and Scherer 2016). These methods do not make any assumption about the data distribution and, as such, are likely to be increasingly used for BCI in the future and to give potentially better performances. Proving superior performances nonetheless require a proper evaluation of the algorithms and notably the use of separate training and testing data sets.

As for classifiers, the other algorithms used in the data processing pipeline, for example, preprocessing or feature extraction, should also match the data properties. For instance, the common spatial pattern (CSP) spatial filter (Blankertz 2008) can be used to classify oscillatory activity data (EEG rhythm band power) but is suboptimal to classify event-related potentials such as the P300 (Lotte 2014). For the latter, dedicated spatial filters such as xDAWN should be used (Rivet et al. 2009). For a complete review of the signal processing and machine learning methods used in BCI, please refer to Blankertz (2018) and Chevallier (2018).

And for our experiment? We do not want the calibration of the classifier to be too long. Thus, we will choose a linear classifier, as it requires fewer training data. Moreover, to obtain features
respecting the properties of a standard LDA or SVM classifier, we will operate the following transformation of the features in order to obtain a normal-like distribution: (1) compute the power of the signal and (2) take the log of this power.

32.3.2 Why Is It Important to Separate the Training Data Set from the Testing Data Set?

When evaluating BCIs offline, the classification system should be calibrated using only the data from a training data set, and the resulting calibrated system should then be evaluated on completely different and independent data, which represent the testing data set (Lemm et al. 2011). Such evaluation ensures that the classifier can indeed generalize to unseen data and has not just memorized the training data set class labels. Failure to use distinct and independent training and testing data sets may lead to much higher classification performance than the “real” performance that would be obtained on unseen data (Olivetti et al. 2010) or even to better than chance performance on random data with no class information (Dominguez 2009).

Therefore, not a single parameter of the machine learning algorithm should have been calibrated with the knowledge of the testing data set. This means the choice of channels, features, hyperparameters, and normalization should be done using only the training data. In particular, when using cross-validation (CV),* all these should not be selected on all the data before applying CV to assess the classifier. Rather, they should be selected separately for each fold of the CV. Misuse of this procedure could result in erroneously concluding that some mental states can be discriminated from brain signals, when they cannot. Such a bias was revealed by Luu and Chau (2008), who claimed that it was possible to predict from functional near-infrared spectroscopy (fNIRS) which object among two a user was going to choose before they could see them. The authors claimed that they could discriminate the preferred object from the nonpreferred one, based on fNIRS signals preceding their presentation, with 80% of average CV classification accuracy (Luu and Chau 2008). Unfortunately, feature selection was performed not for each fold of the CV, as it should have, but only once on all the data. Dominguez (2009) revealed that by applying the exact same incorrect CV evaluation procedure on completely randomly generated noise—with no class information—he could obtain the same classification accuracy. Subsequently, when correcting their approach, and performing feature selection only on each training CV fold, the authors of the original paper obtained a classification accuracy as low as 56% (vs. 80% before), that is, essentially chance level, thus disproving their initial claim (Chau and Damouras 2009). This stresses once more the need for an independent testing data set, that is, a data set that no part of the machine learning algorithm has ever used.

And for our experiment? In order to assess whether remembering a positive emotional souvenir can be distinguished in EEG signals from remembering a negative one, we will use machine learning algorithms such as feature selection, spatial filter optimization, and LDA/SVM. To assess the achievable classification accuracy, we will use K-fold CV (a typical value of K would be 10). Therefore, we will run the training algorithm, the feature selection algorithm, the spatial filter optimization algorithm, and the LDA/SVM on K − 1 folds and evaluate the obtained parameters on the remaining fold (i.e., the testing data set). This training/testing procedure should thus be done K times, for the classifier training, feature selection, and filter optimization.

32.3.3 How to Determine the Chance Level for the Classified Data

In the example introduced just above, we state that 56% of classification accuracy was basically chance level. But how do we know from which accuracy onward we can consider a result to be above

