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

Abstract

When researchers evaluate brain-computer interface (BCI) systems, we want quantitative answers to questions such as: How good is the system's performance? How good does it need to be? and: Is it capable of reaching the desired level in future? In response to the current lack of objective, quantitative, study-independent approaches, we introduce methods that help to address such questions. We identified three challenges: (I) the need for efficient measurement techniques that adapt rapidly and reliably to capture a wide range of performance levels; (II) the need to express results in a way that allows comparison between similar but non-identical tasks; (III) the need to measure the extent to which certain components of a BCI system (e.g. the signal processing pipeline) not only support BCI performance, but also potentially restrict the maximum level it can reach. For challenge (I), we developed an automatic staircase method that adjusted task difficulty adaptively along a single abstract axis. For challenge (II), we used the rate of information gain between two Bernoulli distributions: one reflecting the observed success rate, the other reflecting chance performance estimated by a matched random-walk method. This measure includes Wolpaw's (1998) information transfer rate as a special case, but addresses the latter's limitations including its restriction to item-selection tasks. To validate our approach and address challenge (III), we compared four healthy subjects' performance using an EEG-based BCI, a "Direct Controller" (a high-performance hardware input device), and a "Pseudo-BCI Controller" (the same input device, but with control signals processed by the BCI signal-processing pipeline). Our results confirm the repeatability and validity of our measures, and indicate that our BCI signal-processing pipeline reduced attainable performance by about 33% (21 bits/minute). Our approach provides a flexible basis for evaluating BCI performance and its limitations, across a wide range of tasks and task difficulties.

---

*Correspondence: Jeremy Hill, Wadsworth Center, C640 Empire State Plaza, Albany, NY 12201, USA. jezhill@gmail.com. The authors report no conflicts of interest.

Keywords

Brain-Computer Interface; Neuroprosthetic Control; Control; Performance Assessment; Computer Games; Information Transfer Rate; Information Gain; Relative Entropy

1. Introduction

Many studies over the past few decades have focused on research and development of brain-computer interface systems—see [1, 2] for review. According to the definition in Wolpaw and Wolpaw [2], a brain-computer interface (BCI) is a system that translates activity of the central nervous system into an artificial output signal that can replace, restore, enhance, supplement or improve conventional central-nervous-system outputs. Such systems are also called brain-machine interfaces (BMI) or neuroprosthetics.

BCIs can *replace* important functions normally served by the motor system by allowing people to use brain signals, instead of muscles, to control the functions of a computer or the movements of a prosthetic limb or other external device. Such BCIs have the inspiring potential to improve the lives of people who are paralyzed due to disabling neurological or neuromuscular disorders.

Previous research has included demonstrations of BCI control using neuronal firing rates detected using intracortical implants (e.g. [3–5]), population-level activity measured using subdural electrocorticographic (ECoG) arrays [6, 7], and sensory-motor rhythms extracted from electroencephalographic (EEG) recordings from the scalp [8–11]. These studies are impressive demonstrations of the potential of BCI control. However, one of the most vexing, elusive, widely acknowledged problems of BCI research is that the performance of such demonstrations is actually very low when measured against the demands of real-world tasks, or against the performance of competing control methods for prosthetics and other assistive devices. For example, for BCIs that support continuous movement control, BCI performance is still substantially slower and more variable than muscle-based control [10]. While a success rate of, say, 95% would be considered very impressive in most BCI target tasks, the same performance (i.e. one failure per 20 attempts) falls far short of the human motor system's reliability in performing important tasks of even greater complexity, such as grasping and picking up objects without dropping them.

Thus, a critical question for the future of BCI technology is the degree to which performance can be increased and its minute-to-minute and day-to-day variation can be decreased. Such improvements hinge on the ability to compare and contrast different BCI approaches systematically, to allow the most promising approaches to be identified.

