Supplementary Text and Figures for “*A General Method for Assessing Brain-Computer Interface Performance and its Limitations* ”

N. Jeremy Hill[1*], **Ann-Katrin Häuser**[1,2] **and Gerwin Schalk**[1,3,4,5,6,7]

[1] Wadsworth Center, New York State Department of Health, Albany, NY, USA
[2] Institute for Cognitive Science, University of Osnabrück, Germany
[3] Department of Neurology, Albany Medical College, Albany, NY, USA
[4] Department of Neurosurgery, Washington University in St. Louis, MO, USA
[5] Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
[6] Department of Biomedical Sciences, State University of New York at Albany, NY, USA
[7] Department of Electrical and Computer Engineering, University of Texas at El Paso, TX, USA

[*] Correspondence: Jeremy Hill, Wadsworth Center, C640 Empire State Plaza, Albany, NY 12201, USA. e-mail: `jezhill@gmail.com`

Abstract. This document is part of the supplementary material for the paper *A General Method for Assessing Brain-Computer Interface Performance and its Limitations* .

In Section S1 we present a general random-walk simulation method for estimating the success rate P_0 that we might expect by chance in a brain-computer interface control task, i.e. on the assumption that the user had no voluntary control over the system. The user’s actual success rate can then be assessed relative to P_0 to produce an estimate of information gain. Python code for performing random-walk simulations, and for computing information gain rates, is also provided on the web at `http://schalklab.org/downloads` .

In Section S2, we give a numerical example of the computation of our performance metric RIG_B, along with details of the method we use to compute standard error bars and other confidence intervals on it.

In Section S3 we present an analysis indicating that the session-to-session variation in BCI performance exhibited by our two above-chance subjects is likely the result of fluctuations in the EEG signals themselves, rather than inherent properties of the measurement method.

In Section S4, we discuss our performance assessment approach based on information gain: in Section S4.1 we contrast it with approaches based on Fitts’ Law; in Section S4.2, we explore the relation of our performance metric RIG_B to two measures of information transfer rate (ITR) that are popular in BCI studies, namely Wolpaw’s ITR and channel capacity (a.k.a. Nykopp’s ITR).

S1. Estimating chance performance under the null-hypothesis of no voluntary BCI control

To make a fair assessment of a BCI user's success rate P in performing a task, we must take into account the success rate P_0 that we might expect by chance, i.e. under the null hypothesis that the user has no control over the BCI system. In Section 2.7 of the main paper, we defined one approach by which it is possible to relate the two probabilities, but we did not yet address the question of how P_0 was to be estimated.

In some laboratory BCI experiments, trials can be designed according to very tight constraints: experimenters can ensure that trial outcomes are divided exhaustively into a finite number of non-overlapping targets; they can ensure that each target is encountered with equal probability; and they can minimize any factors that make some targets easier to hit than others. If all these criteria are satisfied, then it is easy to write down P_0 directly: if there are N discrete equiprobable outcomes, then $P_0 = 1/N$.

In the more general case, including many realistic use-cases for BCI technology, it is not so easy to specify P_0 in this way. Even in a simple game, such as the one used in the current study, it is not clear *a priori* how successful players might have been if they had simply moved the cursor back and forth at random, or provided some other kind of uncontrolled input. (Analogously, some early console computer games could be played successfully to some extent by random "button mashing".)

One way to account for this button-mashing problem is to perform random re-simulation of the task during data analysis. The goal is to quantify the performance that would be expected from an input signal whose statistical properties match those of real intentional control signals as far as possible, while lacking any actual intentional control. An example of this, already introduced in Section 2.3 of the main paper, is the Random Baseline condition: we re-ran our original online BCI system, driving it using real pre-recorded brain signals so that the statistics of the input signal would be identical to those used in the real games. Since our game was a deterministic system, we had to introduce a manipulation that took away meaningful control from the signal. In the current study, our solution was to use a time delay. Other studies have replayed the signals backwards, as in our study of a BCI-driven Breakout game [1], or have simply replayed the signals through a *non*-deterministic system, such as the physical pinball machine of Tangermann et al. [2].

The disadvantage of such online-replay methods is that they rely on running a fully-functioning implementation of the online system. An online BCI system typically must perform a large number of tasks that are not directly related to the evaluation of control signals and success rates (for example, interfacing with hardware, presenting visual and auditory stimuli, processing EEG). Therefore, the analysis cannot be performed quickly, cannot be replicated offline, and cannot be repeated an arbitrarily large number of times to increase the precision of the estimate of P_0. It is also no trivial task to engineer an online BCI system to be fully deterministic, to support a reliable replay analysis. An alternative approach is to break the game into discrete trials and re-simulate each trial repeatedly. Some studies that have adopted this approach randomize the input signal's phase spectrum on each re-simulation (see for example [3]). However, we wished to develop a general re-usable method applicable to a wide variety of control scenarios—one which could easily simulate some of the more common game mechanics, such as the fact that the cursor would stop when it hit the edges of the screen. Therefore, our solution was to re-simulate each trial repeatedly using a random-walk method. We describe the approach in the rest of this section. Although the current study only used one-dimensional control, we describe the general multi-dimensional case. As described, the approach is suitable for any control task in which targets must be hit and/or avoided, the cursor is prevented from moving through certain barriers, and the trials are defined such that targets' behavior is not dependent on the cursor's behavior within a given trial. The random-walk approach would also allow further game mechanics and physical constraints to be simulated relatively easily.

We define the **scope** of a simulation to be the set of trials over which a single estimate of P_0 must be computed—for our current study, the scope comprised all 3 measurement/adjustment phases performed by the same person in the same session in the same controller condition. Each trial is simulated S times, and the success rates for all trials within the same scope are averaged to arrive at an estimate of P_0. We used $S = 1000$ and the number of trials within one scope was between 45 and 129. Hence each of our P_0 estimates was based on 45,000–129,000 simulations.