* CV is a procedure that divides a data set randomly into K folds of equal size, and then spatial filter optimization, feature selection, and classifier training are performed on all but one part, and the obtained predictions are compared to the real labels of the remaining part.
chance? It is indeed an important question to determine whether a classification result, the decoding accuracy, deviates from chance level or not. Chance level refers to the rate achieved by random classification. For a two-class problem, the theoretical chance level is 50%; for a four-class problem, it is 25%, and so on. However, achieving a classification result of 70% in a two-class scenario may or may not indicate a valid above chance classification accuracy. It is important to recognize that the theoretical chance level is valid only for a large number of samples (or trials). Imagine flipping a coin once; the result could be either head or tails. Now assume you flip the coin four times in a row. By chance alone, it may be that the outcome is four times head. At least, and even though the coin would not be biased, it is not at all certain that the result will be two times head and two times tail. Because of the small number of trials, a strong deviation of the observed rate from the theoretical chance level can occur. Only with a sufficient number of trials does the outcome of heads become more and more unlikely—and for a sufficient number of observations, the frequency of heads and tails will approach the theoretical chance level of 50%. Thus, if the theoretical chance level can be exceeded by chance alone, how could it be determined whether a particular classification result is significantly above chance? Analytical (binomial statistic) and empirical approaches (permutation tests) are available to answer this question, as summarized in Combrisson and Jerbi (2015) and Müller-Putz et al. (2008). By assuming that classification errors follow a cumulative binomial distribution, one can apply the inverse binomial cumulative distribution to figure out whether a particular classification result is above a particular significance threshold. The critical number of correctly classified trials that could arise by chance alone is determined by

\[ \text{Crit}_{\text{trials}} = \text{binoinv} \left( 1 - p, n, 1/c \right) * 100/n. \]

Here, \( p \) refers to the significance threshold (e.g., \( p = 0.05 \)), \( n \) indicates the number of trials, and \( c \) denotes the number of classes (assuming equal class occurrence). The binoinv function is available for instance in MATLAB® (Mathworks Inc., Natick, Massachusetts). According to this equation, for a two-class problem, \( p = 0.05 \) and \( n = 20 \) trials, the critical number of correctly classified trials is 70%. Hence, only a classification accuracy exceeding 70% can be interpreted as a result significant above chance level. A corresponding look-up table is provided by Müller-Putz et al. (2008) and Combrisson and Jerbi (2015) for different \( p \) values; two-, four-, and eight-class scenarios; and different trial counts. As discussed by Combrisson and Jerbi (2015), the analytical approach has some theoretical limitations, whereas the empirical approach comes at the expense of high computational costs. However, for random noise data, both suggest similar thresholds.

The procedure just described refers to random data, typically resulting in balanced class labels for a sufficient number of trials. However, in unbalanced two-class situations where one class occurs in the majority of all trials (e.g., 90% nontarget class trials, 10% target class trials in paradigms of reactive BCIs such as the P300 speller, when measuring single-trial P300 detection performances), the calculation of accuracies across all trials can be highly misleading. If one always goes with the majority vote, the resulting accuracy may falsely indicate very high recognition rate: Imagine a classifier always voting 1, then the accuracy would be incorrectly calculated as 90%! In those cases, the confusion or error matrix should be reported, which makes it easy to evaluate which classes are confused. One simple solution is the calculation of a corrected accuracy given as the mean across all recognition rates for all classes. In the above example, for class 1, the recognition rate would be 100%; for class 2, the recognition rate would be 0%, and the average of both is then 50%—the corrected classification accuracy. Another frequently used performance metric in that case is the area under the receiver operating characteristic curve (Bradley 1997). The interested reader can refer to the tutorial in Thompson et al. (2014) for more information on classification performance metrics. Alternatively or additionally, the end application performances can also be reported, for example, the percentage of correctly selected characters in the P300 speller, which is a balanced problem.

And for our experiment? Once our CV process performed, we will have to answer the following question: Can remembering a positive emotional souvenir be distinguished from remembering a
negative one, in EEG? The answer depends on whether the obtained classification accuracy is above or below chance level. If it is above, it means that these two tasks can actually be distinguished; otherwise, it means that the classifier did not manage to separate them. To know the chance level, it is enough to have a look at the tables introduced here-above and look at the chance level for two classes (because here we have positive vs. negative emotional souvenirs). The chance level will also depend on the number of trials per class. This should thus encourage collecting as many trials as possible (the higher the number of trials, the lower the chance level).

32.3.4 To What Extent Can Commercial Algorithms Be Trusted?

Consumer-grade EEG and BCI systems, such as the popular Neurosky (www.neurosky.com), Emotiv (www.emotiv.com), or many other devices, are increasingly used. Many commercial BCI systems come with ready-to-use algorithms to detect mental states such as attention, emotions, or meditation. Such algorithms are often used as they are, as a ground truth value, notably in human–computer interaction research. This could sometimes be an issue when designing rigorous BCI experiments for a number of reasons.