Whenever a BCI demonstration is published, researchers would like to be able to quantify its performance in a way that allows meaningful comparison with other data. The first question is "how good is it?" This can be posed in a number of different ways—for example: how good is the BCI relative to other assistive technology that might be used to do the same job? How good is the BCI relative to the level of performance necessary to perform useful real-world tasks safely? How good is the BCI relative to competing BCI approaches that have

similar goals? The second question is "how good can it get?" If BCI performance is currently not close to the desired level, then is it at least theoretically possible for performance to improve—perhaps by user training—until the desired level is reached? Or, might there be some fundamental limitation, intrinsic to the way the brain signals are elicited, measured and translated, that will prevent the BCI's performance from ever exceeding a certain level?

Unfortunately, three critical shortcomings of current performance measurement approaches greatly impede such systematic evaluations. First, most current methods use fixed levels of task difficulty and thus cannot readily be applied across the whole possible spectrum of BCI performance—for example, from current levels of performance to the levels we would like to see for real-world BCI usage. Second, current methods do not readily provide metrics that allow performance comparison across similar but non-identical tasks. For example, two laboratories may both report results on control of a prosthetic arm, but the contraints within which the arm moves, and the task it is required to perform, will likely differ, so that it is unclear how the performance results may be compared. Third, current methods cannot determine to what extent limitations in BCI performance may be due to the intrinsic BCI methodology rather than the underlying abilities of the user. As an example, current BCI methods integrate information from comparatively long time periods (typically 50–500 milliseconds) to extract brain signal features such as single-neuron firing rates, population-level ECoG activity, or the amplitude of EEG oscillations. This temporal smoothing is necessary to increase the signal-to-noise ratio to a level that supports reasonable BCI performance, by the standards we can reach today. However, any such smoothing operation imposes a limit on the maximum rate at which the system can transfer information. Therefore, we must consider the possibility that such necessary elements of current BCI approaches actually impose fundamental limits on the level of performance that BCI systems can *ever* reach, regardless of such factors as the amount of time invested in user training.

Due to these shortcomings of current performance assessment methods, we do not know where such fundamental limits lie relative to the practical demands of everyday tasks, and we are ill-equipped to quantify users' progress meaningfully during training. Furthermore, scientists who are setting out to improve BCI performance must either compare performance only within narrow task parameters, or resort to subjective choices and personal preferences rather than objective and widely applicable criteria. This has left substantial room for several long-standing debates about the relative performance characteristics of different BCI approaches (for example, invasive vs. non-invasive methods). In consequence, a central need in BCI research is to establish a generally applicable methodology that can provide the basis for objective comparisons. In this paper, we describe and demonstrate a package of methods that supports such objective comparisons. It addresses the following three important challenges:

Challenge I: Build an Efficient Adaptive Performance Measurement System. The first challenge was to establish a performance measurement scale, and a procedure for making measurements efficiently on the scale, with which we can adaptively capture performance at all possible levels. We also wanted to equalize the degree to which a user's capabilities were

challenged, and the user's consequent success rate, as far as possible across users and contexts. We addressed this challenge by basing our task implementation on a single abstract **task difficulty** variable that could be adjusted to make our task easier or harder to perform. Though the task difficulty variable could be linked to multiple parameters of the task, the crucial aspects of the design were (i) that the conditions experienced by very unskilled subjects and the conditions experienced by very proficient subjects were distinguished only by changes in the single underlying variable, and (ii) that the variable could be adjusted automatically without any intervention from the investigator. To adjust task difficulty automatically, we implemented an **adaptive staircase procedure** that was originally developed in the field of psychophysics—specifically, we used Kaernbach's weighted up-down method [12]. We used the staircase procedure's built-in method for within-study assessments of user performance—this returns a value on the axis of task difficulty.