The information and assumptions required to simulate each trial are as follows:

(i) The game is assumed to consist of a cursor (under the player's control), targets (which the cursor must either hit or avoid) and barriers (which may limit the cursor's movement).

(ii) The game is broken down into trials. Each trial is a short episode whose outcome can be either successful or unsuccessful. Each trial will be re-simulated S times in isolation from other trials.

(iii) The cursor's size W_c and position X_c at the beginning of every simulation of trial i match the size and position at the start of trial i of the actual game. The cursor size is assumed to be constant throughout the trial.

(iv) The timeline of each trial is considered as a series of discrete time steps. Like many computer games and computerized tasks, our game was already implemented as a series of discrete steps, so we used its intrinsic step duration of 31.25 ms (8 digitizer samples).

(v) For each trial, barriers are defined through which the cursor cannot pass. Whenever a step would cause the cursor to overlap with a barrier, the step is truncated such that the cursor just touches the barrier. In our game, the only barriers were the edges of the screen.

(vi) For each trial, one or more target regions are specified. Each target has a fixed center position X_t and fixed width W_t in each spatial dimension.

(vii) Each target may have a positive or negative valence. Hitting any target causes the trial to end immediately—hitting a positive target means the trial ends successfully, and hitting a negative target means it ends unsuccessfully.

(viii) For each target j, a critical time window is specified: the minimum step number $T^{(j)}_{\min}$ and maximum step number $T^{(j)}_{\max}$ on which it can be hit.

(ix) A trial expires without a hit if the number of steps taken exceeds the maximum $T^{(j)}_{\max}$ across all targets j. An expired trial is considered successful if there were no positive targets (i.e. the only aim was to survive, avoiding negative targets), or unsuccessful if the trial contained one or more positive targets.

(x) At any time step, the **overlap** between the cursor and the target can be computed as

$$\frac{\min\left(X_t + \frac{W_t}{2}, X_c + \frac{W_c}{2}\right) - \max\left(X_t - \frac{W_t}{2}, X_c - \frac{W_c}{2}\right)}{\min\left(W_t, W_c\right)}.$$

For each target, a critical overlap value from 0 to 1 is specified, to define what constitutes a hit. In the fueling and adjustment phases of our game (see Section 2.4 of the main paper), our positive targets (the raindrops that the player had to collect) had a critical overlap value of 1: that is, a hit was only registered if the overlap was ≥ 1, indicating that the cursor completely swallowed the target or vice versa. In the flying phase, our negative targets (the rockets that the player had to avoid) had a critical overlap value of 0: that is, a hit was registered as soon as the overlap was ≥ 0, indicating that the edge of the cursor had touched the edge of the target.[‡]

Each simulation of each trial is conducted as a **Rayleigh flight**, i.e. a random walk whose step sizes are normally distributed (except where steps are truncated to prevent the cursor passing through a barrier). For each time step, a random step vector is drawn from a normal distribution with mean $\mathbf{0}$ and covariance matrix Σ. The covariance Σ is estimated from the real data, by taking the sample covariance of all the non-truncated steps in a given scope. Thus, the variance of the distribution of step sizes (and, for multiple spatial dimensions, the correlation between components of the step along different dimensions) is matched to the distribution observed in the original game. We also match the path's *smoothness*, as characterized by its autocorrelation at an offset of 1 step.[§] To match this, for each spatial dimension k we estimate r_k, the correlation coefficient between all consecutive pairs of non-truncated steps in the original scope. We

[‡] For > 1 spatial dimensions, our default implementation of this collision-detection algorithm assumes that the critical overlap value must be equalled or exceeded in all spatial dimensions simultaneously for a hit to be registered—strictly, this is realistic only for rectangular-edged cursors and targets oriented parallel to the frame of reference, but it may be a close enough approximation for many purposes.

[§] Note that the choice of trajectory parameters (covariance and smoothness) might reasonably be expanded in future to encompass more sophisticated simulated models of BCI users—see [4, 5] for an approach to such models.

then apply a digital causal filter to the time-series of random step sizes, with numerator coefficient $b_0 = \sqrt{1 - r_k^2}$ and denominator coefficients $a_0 = 1$ and $a_1 = -r_k$. The estimation and filtering procedures are repeated separately for each spatial dimension. The filtering operation does not affect the variance or the spatial covariance of the step sizes. Therefore, the result is a time series of random step vectors whose covariance and unit-offset autocorrelations are matched to the original measurements made in the current scope. Finally, this series is numerically integrated into a cursor trajectory, starting at the initial cursor position for the trial in question. During integration, if a step would cause the cursor to have > 0 overlap with a barrier in all dimensions, the step is truncated such that the overlap is exactly 0 in each dimension. The trajectory is then checked against the target positions to determine whether the simulated trial has ended successfully or unsuccessfully according to rules (vii) and (ix).[‖] An example of the random walk trajectories is given in Figure S1. Assuming equal numbers of simulations per trial, the proportion of successful simulations within one scope (for example, the overall proportion of pink to gray traces in Figure S1) is taken as an estimate of P_0 for that scope.

At `http://schalklab.org/downloads` we provide Python code for performing random walks of the kind we have described in this section.

[‖] Note that, although rule (vi) stipulates that targets remain fixed throughout a trial, a target moving along a deterministic path can easily be simulated as multiple targets with different positions and consecutive critical time windows.