First, many (but fortunately not all) of these algorithms claim to be able to measure such mental states but have never been scientifically validated. It does not mean that they do not work, but rather that one does not know if they work. As such, if they were not independently and rigorously validated in a scientific journal publication (whose rigor standards are usually higher than that of conferences), such algorithms cannot be used as reliable measures. For instance, an independent evaluation of the Emotiv emotion recognition algorithms revealed that “the data is unreliable and incoherent” (Jorgensen et al. 2017).

Second, such algorithms are most often black boxes, meaning one does not know how they work and which features they use. This makes a study using them potentially unreproducible with other EEG devices. Moreover, it also prevents us from assessing whether these algorithms are really BCI, based only on signals from cortical origin, or whether they are based on confounding factors such as EOG/EMG (see also Section 32.2.3). For instance, Neurosky and Emotiv algorithms are believed to or even admit they use EMG/EOG and thus are not real BCI (Singer 2008).

Even for algorithms that are scientifically validated and purely based on signals from cortical origin, it should be noted that EEG signals are changing heavily because of their context of use (Brandl et al. 2016; Mühl 2014a). As such, they should be validated again in their target context, if this context is different from the one in which the validation was performed. Finally, even for such algorithms, like any other BCI system, they are not perfect and make frequent mistakes when estimating users’ mental states. Therefore, they should be treated as such, and not as perfect mental state decoders.

Overall, using commercial algorithms is thus not a problem per se and can even be very useful and convenient, but it should be done with care. In particular, they should be used in rigorous scientific BCI experiments only if they were scientifically validated, including in the target context of use, and if the algorithm is known (i.e., not a black box).

And for our experiment? Since some commercial algorithms claim to be able to recognize various emotion-related mental states, or even any user-defined mental states in EEG, we might be tempted to compare the performances they obtain for positive versus negative emotional souvenir, to the one we obtain with our own data and algorithms. However, we should refrain from doing so, at least with algorithms being black box, because we do not know whether they rely on pure EEG signals, or whether they use EMG/EOG as well. In the latter case, any comparison would be unfair and meaningless since we use only EEG.

32.3.5 How to Determine the Relevance of Neurophysiological Markers

One crucial issue when designing BCIs is the choice of features used to encode messages. Features that do not contain relevant information add noise to the system. If the machine learning algorithm
is not robust, then adding noise or redundant information may decrease the BCI decoding accuracy. One way to prevent this from happening is feature reduction or selection. This means that only features that lead to high decoding performance are given to the machine learning algorithm. Note that to ensure that such features do not lead to overfitting, that is, that they can generalize to new unseen data, it is necessary to assess them on a different data set than the one on which they were selected, as indicated in Section 32.3.2. The most crucial point, however, is that the selected features are neurophysiologically meaningful. Machine learning methods generally cannot judge whether the used features are neurophysiologically meaningful. For instance, artifacts not originating from the brain (e.g., EMG) may be strongly correlated with the BCI task and would be easier to detect given their higher amplitudes. An algorithm exploiting them would, however, not be considered as a pure BCI, as discussed in Section 32.2.3. A relevant concept in machine learning is “Garbage in, Garbage out,” which means that if the used features are meaningless, even the best machine learning algorithm will not be able to find patterns that can be discriminated.

One way to check whether there are significant differences between conditions in the spontaneous EEG is to compute time/frequency event-related desynchronization (ERD—relative amplitude decrease in a specific frequency band over defined brain areas) and event-related synchronization (ERS—relative amplitude increase) maps (Pfurtscheller and Da Silva 1999). These maps show statistically significant changes as a function of time. In other words, they show which oscillatory components undergo significant amplitude changes. If identified components are in agreement with patterns reported in the literature, then the process was successful. For instance, for a motor imagery experiment, in agreement with the literature would mean ERD in the alpha and beta range over sensorimotor areas (Pfurtscheller et al. 1997), while for a steady-state visual evoked potential experiment, it would mean ERS at the stimulation frequencies over occipital areas (Vialatte et al. 2010). As mentioned before, if the experiment consists in exploring a new mental task for BCI, then such analyses are even more necessary to ensure that the machine learning algorithm is not in fact using artifacts (see also Section 32.2.3).

Finally, it should be stressed that the weights obtained by training the machine learning algorithms cannot necessarily be used directly to identify the involved brain areas. Indeed, most classifier weights are actually “filters” and can thus give high weights to sources of noise in order to cancel them (Haufe et al. 2014). Rather, the weights should be transformed into “patterns” before interpretation (see notably Haufe et al. 2014 for the linear case).