Challenge II: Develop a Transferable Performance Metric. The second challenge was to express the results not in arbitrary task-difficulty units, but on a universal, familiar scale that will allow comparison of results across studies. Though many metrics exist for quantifying BCI performance (see [13–15] for reviews), many of these are highly specific to the context of particular tasks, particularly when the task requires movement control rather than item selection. There is little consensus regarding the measures that should be used to compare performance in one task (for example, a monkey feeding itself with a robot arm, as in [4]) with performance in another (for example, a tetraplegic human guiding a mouse cursor, as in [3]). Our strategy was to develop a **relative entropy** or **information gain** measure, quantified in bits per unit time. This measure reflects the extent to which a user's performance exceeds the performance we would expect by chance, under the null hypothesis that the user has no control over the BCI system. Importantly, this measure is identical to that proposed by Wolpaw et al. [16] in the specific case of equiprobable item selection (i.e. when the chance-level success probability is simply the reciprocal of the number of items). However, it can also be applied to movement control or other tasks that are different from item selection (i.e. tasks in which there is no easy way to determine *a priori* the performance we would expect by chance). To address the problem of estimating chance performance in this wider range of tasks, we develop and apply a general trial-by-trial random-walk simulation method—a strategy that has been adopted by some others in BCI movement control[17, 18].

Challenge III: Measure Limitations in BCI Performance. The third challenge was to use the measurement and evaluation techniques to assess not only what a BCI signal processing pipeline *enables* us to do, but also what *limits* it imposes on performance. The performance measurement methodology that resulted from our solutions to the first two challenges allowed us to address this challenge. We did so by conducting a within-subject performance comparison between a **Direct Controller** and a **Pseudo-BCI Controller**. The Direct Controller was a hardware input method with which a healthy user could attain a high level of performance in the task via conventional motor control. The Pseudo-BCI Controller used the same input device as the Direct Controller, but its control signal was processed using the signal processing pipeline that we used for BCI control. We refer to the difference in

performance between the two conditions as the **false performance ceiling** for the signal processing pipeline. It reflects the extent to which a particular system component restricts the performance a BCI user can achieve—even, perhaps, irrespective of the amount of training the user receives.

Our experimental demonstration of these approaches was a cursor task in which subjects had to catch falling targets. Both cursor width and target speed varied as a function of the underlying task difficulty variable. Our subjects modulated their sensory-motor rhythms to control this 1-dimensional computer game. While we used this particular task design and BCI approach in our validation experiments, the same principles and methods could readily be applied to any BCI-controlled system, whether invasive or non-invasive, whether 1-, 2- or 3-dimensional, and whether the effector is virtual or physical.

2. Materials and Methods

2.1. Subjects

Four healthy subjects took part in the experiment: two male and two female, all right-handed, aged 21, 28, 55 and 55. All subjects had normal or corrected-to-normal vision and no history of neurological defects. Some of them had previously taken part in EEG studies of BCIs based on event-related potentials (P300 speller systems) but none of them had had prior experience with BCI systems based on sensory-motor rhythms. Subjects gave informed consent according to a protocol approved by the Institutional Review Board of the Wadsworth Center. Each subject participated in ten 90-minute sessions on separate days (total: 60 subject-hours). One additional pilot subject also performed the experiment during development. The pilot subject's results are not reported, because we frequently re-tuned the method's parameters over the course of this subject's sessions, which prevented valid comparisons with other data. Apart from the pilot, there are no unreported subjects (subjects were not dropped from the analysis on the basis of performance).

2.2. Hardware and Software Setup

EEG recordings were made using a 16-channel g.USBamp series B amplifier (g.tec medical engineering GmbH, Austria) in conjunction with a 16-channel EEG cap (Electrocap, Inc.). The cap used gelled 9 mm tin electrodes at positions F3, Fz, F4, T7, C3, Cz, C4, T8, CP3, CP4, P3, Pz, P4, PO7, PO8 and Oz of the extended international 10–20 system of Sharbrough et al. [19], with the reference at TP10 (the right mastoid) and the ground electrode at TP9 (the left mastoid). The amplifier performed appropriate anti-alias filtering before digitizing with a resolution of 24 bits and downsampling to 256 Hz.