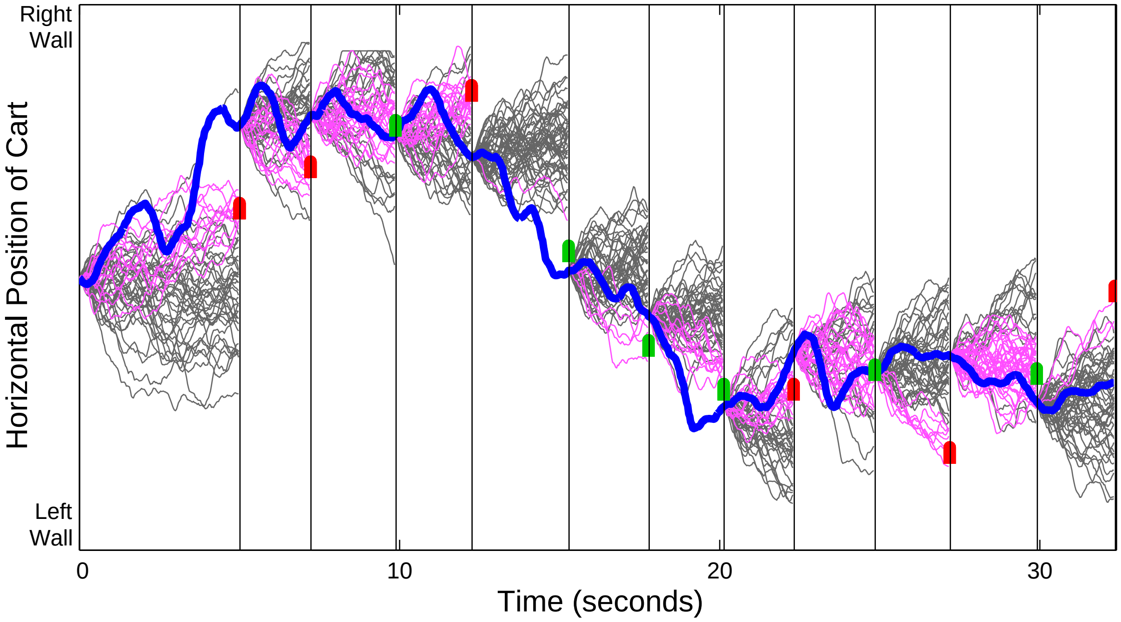

Figure S1. This figure shows the same example performance measurement sequence as Figure 1 of the main paper. The thick blue line indicates the center position of the cursor at each time step. The vertical gray lines indicate the times at which trials end. The small rectangular patches indicate the spatial extent of the targets and the temporal window during which the cursor had to catch them. Light green patches indicate targets that were hit, whereas dark red patches indicate targets that were missed. Thinner gray and pink lines indicate the first 50 random-walk simulations for each trial: gray lines indicate misses during simulation, and pink lines indicate hits during simulation.

S2. Error bars and confidence intervals on RIG$_B$

Having observed PM successes out of M trials, a popular approach to statistical inference on this outcome is to use the Jeffreys interval [6]. The Jeffreys interval is computed by taking quantiles of a beta distribution with shape parameters $PM + \frac{1}{2}$ and $(1 - P)M + \frac{1}{2}$, with special handling when the observed proportion of successes is 0 or 1. For example, if $I(a, b)$ denotes the cumulative distribution function for the beta distribution with shape parameters a and b, then the limits of a two-tailed 95% confidence interval are given by

$$P^{[0.025]} = \begin{cases} 0, & \text{if } P = 0; \\ I^{-1}_{0.025}\left(PM + \frac{1}{2},\ (1-P)M + \frac{1}{2}\right) & \text{otherwise, and} \end{cases}$$

$$P^{[0.975]} = \begin{cases} 1, & \text{if } P = 1; \\ I^{-1}_{0.975}\left(PM + \frac{1}{2},\ (1-P)M + \frac{1}{2}\right) & \text{otherwise.} \end{cases}$$

Since our measure RIG$_B$ is a monotonically increasing function of P for a fixed success-by-chance probability P_0, one could then simply substitute the limit values $P^{[0.025]}$ and $P^{[0.975]}$ for P in our formula for RIG$_B$. Note, however, that this approach disregards uncertainty in the estimation of P_0 itself. Strictly, the predictive distribution of P_0 is also a beta distribution, with shape parameters $P_0SM + \frac{1}{2}$ and $(1 - P_0)SM + \frac{1}{2}$, where S is the number of random-walk simulations performed per trial. The joint distribution of P and P_0 is a product of the two beta distributions, and a confidence interval for a function of both P and P_0 may be estimated by sampling—for example, by dense two-dimensional grid sampling, or by a Markov-chain Monte Carlo method such as the Metropolis-Hastings algorithm [7].

We compared confidence interval boundaries computed according to the fixed-P_0 assumption with those computed by 1000×1000 grid sampling and by the Metropolis-Hastings method using 5000 iterations, for various values of N, P and P_0. We found that, as long as the number of random-walk simulations exceeds around 200 per trial, the assumption of fixed P_0 changes the confidence interval boundaries only very slightly (typically by less than 1% of the interval width) relative to the values found by the sampling approaches.

Therefore, we adopt the approach of simply plugging in the Jeffreys interval limits for P into our formula for RIG$_B$ on the assumption that our estimate of P_0 is correct. A standard error bar can be computed this way, using $P^{[0.159]}$ and $P^{[0.841]}$ to achieve the same coverage that one would expect from the mean ± one standard error in an estimate of a normally-distributed variable.

Numerical example: suppose a BCI user were to perform 60 trials, each lasting 1 sec on average, and succeeds on 45 of them, so that we have $P = 0.75$, $M = 60$ and $\bar{t} = 1$ sec. Then suppose that, from our random-walk simulations, we estimate $P_0 = 0.3$ as the probability of success by chance. Our estimate of the rate of information gain is RIG$_B = 0.620$ bits per second. Finally, suppose we want to compute an error bar whose coverage would be equivalent to the mean ± 1 standard deviation of a normally-distributed variable, i.e. a confidence interval that covers quantiles 0.159 to 0.841 of the predictive distribution. The confidence interval on P is $[0.690, 0.801]$ and hence the interval on RIG_B is $[0.466, 0.775]$.