For our experiment? In addition to the offline classification of EOG/EMG introduced in Section 32.3.1, we will perform a time/frequency ERD/ERS analysis in order to determine the neurophysiological features involved. We hope that some neurophysiological features related to emotions (positive vs. negative valence) based on the literature, such as the frontal asymmetry (Dolcos et al. 2004; Schmidt and Trainor 2001), will be involved whereas frontal and occipital high frequencies (gamma) will not be involved, as they are most likely related to EMG (Goncharova et al. 2003).

### 32.3.6 Why Is It Important to Correct for Multiple Comparisons?

The aim of most experiments is to test and compare alternative hypotheses. In BCI, typical hypotheses may arise from questions such as the following: When is the sensorimotor desynchronization significantly different from baseline? Where on scalp can the beta rebound be best captured after motor imagery? In which frequency band can we observe a significant difference between two mental states? Or what signal features best discriminate between the desired alternative commands?

Although quite different, those questions share a common goal (feature identification) and a common procedure (statistical hypothesis testing). Importantly, when the data space is large, such as with EEG data that unfold in space, time, and frequency, feature identification involves multiple hypothesis testing. As the size of the space of possible features increases, the number of tests or comparisons to be performed increases. And as the number of tests increases, so does the risk of
concluding to a significant effect in at least one of those comparisons, by mistake. The latter means that an effect could be suggested to be significant whereas this effect does not truly exist.

Controlling for that risk (called the risk alpha or type I error) is very important since wrongly identifying an effect would typically yield a choice of BCI features that would not work in practice. The risk alpha at the level of multiple tests relates to the risk alpha at the level of a single comparison. In classical statistics, there is a wide agreement to keep that risk below 5%. This number has been chosen somewhat arbitrarily and could be chosen otherwise. What is important is to define that limit before testing, so as not to bias this choice and the ensuing conclusion; 5% means that the probability $p$ of mistakenly rejecting the null hypothesis is equal to or less than 0.05.

A classic example of multiple comparisons is the two-sided $t$ test, when comparing the means of two populations, say $\mu_1$ and $\mu_2$. The null hypothesis states that $\mu_1 = \mu_2$. The test to reject this null is two-sided when one tests for both alternatives: $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$. In that case, the risk alpha for each comparison is typically set to 2.5% so that the risk alpha of mistakenly rejecting the null is 5%. Hence, the family risk is equal to the number of comparisons ($n = 2$) times the risk alpha at the single test level (2.5%).

The same rationale applies to $n$ multiple comparisons with $n > 2$ so that the risk for a single test is set to 0.05/$n$ and guarantees that the family risk remains 5%. This is known as the Bonferroni correction. However, this correction is often too conservative, since it relies on the a priori assumption that the multiple comparisons are mutually independent, which is rarely true in practice, at least when dealing with neurophysiological or neuroimaging data. For instance, scalp EEG data are known to be spatially blurred so that nearby sensors will likely display highly similar activities. These spatial correlations should be accounted for in order to increase the sensitivity of the statistical analysis. Similarly, EEG signals exhibit strong temporal correlations and most reliable significant effects typically expand over several tens of milliseconds. Methods have been developed so as to optimize the corrections for multiple comparisons in the particular context of brain functional data. The most popular ones are available in main academic software packages (see, for instance, Litvak et al. 2011 and Oostenveld et al. 2011 and other articles in that same special issue). These are the corrections based on random field theory (see Worsley 2006) to control for the familywise error rate or on approaches to control for the false discovery rate (see Nichols 2006). Note that the former is typically made for statistical inference on parametric images. However, nonparametric or permutation approaches are also very much used for the analysis of electrophysiological data (Maris and Oostenveld 2007).

To sum up, as the number of comparisons increases, the risk of mistakenly rejecting the null hypothesis in one of these comparisons increases. A correction is needed to reliably identify a useful feature for BCI. Be it at the individual or group level, feature identification will often require multiple testing. This reminds us very importantly that whenever prior knowledge is available to reduce the search space to the most likely relevant features, the correction for multiple comparisons will be less drastic and the ensuing identification will be more sensitive.