Data acquisition and signal processing were performed using the BCI2000 software platform [20, 21] v.3.0. Stimulus presentation was implemented in Python using the 'BCPy2000' add-on to BCI2000 [22]. The software executed on a Lenovo ThinkPad T61p laptop with a 2.2 GHz dual-core processor.

Two Wii Remote controllers or "Wiimotes" (Nintendo Co. Ltd., Japan) were connected to the computer via Bluetooth. Signals from their accelerometers were acquired with BCI2000 and synchronized with the EEG signals.

Data analyses were performed using custom Matlab code.

2.3. Controller Conditions

As we will describe in more detail in Section 2.4, the task involved one-dimensional control of a cursor, which had to be moved left and right on the screen in order to catch or avoid falling targets. The velocity of the cursor could be controlled in various different ways, which we will describe below.

In designing these different controller conditions, we set out to address two of the challenges described in the Introduction. First, we aimed to address Challenge I, the need for a measurement scale that allows us to assess the performance of a **BCI Controller** relative to the lowest possible floor (random chance performance) and a high ceiling (close to the performance achieved in daily tasks by a healthy human motor system). We designed a **Random Baseline** and a **Direct Controller** condition, respectively, to measure these two reference points. Second, we aimed to address Challenge III, the need to assess the limitations that current BCI methods might impose on the level of performance that a subject can reach. We addressed this by computing the difference between Direct Controller performance and performance in a condition we call the **Pseudo-BCI Controller**.

We refer to the BCI Controller, Direct Controller and Pseudo-BCI Controller as the **active** controller conditions because they all required the active participation of the subject (in contrast to the Random Baseline condition). In each 90-minute session, the subject played 3 games in each of the 3 active conditions, for a total of 9 games. The controller conditions were as follows:

- **BCI Controller**: the cursor was controlled by motor imagery of the hands: imagined left-hand movement caused the cursor to move left, and imagined right-hand movement caused the cursor to move right. As described in Section 2.6, EEG signals were translated into control signals via spatial filtering followed by temporal windowing, detrending, auto-regressive spectral amplitude estimation, differential linear weighting of amplitudes in chosen frequency bands, and normalization.

- **Direct Controller**: the subject held a NintendoWiimote in each hand. The cursor velocity was proportional to the total power of accelerometer fluctuations in the right Wiimote minus the total power in the left: hence, the more the subject shook the left-hand Wiimote, the faster the cursor would move to the left, and the more the subject shook the right-hand Wiimote, the faster it would move to the right. This condition was intended to be comparable with the BCI condition in the sense that control was still based on the difference between activity of the left and right hands. However, within this constraint, the purpose of the Direct Controller condition was to investigate our system's ability to measure performance levels that were as high as possible.

- **Pseudo-BCI Controller**: the subject held the Wiimotes and shook them as in the Direct Controller condition. However, translation into cursor velocity was different: the accelerometer power in each Wiimote inversely modulated the amplitude of an

artificial white noise signal, which was then passed through exactly the same signal processing pipeline that was applied to brain signals in the BCI controller condition, i.e. starting with the temporal windowing stage and ending with a separately-calibrated normalization stage. The purpose of this condition was to provide a contrast with the Direct Controller condition, by which we could evaluate the extent to which high performance was limited or otherwise affected by the signal processing pipeline used for BCI. The white noise played an analogous role to the sensory-motor rhythm in a real subject's EEG, in that it acted as a carrier for the amplitude modulation that encoded movement intention. Since white noise has energy at all frequencies, the modulation signal could be extracted by the processing chain we already optimized for the subject's BCI data, regardless of which frequency happened to have been chosen during optimization.

- **Random Baseline**: this condition was performed after the subject had left. It involved playing back the subject's EEG for each BCI game, but with a 3-minute time-shift, and running this through the BCI signal processing pipeline to generate a control signal to drive the game. Therefore, although the control signal was determined by input signals whose distribution of amplitudes and other temporal properties were identical to those of the original EEG used in the BCI Controller condition, the time-shift removed the temporal relationship between intended and required movements. The purpose of this condition was to establish a baseline for the performance that might be expected if one randomly moved the cursor left and right with similar speed, frequency and amplitude to the movements achieved by the subject in the BCI Controller condition.