S3. Variation in BCI performance from session to session

In Section 3.1 and Figure 2 of the main paper, we report session-by-session performance measurements of four subjects in four different controller conditions. Two of the subjects, B and C, performed better than change with the BCI Controller, but their performance (particularly that of Subject B) was highly variable from session to session when contrasted with performance in the other controller conditions. The within-session variability (among the three repetitions of the performance measurement procedure performed on each session) was small relative to the between-session variability, and we interpreted this to mean that the between-session variability was not due to any intrinsic property of our performance measurement system, but rather due to day-to-day variability in the signal-to-noise ration of the EEG signals themselves.

The data from each session's calibration trials support this interpretation. As an illustration, see Figure S2, where SCD values (signed coefficient of determination, otherwise known as signed r^2) are plotted on the scalp to indicate how well bandpower features distinguish left-hand from right-hand calibration trials performed by Subject B. The scalp maps show SCD for bandpower features at a range of frequencies (from left to right), for four consecutive sessions (top to bottom). Note that the features are highly variable from session to session not only regarding the magnitude of the SCD (which indicates the degree of separability of the two classes) but also with regard to which frequency band is most influential, and even which hemisphere. Note in particular that session 4 (third row) was a bad day for this subject's EEG signals: the MRE_d during online BCI performance was no better than the Random Baseline, and this is corroborated by the particularly low SCD values during the calibration phase.

This interpretation is further corroborated if we summarize each calibration session with a single r^2 value, by summing of r^2 across channels C3, C4, CP3, CP4, P3, and P4 and across frequencies from 9 Hz to 24 Hz. If we then compute the Spearman correlation, across all of the sessions performed by subjects B & C, between these summary r^2 values and mean MRE_d values, we find a significant correlation ($\rho = 0.52, p = 0.0216, N = 19$). Note that the r^2 values themselves should be expected to be noisy, since they are each based on only 40 trials per session, and that the calibration trials were performed separately from the actual performance measurement sequences. Nonetheless, the day-to-day variations in the signal-to-noise ratio of the subjects' sensory-motor rhythms clearly predict variations in MRE_d to a significant extent.

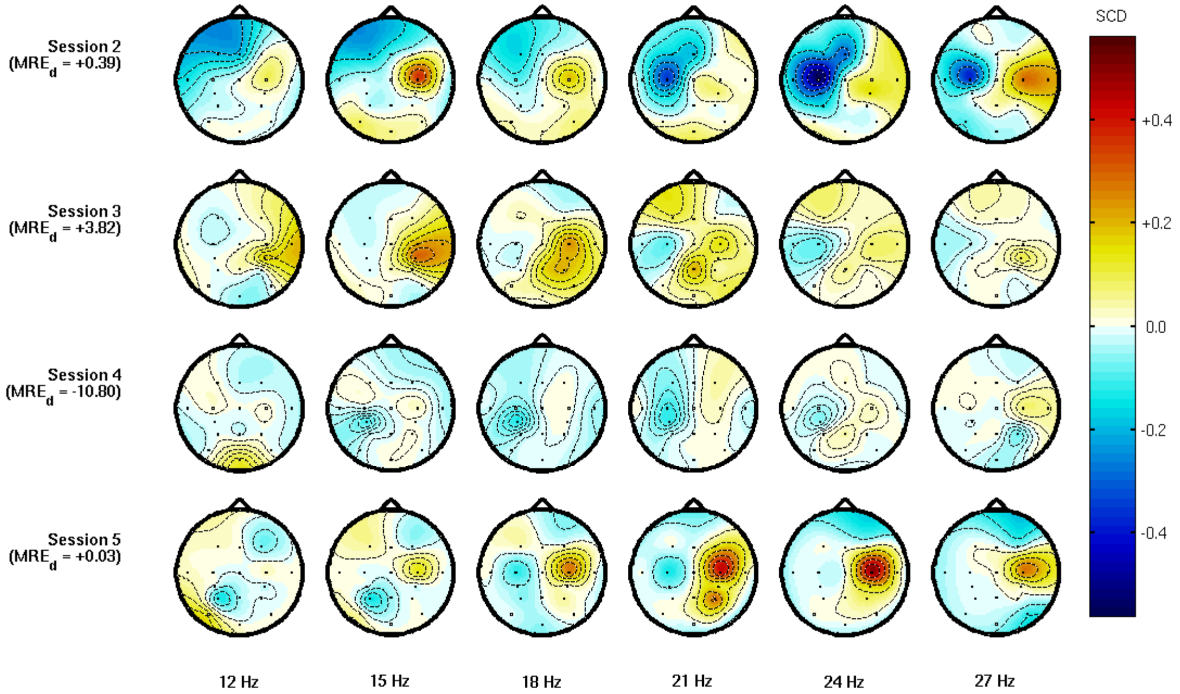

Figure S2. This figure shows example scalp maps from Subject B, illustrating session-to-session variability in BCI signal quality. Signed coefficients of determination (SCDs, or signed r^2 values) are plotted on the scalp for bandpower features computed after surface-Laplacian spatial filtering. The SCD values indicate the signal-to-noise ratio in the discrimination of left-hand from right-hand imagined movement trials during the calibration trials at the beginning of each session. Each column corresponds to a 3-Hz frequency in the range from 9 Hz to 24 Hz. Each row corresponds to a session, from number 2 to number 7. The corresponding mean of the adaptive performance measure MRE$_d$ is also given for each session (these values correspond to the first four points on the blue solid line in the second panel of Figure 2 of the main paper).

S4. Discussion of performance measurement approaches

In the Introduction of the main paper, we described one of our challenges, Challenge II, as the need to develop a general performance metric that could allow comparison of performance between different task contexts.

Since our approach to this challenge, an analysis based on information gain, is applicable to a wide variety of tasks and contexts, it is worth discussing its properties in more detail relative to those of related approaches.

In Section S4.1, we contrast our approach with a popular approach based on Fitts' Law. In Section S4.2, we explore the relation of our metric RIG_B to two measures of information transfer rate (ITR) that are also frequently used in BCI studies, namely Wolpaw's ITR and channel capacity (a.k.a. Nykopp's ITR).