And for our experiment? Here, to complete the time/frequency analysis, we would like to know if, in accordance to the literature, the frontal asymmetry varies depending on the valence of the emotional memory and if this asymmetry enables us to classify reliably enough our data. Thus, we will divide the frequency range and subbands, for instance, delta (1–4 Hz), theta (4–8 Hz), low alpha (8–10 Hz), high alpha (10–12 Hz), low beta (12–24 Hz), and high beta (24–30 Hz). Then, we will compare the frontal asymmetry for each of these frequency bands. Because we perform several comparisons, we will apply a correction, for instance, the false discovery rate, in order to adjust the significance threshold.

### 32.4 EXPERIMENTAL DESIGN AND THE USER COMPONENT

#### 32.4.1 What to Have in Mind When Designing a New BCI Experiment

Along the long road to develop and validate a new application, BCI experiments may have different purposes. At an early stage, offline experiments may be required to explore the neural correlates
of some targeted mental processes and their potential usefulness. For instance, significant ERDs or ERSs in specific frequency bands will be investigated in a population of subjects, between two conditions that we wish to distinguish (e.g., low vs. high mental workload).

Then, online experiments are typically designed to evaluate or compare the performance of a BCI or of one of its component (a neurophysiological marker, an algorithm, a feedback, etc.). In that case, typically, the same subjects will be tested under two or more conditions in order to answer questions such as the following: Is it useful to include the early N170 visual component in the classification to improve P300-based spelling? Which classifier provides the best performance? Does it make a different if we move from a 2D to a 3D visual feedback? Note that some of these questions can be partially answered with offline experiments in which many tests can be performed a posteriori, based on the same data. However, a full evaluation will require an online study where the different conditions to be compared will have to be evaluated in separate trials.

BCI experiments may also aim at assessing a learning curve over several sessions or at comparing groups of users. Although not mutually exclusive, these questions are very different from each other and point toward different design parameters. In particular, the former will require defining the number of sessions for each subject, while the latter will require defining the number of subjects in each group.

Finally, at a later stage of development, validation of a BCI may take the form of a randomized controlled trial (e.g., for neurofeedback training applications), which will have to be carefully designed to efficiently demonstrate and help quantify the desired effect. An important question, in particular, will be to control for putative confounders and ensure that the observed effect is indeed produced by the intended manipulation (e.g., in BCI, that the control is based on brain and not muscular activity, or in neurofeedback, that the effect is specific to the modulation of the targeted neural activity).

Hence BCI experiments cover pretty much the whole spectrum of possible designs that one may encounter in empirical science. The crucial question of how to optimally design a BCI experiment can thus rely on principles derived from applied statistical works in the fields of experimental psychology, cognitive neurosciences, and neuroimaging (e.g., Daunizeau et al. 2011; Henson 2006).

Put simply, designing an experiment first requires one to clearly state the alternative hypothesis to be tested. If properly done, this greatly constrains the experimental conditions one should consider and naturally points toward confounds that should be carefully controlled. In other words, this early and seemingly simple first step is essential to enable finding natural and proper answers to most important design questions that come next about the control group (Section 32.4.2), the control condition (Section 32.4.3) and the appropriate statistical tests (Section 32.4.4).

An often overlooked aspect in BCI though, like in many other fields, relates to the sampling issue. How many subjects should I test? And how many trials per subject should I record? Answering those questions is crucial for guaranteeing the reproducibility of BCI results.

Interestingly, the theoretical field of design optimization is still very active and BCI already motivated methodological innovations in that field, be it for maximizing design efficiency with respect to hypothesis testing or parameter estimation, by optimizing the stimuli or the number of samples (trials and subjects) (see, for instance, Sanchez et al. 2016 and Melinscak and Montesano 2016).

And for our experiment? Here, our objective is to “explore the neural correlates of some targeted mental processes and their potential usefulness,” the mental process being remembering positive/negative emotional souvenirs. Given the high between-subject and intrasubject variability in terms of BCI performance, in order to obtain a statistically significant response to our hypothesis, we need an important number of participants (at least 20 would be good) and many trials (a typical BCI session would be 4 runs of 20 trials per class, i.e., 80 trials per class in total).