The four controller conditions are illustrated schematically in the right panel of Figure 1.

2.4. Basic Gameplay

Each 90-minute session comprised 9 **game cycles**: 3 in each of 3 active controller conditions. Each game cycle consisted of multiple **fueling** and **flying** phases, followed by a final **measurement/adjustment** phase that provided a single measure of performance for the cycle and adapted the task difficulty for future cycles.

The fueling and flying phases were designed to accustom the player to the current playing conditions, stabilize their performance, and provide enough variety and goal-directed motivation to prevent their becoming bored. Players first had to collect fuel for their spaceship: the cursor took the shape of a fuel cart, which the player moved left and right along the bottom of the screen to catch drops of water that fell from a randomly-moving cloud at the top. Once the fuel cart had caught 10 drops, the player had to fly their spaceship towards a planet: the cursor, in the shape of a spaceship, stayed at the bottom of the screen and the player had to move it left and right to avoid missiles that scrolled down the screen towards it. If the spaceship struck a missile, the player was sent back to the beginning of the fueling phase, but the distance covered in the journey towards the planet was recorded, and the next flying phase would resume from the furthest point reached. The game cycle entered its final phase when 4 minutes of attempted fueling and flying had elapsed, or when the planet was reached (which typically took between 1 and 2 minutes when the subject's

control was good). The final phase, for measurement and adjustment, is described in the next section.

For each controller condition, three game cycles were performed consecutively with short breaks between them, resulting in three separate performance measurements per session per controller condition.

All three phases are exemplified in Movie S1 in the Supplementary Material: a subject is shown performing a measurement/adjustment phase first, followed by the first fueling phase and then the first flying phase of the subsequent game cycle, all in the BCI Controller condition.

2.5. Measurement and Adjustment

The concluding phase of each game cycle was similar to the fueling phase described above, in that the player had to move the cursor left and right to catch falling water droplets from the randomly moving cloud. However, during this phase the difficulty of the task was adjusted using the weighted-up-down psychophysical staircase procedure of Kaernbach [12]. The task difficulty level d (expressed in arbitrary units) was increased by an amount S_{up} every time the player caught a droplet, and decreased by an amount S_{down} every time the player missed. We set $S_{up} = 1.0$ and computed S_{down} according to Kaernbach's formula $S_{up}/S_{down} = (1-p)/p$, where p is the target hit rate on which the procedure converges (we used $p = 0.65$). Over a broad range of values, the task difficulty value d was mapped logarithmically to the size of the cursor: a unit increase in d meant a 10% reduction in width, although the cursor was never made smaller than 1/20, or larger than 1/2, of the width of the screen. To allow the range of difficulty levels to extend beyond these limits, task difficulty also determined the speed with which water droplets fell: whenever the cursor was at minimum size, or larger than 1/5 of the screen, a unit increase in d translated into a 10% increase in speed. During pilot testing we found subjectively that this had the additional advantage of increasing the pace of the game for more-proficient players, thereby preventing players from becoming bored.

The staircase procedure continued until the 8th reversal, i.e. until the change in d reversed direction 8 times. Discarding the first two reversals, the median of the d values at the last 6 reversals was computed: this is known as the **mid-run estimate** of task difficulty, which we denote by MRE_d. We recorded MRE_d as a measure of performance for the current game cycle, and used it as the starting difficulty level (in respect of both cursor width and speed) for the next game cycle.

An example of an adjustment phase is illustrated in the upper and lower left panels of Figure 1. The upper panel shows the time course of the cursor's position and width over the course of the game, relative to the spatio-temporal windows that the cursor must hit in order to catch the targets. Hits and misses cause step changes in the task difficulty variable d, plotted in the lower panel. The lower panel also illustrates how MRE_d is computed.