S4.1. Information gain vs. Fitts' Law

One very influential approach to measuring control performance is an analysis based on Fitts' Law. Based on the work of Fitts [8], this law states that the average time taken to complete a movement, $\bar{t}$, is proportional to the **index of difficulty** of the movement ID_F, measured in bits. Thus, if $\bar{t} = a + b\,\mathrm{ID}_F$, experiments are designed to require movements of different lengths to allow estimation of a and b. An **index of performance** IP_F, in bits per unit time, is then calculated—usually as $1/b$—to compare performance across studies, conditions, subjects and interfaces. ID_F is usually defined according to the Shannon formulation due to MacKenzie [9]: $\mathrm{ID}_F = \log_2{(1 + D/W)}$, where D is the length of the required movement, and W is either the width of the target or some multiple of the standard deviation of terminal positions (measured in either case in the direction of movement). Thus, $1/W$ reflects either the required or the achieved precision of successful movements.

Fitts' Law analyses (FLA) have been used extensively in human-computer interaction research, since the linear relation has been found to hold very reliably and with strikingly little residual error under many conditions. The approach is part of the ISO standard for evaluating the performance of pointing devices [10]. In the BCI literature, Felton et al. [11] reported that Fitts' Law held for selected abled-bodied and disabled subjects, and selected difficulty levels, in the control of one-dimensional velocity of a cursor by modulation of EEG sensory-motor rhythm power. A trial was labeled successful if the subject caused the cursor to dwell within a target zone of width W for at least 500 ms, and unsuccessful if the subject did not achieve this before 15 seconds had elapsed. Simeral et al. [12] used a modification of the ISO task in combination with FLA when assessing the performance of a tetraplegic person using the BrainGate 2 system to steer a cursor in two dimensions and "click" on targets. Movement and "clicking" were achieved by modulating the spike rates of neurons in motor-cortex, as measured using an implanted microelectrode array. The authors observed significant linear correlations between trial duration and ID_F in most sessions. Gilja et al. [13] used FLA to analyze the performance of rhesus monkeys who used the firing rate of neurons in motor-cortex to steer a cursor in a two-dimensional center-out 8-target task and acquire them within 4 s by dwelling for 500 ms.

We have chosen to develop alternative methods of performance assessment for BCI, because of the following three limitations of FLA:

i). Assumption of negligible rates of failure Generally, FLA assumes that the task can be performed with a negligible rate of failure. Although methods have been proposed for making approximate corrections to ID_F based on empirical error rates [14], the analysis typically begins by excluding unsuccessful trials. This had a large impact on the study by Felton et al., even in their very simple task. Subjects were selected according to their performance: only those who achieved $\geq 90\%$ success rates were included in the analysis. Of the 12 disabled subjects, 7 were removed because they failed to reach this criterion. The able-bodied subjects were recruited based on previous high BCI performance, and had to pass this criterion level again for inclusion in the results. Even then, trials at the highest ID_F had to be excluded due to large error rates. All this data selection was presumably necessary to make Fitts' Law fit—this does not necessarily invalidate the results, but it clearly indicates that they only capture a fraction of the information we would like to know about performance. The two invasive studies benefitted from better signal-to-noise ratios, and also longer time periods during which the subjects could be trained and the task difficulties adjusted. Gilja et al. [13] were able to adjust the difficulty of their task such that both their monkeys achieved success rates of $> 90\%$, and Simeral et al. [12] observed an average success rate of 91% with their implanted human subject.

In the Introduction of our main paper, we described Challenge I as the need for methods that capture a wide range of performance levels on the same scale. We decided that the assumption of negligible rates of failure made FLA unsuitable for addressing this challenge, since BCI data in general contain a large proportion of errors, and these contain information we would like to use.

ii). Suitability for a limited range of tasks The second limitation is that the near-perfect-performance assumption of FLA means that it is only suitable for situations in which the subject can increase or reduce the speed of movement at will, and can thereby trade off speed for accuracy. It works well for motor pointing tasks and for tunnel navigation tasks (such as sub-menu selection with a mouse) because it is generally possible for subjects to achieve near-perfect performance in these tasks by slowing down. However, it remains to be seen whether this property holds for BCI systems, even for the subset of trials that are performed correctly. In Felton's study this is unclear because, apart from the fixed timeout at 15 seconds, there was no action or contingency analogous to a mouse-click that brought the trial to an unequivocal end, whether successful or unsuccessful. Instead, the longer the trial continued, the more opportunities there were for the cursor to dwell on the target for the required time, whether as a result of control or of chance. Hence, such a design automatically introduces a positive effect of target difficulty on the duration of correct trials, even if the subject has insufficient control over speed to trade it for accuracy. Simeral et al. also seem to express doubt that their subject had this type of control over the BrainGate, reporting an increased error rate for smaller as compared with larger targets.

In the Introduction of the main paper, we described Challege II as the need for a general performance metric that can be used to compare performance fairly across a wide range of tasks. We decided that FLA was unsuitable to address this challenge due to its applicability only to tasks in which speed can be traded for accuracy. It also excludes tasks like the current study's fixed-paced raindrop-catching task in which, even if the subject is able to perform some kind of beneficial speed-accuracy tradeoff, the effects are not manifested in differences in measured trial duration.

iii). Lack of invariance to nuisance parameters The third limitation is that it is difficult to compare control performance across studies using the FLA bit rate if there is any variation in "nuisance" task parameters, i.e. in parameters other than the two (target width and distance) that are reflected in ID_F. An obvious example of such a parameter in Felton's design is the required dwell-time, which was set to 500 msec. It is easy to see that manipulations of dwell-time might affect the difficulty of the task (albeit perhaps in non-trivial ways). Yet, this important change in actual difficulty will not be captured by the difficulty index ID_F on which bit rate estimates are based, since ID_F depends only on distance moved and target size. Thus, it is unclear whether performance of, say, 0.4 bits/sec measured with a required dwell-time of 500 msec represents better or worse BCI control than 0.3 bits/sec when the required dwell-time is 1 sec. Generally, when we consider the wide range of tasks for which neuroprosthetic devices might be used, we imagine that there will be many such nuisance parameters. Though it is possible to reduce this problem by standardizing each single-purpose task (as in the ISO standard for pointing devices, for example), this does not help us to compare control performance across tasks. Even for individual tasks, it is likely that many researchers will find their own good reasons for making study-specific departures from the standards, as Simeral et al. did in modifying the ISO standard.