### 32.4.2 Why and How to Have a Good Control Group?

Broadly speaking, one can distinguish between feasibility studies and controlled studies. As the name indicates, feasibility studies, also referred to as proof-of-concept studies, analyze the viability
of an idea. Often, feasibility studies are designed to pave the way for future controlled studies. To take but one example, Gharabaghi et al. (2014) demonstrated that closing the loop between mental states, cortical stimulation, and haptic feedback is feasible. Building on this, larger (clinical) controlled studies are necessary to evaluate the utility of this approach for the rehabilitation of lost motor function after stroke. It is not unusual that, at the beginning of a scientific achievement, a higher rate of feasibility studies is performed, which, at success, are followed by controlled studies. Contrary to most feasibility studies, controlled studies comprise an experimental group and a control group, whereby the nature of the control group largely depends on the research question. For instance, if one wants to assess the effect of age on the accuracy of a motor imagery–based BCI, the experimental group could comprise older individuals and the control group could comprise younger individuals. If one is, however, interested in the consequences of stroke on the ability to steer a motor imagery–based BCI, the experimental group could consist of stroke patients and the control group could consist of healthy individuals. If, to name another example, one wants to examine the rehabilitative effects of motor imagery–based BCI training after stroke, both experimental and control group should consist of stroke patients, whereby the experimental group receives real feedback and the control group receives sham feedback during the BCI training (for details on the aspects of sham feedback, see Section 32.4.3). In all cases, the experimental group and the control group differ ideally only with regard to the independent variable. This can be achieved by matching individuals on variables of putative importance but of noninterest, such as gender, age, or education, or, ideally, if possible, by employing a within-subject design. Taken together, well-controlled studies have the potential to isolate the effect of the independent variable of interest on dependent variable(s).

And for our experiment? We are proposing a feasibility study: the goal being to investigate whether or not it is possible to control a BCI using the remembrance of positive/negative emotional souvenirs. This step being preliminary, a control group is not mandatory. Nonetheless, we also aim to assess the relevance of these tasks in comparison to more standard motor imagery tasks such as imagining left- and right-hand movements. This is why we will propose a within-subject design with two conditions: in some runs, participants will perform the new tasks (remembering positive vs. negative emotional souvenirs), while in other intermixed runs, they will perform standard left- and right-hand motor imagery tasks. To keep a high number of trials per class (as required), participants will take part in two sessions of four runs.

32.4.3 HOW TO AVOID BIASES: CONCEPTS OF COUNTERBALANCING, SHAM CONTROL, DOUBLE BLINDNESS, AND RANDOMIZATION

In order to avoid biases, a couple of concepts can be useful. One of these is to compare the effects of real feedback with the effects of sham feedback. The so-called sham-controlled designs enable one to better evaluate the effectiveness and specificity of the feedback. In other words, the inclusion of a sham-control group is crucial to control for nonspecific factors such as motivation, expectancy, and practice effects. Based on these principles, there seems to be a general agreement that the inclusion of a sham-control group is of advantage. However, at present, there are no common criteria for the optimal sham-control condition. The existing sham-control conditions can be mainly assigned into five groups: (1) no feedback (Kadosh et al. 2016; Zich et al. 2015); (2) feedback based on activity stemming from a different brain region (Harnelech et al. 2015; Lee et al. 2012; Paret et al. 2016; Yao et al. 2016; Zotev et al. 2016); (3) feedback based on the activity from a different point in time, for example, different trial or session (Braun et al. 2016; Okazaki et al. 2015); (4) feedback based on activity from a different user (Chiew et al. 2012; Engelbrecht et al. 2016; Escolano et al. 2014; Kober et al. 2014; Ros et al. 2013; Witte et al. 2013); and (5) feedback based on artificially created irrelevant randomized signals (Arnold et al. 2012; Mihara et al. 2012, 2013). Furthermore, Gevensleben et al. (2014) designed for their learning study a particular exceptional sham-control condition. In brief, feedback was based on data from a previous study, providing a variety of different feedback curves, which were additionally weighted by coefficients to control the development
Mind the Traps! Design Guidelines for Rigorous BCI Experiments

over time (Gevensleben et al. 2014). To the best of our knowledge, there is no ideal sham-control condition* at the present time, which is why it is even more important to indicate in detail what kind of sham-control condition was used. Furthermore, the ethical concerns of sham-control conditions should be considered; this particularly applies for clinical research where standard treatment exists (La Vaque and Rossiter 2001; Vernon et al. 2004).

For sham-control designs, but also other group comparisons (e.g., two different mental strategies), the question arises whether to employ a within-subject design or a between-subject design. Each has its advantages and disadvantages; for instance, while a between-subject design avoids order and carryover effects, interindividual variation introduces nonspecific differences between the experimental groups. In each case, it is advisable to (pseudo-)randomize and counterbalance the conditions, ideally in a double-blind manner. While no randomization and counterbalancing can introduce order effects, no blinding can compromise the objectivity of the evaluation.