The measurement/adjustment phase is exemplified at the start of Movie S1 in the Supplementary Material.

2.6. Calibration and Signal Processing

Our procedures for calibration and signal processing are similar to those used in previous studies of cursor control using non-invasive BCI systems based on sensory-motor rhythms [8, 10]. We set up the BCI system in three phases: an initial cued motor-imagery calibration measurement phase; second, a phase in which feature-extraction and classification parameters were chosen for the current subject, in the context of a particular signal processing pipeline; finally, a second calibration phase in which the control signal was centered and standardized. The three phases and the signal processing pipeline itself were as follows:

Calibration Phase I (BCI): Before the first game cycle of each session's BCI Controller condition, the subjects performed 40 cued motor-imagery trials in response to text prompts on the video screen: 20 left-hand and 20 right-hand, in random order. Subjects performed motor imagery for 6 seconds on each trial and then relaxed for 2 seconds.

Signal Processing (BCI, Pseudo-BCI and Random): Signals were processed, both offline and in real time, using the BCI2000 software system. First, they were spatially filtered using a surface-Laplacian filter matrix, buffered in a 500 msec moving window (moving in steps of 31.25 msec), and linearly detrended. At each time step, spectral amplitudes were then estimated in 3 Hz bins using an auto-regressive model of order 20. Based on the motor imagery trials from Calibration Phase I, BCI2000's OfflineAnalysis tool was used by the experimenter to select the electrodes and frequency bins that would be positively or negatively weighted in the linear sum that produced the final control signal. A positive weight on bandpower meant that a reduction in bandpower due to event-related desynchronization (ERD) would move the cursor to the left, and negative weight meant that ERD would move the cursor to the right. The choice was limited to electrodes C4, CP4 and P4 for positive weightings (since we assumed these locations would best capture left-hand motor imagery signals) and to C3, CP3 and P3 for negative weightings (right-hand motor imagery). The choice of frequency bins was restricted to the 9–24 Hz range (μ and β bands).

Calibration Phase II (BCI and Pseudo-BCI): The subject then performed 20 further calibration trials, using the setup determined at the end of Phase I, but with BCI2000's Normalizer system turned on [21]. This system maintained a rolling buffer of control signal values, which was updated every trial, with balanced contributions from imagine-left and imagine-right trials. The Normalizer system used these data to compute, and to update after every trial, a linear offset and gain value that standardized the balanced control signal to mean 0 and variance 1. From the sixth trial onwards, the cursor was continuously visible, moving according to the standardized control signal from the real-time motor-imagery processing pipeline. At the end of this phase, the Normalizer returned the final offset and gain values that were then fixed for the remainder of the session. This calibration phase was performed separately for the BCI Controller and Pseudo-BCI Controller conditions.

2.7. Evaluation Criterion

In the Introduction, we described Challenge II as the need for a transferable performance metric that allows comparison between different experimental setups. To address this

challenge, we define a criterion that we call the rate of information gain (RIG$_B$), measured in bits per unit time. Specifically, we measure information gain between two Bernoulli distributions. A Bernoulli distribution is the simplest possible probability distribution, consisting of just two numbers: the probability of hitting a desired target and the complementary probability of missing it. Thus, our measure can apply to an assessment of any a set of events ("trials"), provided that each event can be judged unequivocally to have succeeded or failed. The BCI user's observed probability of success is denoted by P. We assume that there is some method of estimating P_0, the rate of success according to chance (i.e. under the null hypothesis that the BCI user has no control over the system). RIG$_B$ is computed by dividing the information gain in bits per trial by $\bar{t}$, the mean duration of a trial:

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

A numerical example, along with details of the method we use to compute standard error bars and other confidence intervals on RIG$_B$, can be found in the Supplementary Material in Section S2.