This third limitation was also a factor in rejecting FLA for Challenge II: inability to account for nuisance parameters further limits the scope for comparing experiments, whereas we would like to design an index of difficulty that already accounts for a wide range of possible task parameters.

These issues highlight an urgent need for alternative methods of performance assessment—a need which Simeral et al. also highlight in their paper.

Our approach to task difficulty was complementary to that of FLA. Like FLA, we aimed to equalize the success rate P regardless of users' ability, but to do so we adapted our task difficulty using a staircase procedure that, unlike the FLA approach, ensures performance is *below* ceiling. We then based our performance measurement on the empirical success rate relative to the performance P_0 of a chance model such as the random-walk simulation model introduced in Section 2.8 of the main paper and explained in greater detail in Section S1. Our simulation method for estimating P_0 aimed to incorporate as many as possible of the experimental conditions and constraints experienced by the user, so that the outcome reflected a large number of parameters that would be nuisance parameters in FLA. The difficulty of a particular trial, or set of trials, is indicated by the resulting P_0 estimate, and can be expressed in bits as an index of difficulty $\mathrm{ID}_B = -\log_2 P_0$.

We also required an index of performance, in bits per unit time. Important considerations for different ways to calculate bit rate are discussed in the next section.

S4.2. Information gain vs. channel capacity

In Section 2.7 of the main paper we introduced an index of performance for assessing observed success rate P, relative to estimated chance performance P_0, over the course of a number of discrete trials of average duration $\bar{t}$. The formula is reproduced here in equation (1) for convenience:

$$\mathrm{RIG}_B = \mathrm{sign}(P - P_0) \ \times \ \left[P \log_2 \frac{P}{P_0} + (1-P) \log_2 \frac{1-P}{1-P_0} \right] \ / \ \bar{t} \ . \qquad (1)$$

Our term RIG_B stands for the $\underline{\text{r}}$ate of $\underline{\text{i}}$nformation $\underline{\text{g}}$ain between two $\underline{\text{B}}$ernouilli distributions. The term in square brackets is the **information gain** term, computed as the Kullback-Leibler divergence of the Bernoulli distribution $\{P_0, 1 - P_0\}$ reflecting chance performance, from the Bernoulli distribution $\{P, 1-P\}$ reflecting the empirically-observed performance.

This information-gain term is also a simple generalization of the well-known formula for information transfer proposed by Wolpaw et al. [15]. Wolpaw's criterion has been used in many studies to evaluate the performance of BCI systems that produce one of a fixed number N of discrete outputs (target classes or symbols). Wolpaw et al. originally defined the information transfer measure B, in bits per trial, as in equation (2) below, but it is easy to substitute $P_0 = 1/N$ and show their measure's equivalence to our information gain term (3):

$$\begin{aligned} B \ &= \ \log_2 N + P \log_2 P + (1-P) \log_2 \frac{1-P}{N-1} && (2) \\ &= \ P \log_2 \frac{P}{P_0} + (1-P) \log_2 \frac{1-P}{1-P_0} \quad \text{where} \ \ P_0 = \frac{1}{N}. && (3) \end{aligned}$$

This formula has attracted some criticism, which we will discuss below. Nonetheless, it has proven very popular in the BCI literature since it provides a simple way of comparing performance across studies. Its appeal lies in the fact that it takes account of two factors (number of target classes, and rate of target selection) that vary widely between studies and hence make their results incomparable based on P alone.

Wolpaw et al. acknowledged that equation (2) rested on certain assumptions. Other authors—for example, Schlögl et al. [16]—correctly point out that the assumptions often do not hold in many BCI tasks. The assumptions are: (i) that trial outcomes are divided exhaustively into a finite number N of discrete non-overlapping targets; (ii) that the user attempts each of the N targets with equal frequency; (iii) that the user's hit rate is equal regardless of which target is attempted; and (iv) that misses, when they occur, hit each of the unwanted targets with equal probability. While it is technically correct to point out that these assumptions often fail, it is misleading to argue that this is a weakness in all such cases. It is true that, if the assumptions are broken, equation (2) no longer exactly returns the BCI system's **channel capacity** (a measure that is sometimes called the "Nykopp information-transfer rate" due to its adoption in the context of BCI in the Masters thesis of Nykopp [17]). However, this is not necessarily a problem: for example, we will argue below that when assumption (iv) is violated, equation (2) remains a valid evaluation criterion and is in fact *more* relevant than channel capacity in such cases.

For the purposes of the current study, the main advantage of reformulating the Wolpaw et al. criterion in the form of RIG_B is that it frees us from assumption (i) since it can naturally be applied to tasks in which the targets are not exhaustively binned into N discrete categories. Instead, its interpretation as information gain between two Bernoulli distributions of hits and misses means that we require only a rule for distinguishing successful trials from failures and a suitable null-hypothesis model that provides an estimate of the chance level success rate P_0.