And for our experiment? Here, we are performing offline analyses. In other words, we do not classify online the data and therefore do not propose a closed-loop BCI. Thus, participants will not be provided with a feedback. Furthermore, as we propose a within-subject design with the comparison of two pairs of mental imagery tasks, we have to choose to either randomize or counterbalance the conditions in order to avoid order effects. Given the low number of participants, it will be more relevant to counterbalance the conditions. The experiment will be conducted in a single-blind manner: the experimenter is blind to the order of the two mental imagery tasks.

32.4.4 How to Select the Appropriate Statistical Tests

The type of statistical test to use depends on the research question, that is, more precisely on the hypothesis. Typically, a dependency analysis is performed: we want to find either differences (univariate analyses) or correlations (bivariate analyses). There are two cases if one looks for differences. The samples are either independent or related (paired). For instance, if we have a variable with two modalities, A and B, (1) in a between-subject design, group 1 will use modality A and group 2 will use modality B: the samples will be independent; (2) in a within-subject design, all the participants will use both modalities: the samples will be related/paired. Then, it is also important to pick the method based on the distribution of the data: if the data have a normal distribution, it will be possible to use parametric tests; otherwise, nonparametric tests should be used (although it is worth noting that parametric tests are fairly robust to deviations from the normal distribution). Concerning bivariate analyses, if the data’s distribution is normal, a Pearson correlation analysis should be performed; otherwise, a Spearman rank correlation should be used. Concerning univariate analyses, for a better readability, we propose a table (Table 32.1) to find the appropriate statistical test depending on (1) the number of variables, (2) the fact that the samples are independent or not, and (3) the distribution of the data.

Then, when we have three or more groups, if the analysis shows significant effects (for instance, $p < 0.05$), post-tests can be performed for which correction for multiple comparisons should be applied (see Section 32.3.6).

And for our experiment? In our experiment, we want to investigate the difference of BCI performance between the tasks “remembering positive/negative emotional souvenirs” and “imagining left/right-hand movements.” Thus, we will do a univariate analysis. Also, we have one variable “MI tasks” with two modalities: “remembering emotional souvenirs” and “performing motor imagery”; we use a within-subject design and counterbalance the conditions to avoid order effects. Thus, we will average the performance obtained at the four runs of each condition in order to obtain one measure of performance for each pair of tasks, for each participant. We will analyze the distribution of the data, but given the small sample (20 participants), we will most likely have to use a

* Indeed, the abovementioned control conditions do not control exactly for the same aspects/effects; therefore, it would be hard if not impossible to imagine a sham condition that would control for all of them.
nonparametric test. Thus, we will do a univariate nonparametric test for two paired samples: a Wilcoxon signed-rank test.

### 32.5 SUMMARY AND CONCLUSION

To conclude, we propose, in Table 32.2, a summary of the key points to be considered to design a rigorous BCI experiment. Once again, we do not claim to be exhaustive, but hope that we tackled the key points that will enable the reader to understand how to avoid common pitfalls when designing a BCI experiment. Nonetheless, it has to be noted that designing a rigorous experiment is often not enough to guarantee the significance and relevance of the latest. It is also of the utmost importance to question the scientific relevance of the study, to carefully acknowledge the impact of the user training and feedback (Kleih and Kübler 2018; Mladenovic et al. 2018), as well as to consider the impact of social and relational aspects, that is, of the way the study is performed and the relationship between the experimenter and the participant/patient.

### TABLE 32.1

**Type of Statistical Test to Be Performed as a Function of the Study Design (Type and Number of Samples) and of the Distribution of the Data**

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Independent/Unpaired Samples</th>
<th>Paired Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parametric Tests</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Samples</td>
<td>Variances assumed equal: Student t test</td>
<td>Paired t test</td>
</tr>
<tr>
<td></td>
<td>Otherwise: Welch t test</td>
<td></td>
</tr>
<tr>
<td>3+ Samples</td>
<td>Variances assumed equal: n-way ANOVA</td>
<td>ANOVA for repeated measures</td>
</tr>
<tr>
<td></td>
<td>Otherwise: Kruskal–Wallis test</td>
<td></td>
</tr>
<tr>
<td><strong>Nonparametric Tests</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Samples</td>
<td>Mann–Whitney U test</td>
<td>Wilcoxon signed-rank test</td>
</tr>
<tr>
<td>3+ Samples</td>
<td>Kruskal–Wallis test</td>
<td>Friedman test</td>
</tr>
</tbody>
</table>