The details of the procedure for estimating P_0 are crucial, because they allow us to address assumptions (ii) and (iii)—the assumptions that all targets are equal regarding, respectively, the prior probability that the user will aim for them and the user's ability to hit them once they have decided where to aim. As Thomas et al. [18] point out, equation (2) is an unsuitable measure of performance when data-sets are unbalanced—i.e. when one target class is presented more frequently than others, in violation of assumption (ii). It will also be an unfair measure of performance when assumption (iii) is broken. In either case, we can see the problem as originating from a mismatch between the circumstances under which we measure P and the assumptions under which we obtain P_0. It is easy to see that equation (2)'s built-in assumption $P_0 = 1/N$ is not valid for unbalanced data-sets. This problem is automatically addressed by our re-simulation method for estimating P_0, which we described in Section 2.8 of the main paper, and in more detail in Section S1 above. In the current study, any given performance-measurement phase contained a mixture of trials of different difficulty levels, not only because we intentionally varied our task-difficulty variable d, but also due to other factors—for example, the fact that we prevented the cursor from overshooting the edges of the screen, thereby possibly making targets at the edges easier to hit than targets in the middle. The original data-set may contain an unequal mixture of target locations (which can be seen as a violation of assumption ii) and different difficulty levels (a violation of assumption iii). However, in all cases the starting conditions of each original trial are replicated in each simulated session: thus the distribution of different target locations and difficulty levels is the same when we estimate P_0 as it was when we estimated P.

Assumption (iv), the assumption that errors are evenly spread, merits special attention because an understanding of its role is pivotal to the question of what we actually want to measure when we evaluate BCI performance. Instead of a prerequisite for validity of Wolpaw's performance comparison measure, assumption (iv) should be regarded as an *assertion* that the distribution of errors is irrelevant when making performance comparisons. To make the rationale for this intuitive in non-mathematical terms, we will use an analogy with target-shooting.

Consider a shooter whose arrows or bullets all land within an inch of each other, but are consistently several inches too low on the target. Such consistent, narrow grouping indicates high information transfer—we would say that the **entropy** of the distribution of shots is very low. The shooter might reasonably infer from this that the elements of the process that are hardest to perfect—stance, breath control, concentration, action, etc.—are all very good. Something is still wrong, but the shooter can take advantage of the information content of the misses by adjusting the sights (or simply by aiming slightly higher) to bring the group into the center of the target. The greater the information content of the misses, the more precisely such compensatory adjustments can be made—they would not be possible at all if the misses had zero information content (i.e. they were evenly spread in all directions from the center). Therefore, in a practice session, our shooter can be fairly happy with this outcome, and proceed to make the necessary adjustment. By contrast, in a competition, there is no reward for good grouping *per se*: a miss is just counted as a miss, even if it is bunched together with the other misses. This is because competition organizers assume that competitors *have already had their chance* to make all the necessary adjustments and re-calibrations.

In an analogous way, Wolpaw et al. can be seen as implicitly applying competition-shooting criteria rather than practice-shooting criteria to BCI performance measurement. The assumption is not that we would like to, but cannot, incorporate the unaccounted-for information in the pattern of misses into our performance measure. Rather, by adopting the criterion of Wolpaw et al., we assume that the BCI engineers have already had their chance and will have made their best efforts, *before* the stage at which it is appropriate to apply the evaluation criterion, to exploit whatever information is in the misses and to turn them into hits. Such efforts might include adjustments in the way that symbols are encoded by the human interface or decoded by the computer, for example. In real BCI systems, any such manipulation is likely to have an effect on the user, potentially altering the statistics of that user's brain signals in ways that are not trivially predictable. Therefore, it makes sense that the success or failure of these adjustments should be measured as part of the performance metric of the finalized BCI system. By the same token, channel capacity can be seen as a less-relevant BCI measure: it adopts a practice-shooting approach, and predicts the information throughput that *might have been* obtained if the encoding and decoding had been optimally restructured. In doing so, it ignores the question of whether re-coding is in fact practically possible, as well as the fact that any such re-coding would affect the user's behavior and brain signals in unpredictable ways.[¶]

The competition-shooting interpretation implicit in Wolpaw's formula becomes more evident when we rewrite it as equation (3): by applying this measure, we are effectively reducing performance to the simplest and most relevant measure (a Bernoulli distribution of hit and miss probabilities), deliberately disregarding the information content of the misses before computing a measure of information gain relative to the chance model. However, note that equations (2) and (3) *still* apply competition-shooting criteria incompletely. To avoid applying practice-shooting criteria inadvertently, an additional caveat is required when using them, because the minimum value of B does not occur when $P = 0$, but rather when $P = P_0$. Therefore, worse-than-chance performance can appear to have a higher B value than chance performance. This contradicts our intuitive interpretation of competition-shooting criteria, where we assert that hitting is always better than missing. The effect is most obvious when $P_0 = 0.5$: then B is a symmetric function about $P = 0.5$, such that highly-consistent wrong performance (say, $P = 0.01$) yields the same B value as equally consistent good performance ($P = 0.99$). Since a Bernoulli distribution is completely characterized by just a single parameter P, we can implement a simple fix: we monitor whether $P < P_0$ and adjust the information gain term accordingly when this is the case.[+] Specifically, we negate RIG_B in this case—hence the $\mathrm{sign}(P - P_0)$ term in equation (1). This is because, in applying competition-shooting criteria, we are evaluating the BCI from the point of view of a subsequent processing stage that assumes errors have already been corrected as far as possible—in other words, one which *trusts* the direction of the BCI's output. So, when $P < P_0$ we ensure the performance measure is less than 0, to reflect the fact that the BCI's output would be consistently misleading and therefore worse than random.

In a fairly frequently-cited conference paper, Kronegg et al. [20] recommend that equation (2) not be used, based on their claim that it underestimates bit-rates with increasing severity as N increases. Our decision to reject this recommendation can now be understood by considering two factors. The first factor is the fact that Kronegg et al. chose channel capacity, which embodies practice-shooting standards, as the gold standard "bit rate". Equation (2) can indeed fail to approximate channel capacity, but we have argued above that its departure from channel capacity is a desirable feature, because the implicit competition-shooting approach of the former makes it *more* relevant

[¶] In Nykopp's approach [17], potential re-coding is accounted for by the maximization of bit-rate over all possible input alphabets. This is appropriate when two *computer* systems communicate over a noisy channel: then, the sender is free to compress and re-code the input into an arbitrarily different alphabet without changing the properties of the channel itself. However, as we illustrated in Hill et al. [19], re-coding in a *human*-computer interface can have profound effects on the signal-to-noise ratio of the brain signals that are relevant to BCI. Since the BCI system's transmission channel must thus be expected to depend on the encoding in non-obvious ways, optimization over alphabets is not appropriate during data evaluation.