(Continued)

### TABLE 32.2

**Key Points to Design Rigorous BCI Experiments**

| Point | |
|-------|-----------------
| How to choose the appropriate sensor type, location, and number depending on what we want to measure | Acquisition of the Signals | Currently, a trade-off has to be done between the comfort/ease of use and the quality of the signal. Ag/AgCl wet electrodes offer the highest quality of signal. Using the 10–20 system with at least 32 electrodes enables a large coverage of the scalp. Nonetheless, for economical/practical reasons, lighter setups can be used. Given that an unknown number of brain (and non-brain) generators contribute to the scalp EEG, one cannot determine the brain source of any particular EEG signal for sure only based on the activity measure at one electrode location. EEG source localization procedures can be used to confirm ideas about brain sources contributing to the EEG. |
| What can we infer from the activation of one electrode? Concept of spatial resolution | | Ocular (EOG) and muscular (EMG) activity are also measured by EEG sensors. They are typical confounding factors in BCI experiments, when correlated with the EEG patterns used by the BCI. They should thus be controlled for in any study. |
| How muscular activity can interfere with EEG activity and why it should be controlled for to avoid confounds | | |

(Continued)
TABLE 32.2 (CONTINUED)

Key Points to Design Rigorous BCI Experiments

| What kinds of mental states can be used? | Theoretically, any mental state can be used while its neural signature is larger than the background noise. Then, the mental state should be selected depending on the information transfer rate and on the classification accuracy required for the target application. |
| Which classifier can we use depending on the distribution of the data? | When a few data are available for the calibration, linear classifiers such as LDA or SVM should be used (to avoid overfitting). It should be reminded that to be allowed to use these classifiers, a transformation of the features must be performed to obtain a normal-like distribution. |
| Why is it important to separate the training data set from the testing data set? | During offline analyses, cross-validation procedures should use testing data sets that are independent from the training data set to ensure that the algorithm can indeed generalize to unseen data and has not just memorized the training data set class labels. |
| How to determine the chance level for the classified data | The chance level depends on the number of classes and of the number of trials per class. To know the chance level, it is enough to have a look at the tables presented in the papers cited in this section. |
| To which extent can commercial algorithms be trusted? | Using commercial algorithms can be very useful and convenient, but it should be done with care. In particular, these algorithms should be used in rigorous scientific BCI experiments only if they were scientifically validated, including in the target context of use. |
| How to determine the relevance of neurophysiological markers | One way to determine the relevance of neurophysiological markers is to do time/frequency analyses of ERD and ERS and to compare the highlighted features to the literature. |
| Why is it important to correct for multiple comparisons? | As the number of tests increases, so does the risk of concluding to a significant effect in at least one comparison by mistake. Controlling for that risk (called the risk alpha or type I error) is very important since wrongly identifying an effect would typically yield a choice of BCI features that would not work in practice. |

Experimental Design and the User Component

| What to have in mind when designing a new BCI experiment | Designing an experiment first requires to clearly state the alternative hypotheses to be tested. If properly done, this greatly constrains the experimental conditions one should consider and naturally points toward confounds that should be carefully controlled. |
| Why and how to have a good control group? | Control groups are required to prove the efficiency of a new paradigm/approach. The nature of this control group should be defined depending on the hypotheses and on the goal of the experiment. |
| How to avoid biases: Concepts of counterbalancing, sham control, double blindness, and randomization | The so-called sham-controlled designs enable one to better evaluate the effectiveness and specificity of the feedback. Also, it is advisable to randomize or counterbalance the conditions, ideally in a double-blind manner. While no randomization and counterbalancing can introduce order effects, no blinding can compromise the objectivity of the evaluation. |
| How to select the appropriate statistical tests | The type of statistical test to use depends on the research question and more precisely on the hypothesis. Typically, a dependency analysis is performed: either we want to find differences or correlations. Then, to determine which test to perform, 3 questions should be answered: (1) Is the distribution of the data normal-like? (2) Are the samples paired? (3) How many groups are involved? |
ACKNOWLEDGMENTS

CJ and FL were supported by the French National Research Agency with the REBEL project (grant ANR-15-CE23-0013-01). JM is supported by the LabEx Cortex ("Construction, Function and Cognitive Function and Rehabilitation of the Cortex," grant ANR-11-LABX-0042) of Université de Lyon.

REFERENCES


Bradley, A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145–1159.


Mind the Traps! Design Guidelines for Rigorous BCI Experiments