[+] However, note that it is less clear how to choose the best rule for avoiding this pitfall if we were to want to compute an information gain term for non-Bernoulli measures of success, such as an ordinal-valued game score or duration.

for BCI. The second factor is that, even in practice-shooting terms, the model from which Kronegg et al. computed gold-standard channel capacity values was inappropriate for BCI. It assumed that the feature distributions associated with the N classes were arranged ordinally along a single axis: thus, the overlap between class 3 and class 1 was smaller than the overlap between class 2 and class 1, and so on up to class N which overlapped with class 1 least of all. This created an extremely and atypically rich structure of information in the misses: in real BCI systems based on multivariate feature distributions, we cannot assume that the confusability of each pair of classes can be ordered in this highly structured way. Thus, even if we were to accept practice-shooting criteria, the gold standard of Kronegg et al. could be expected to *overestimate* information transfer with increasing severity as N increases, relative to a BCI system based on realistic assumptions.

References

[1] Hill N. J, Moinuddin A, Häuser A.-K, Kienzle S. & Schalk G. (2012). *Communication and Control by Listening: Toward Optimal Design of a Two-Class Auditory Streaming Brain-Computer Interface.* Frontiers in Neuroscience, **6**: 181.

[2] Tangermann M. W, Krauledat M, Grzeska K, Sagebaum M, Vidaurre C, Blankertz B. & Müller K.-R. (2009). *Playing Pinball with Non-Invasive BCI.* Advances in Neural Information Processing Systems, **21**.

[3] Bansal A. K, Vargas-Irwin C. E, Truccolo W. & Donoghue J. P. (2011). *Relationships among low-frequency local field potentials, spiking activity, and three-dimensional reach and grasp kinematics in primary motor and ventral premotor cortices.* Journal of Neurophysiology, **105**(4): 1603–1619.

[4] Quek M, Boland D, Williamson J, Murray-smith R, Tavella M, Perdikis S, Schreuder M. & Tangermann M. (2011). *Simulating the Feel of Brain-Computer Interfaces for Design, Development and Social Interaction.* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, **2011**: 25–28.

[5] Quek M. (2013). *The Role of Simulation in Developing and Designing Applications for 2-Class Motor Imagery Brain-Computer Interfaces.* PhD thesis, University of Glasgow.

[6] Brown L. D, Cai T. T. & DasGupta A. (2001). *Interval Estimation for a Binomial Proportion.* Statistical Science, **16**(2): 101–133.

[7] Mackay D. J. C. (2003). *Information Theory , Inference , and Learning Algorithms.* Cambridge University Press, Cambridge, UK.

[8] Fitts P. M. (1954). *The Information Capacity of the Human Motor System.* Journal of Experimental Psychology, **47**(6): 381–391.

[9] MacKenzie I. S. (1992). *Fitts' law as a research and design tool in human-computer interaction.* Human-Computer Interaction, **7**: 91–139.

[10] ISO. (2002). *Reference number ISO 9241-9:2000(E): Ergonomic design for office work with visual display terminals (VDTs) part 9: Requirements for non-keyboard input devices.* Technical report, International Standards Organization.

[11] Felton E. A, Radwin R. G, Wilson J. A. & Williams J. C. (2009). *Evaluation of a modified Fitts law brain-computer interface target acquisition task in able and motor disabled individuals.* Journal of Neural Engineering, **6**(5): 056002.

[12] Simeral J. D, Kim S.-P, Black M. J, Donoghue J. P. & Hochberg L. R. (2011). *Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array.* Journal of Neural Engineering, **8**(2): 025027.

[13] Gilja V, Nuyujukian P, Chestek C. A, Cunningham J. P, Yu B. M, Fan J. M, Churchland M. M, Kaufman M. T, Kao J. C, Ryu S. I. & Shenoy K. V. (2012). *A high-performance neural prosthesis enabled by control algorithm design.* Nature Neuroscience, **15**: 1752–1757.

[14] Soukoreff R. W. & MacKenzie I. S. (2004). *Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts law research in HCI.* International Journal of Human-Computer Studies, **61**(6): 751–789.

[15] Wolpaw J. R, Ramoser H, McFarland D. J. & Pfurtscheller G. (1998). *EEG-based communication: improved accuracy by response verification.* IEEE Transactions on Rehabilitation Engineering, **6**(3): 326–33.

[16] Schlögl A, Kronegg J, Huggins J. E. & Mason S. G. (2007). *Evaluation Criteria for BCI Research.* In Dornhege G, Millán J. d. R, Hinterberger T, McFarland D. J. & Müller K.-R (Eds.), *Toward Brain-Computer Interfacing*, pages 327–342. MIT Press, Cambridge, MA.

[17] Nykopp T. (2001). *Statistical Modelling Issues for The Adaptive Brain Interface.* Masters thesis, Helsinki University of Technology, Finnland.

[18] Thomas E, Dyson M. & Clerc M. (2013). *An analysis of performance evaluation for motor-imagery based BCI.* Journal of Neural Engineering, **10**(3): 031001.

[19] Hill J, Farquhar J, Martens S. & Schölkopf B. (2009). *Effects of Stimulus Type and of Error-Correcting Code Design on BCI Speller Performance.* Advances in Neural Information Processing Systems, **21**: 665–672.

[20] Kronegg J, Voloshynovskiy S. & Pun T. (2005). *Analysis of bit-rate definitions for Brain-Computer Interfaces.* In *Proceedings of the International Conference on Human-Computer Interaction.*