The term in square brackets in equation (1) is the **information gain** term, otherwise known as **Kullback-Leibler divergence**, Kullback-Leibler information criterion (**KLIC**), or **relative entropy**. More precisely, it is the Kullback-Leibler divergence of a Bernoulli distribution reflecting chance probability of success, from a Bernoulli distribution reflecting the empirically-observed probability of success.[‡] Thus, our information gain term quantifies the extent to which the user's hit-vs.-miss distribution departs from a model that assumes hits happen by chance [23, 24].

In principle it would also be possible to compute information gain for other measures of success—for example, a total *number* of hits obtained in time t, or survival duration in the flying phase of our game, or a correlation between ideal and actual trajectories, or average task completion time, or any other ordinal-valued game score. Such scores will no longer be Bernoulli-distributed, but the Kullback-Leibler divergence of a chance model from the data can still be computed, provided that there is a method for estimating the distribution of the chosen measure under the null hypothesis. The result will also be expressed in bits, although it is not meaningful to attempt to compare the information gain computed from one type of score (for example, one with a Gaussian distribution) with information gain computed from another (say, a Bernoulli-distributed indicator of success). For current purposes, we will stick to hit probabilities as a measure of success, and hence operate on Bernoulli distributions, and thus retain the B subscript on RIG$_B$ to stand for Bernoulli.

For performance levels at or above chance ($P \quad P_0$), our RIG$_B$ is a generalization of the well-known and frequently-used criterion introduced by Wolpaw et al. [16]. Wolpaw's information transfer rate (ITR$_W$) is equal to RIG$_B$ in the particular case where $P_0 = 1/N$, for some finite integer number N of discrete, non-overlapping, exhaustive target classes—as is

---

[‡]Note, however, that relative to a classic Kullback-Leibler divergence, our term is actually scaled by a factor of $(\ln 2)^{-1}$, which serves to convert the units into bits.

the case, for example in many item selection tasks. For the purpose of comparing BCI performance across conditions, users and studies, we find ITR_W to be a more relevant measure than **channel capacity**—a criterion from Information Theory, sometimes referred to as Nykopp's ITR, against which ITR_W has sometimes been compared.

Note that, even for $P_0 = 1/N$, we depart from the ITR_W definition by explicitly negating the measure in the (presumably rare) cases in which $P < P_0$: we find this to be more consistent with the assumptions that implicitly underlie the preference for ITR_W over channel capacity. Specifically, ITR_W and RIG_B are best regarded as measuring performance given two assumptions: that targets cannot be arbitrarily recoded, and that errors have already been corrected to whatever extent is possible. By contrast, channel capacity takes account of potential future recoding and error correction.

In the Supplementary Material, we present a more detailed discussion of the relationship between RIG_B and other performance measures. In Section S4.1, we contrast it with approaches based on Fitts' Law. In Section S4.2, we argue that, in the context of BCI, it is more appropriate to exclude potential recoding and error-correction, and hence to prefer RIG_B or ITR_W over channel capacity. We also explain the ways in which our approach differs from ITR_W, adapting ITR_W to cope with cases in which its default assumptions are inappropriate.

Relative to the classic definition of ITR_W, the main advantage of our formulation of RIG_B for the purposes of the current study is that we no longer rely on the assumption of N exhaustive, mutually-exclusive, equiprobable target classes—an assumption that is often not met, for example in most continuous control tasks. Instead, we require only a method for estimating the rate at which desired targets are hit by chance, according to some model that embodies the null hypothesis of no voluntary control, under conditions that match those experienced by the user. In Section 2.8, we develop such a method.

2.8. Estimating Chance Performance

The performance metric defined above in Section 2.7 requires an estimate of P_0, the success rate we expect by chance under the null hypothesis that the user has no voluntary control over the BCI system. To borrow terminology from Fitts' Law analysis, if equation (1) can be seen as an **index of performance**, the corresponding **index of difficulty** quantifying the difficulty of a given trial in bits, could then be defined as $-\log_2 P_0$. The chance probability estimate may also be used in other performance metrics, such as Cohen's κ [25]—see Billinger et al. [15] for discussion of the importance of taking chance levels into account when reporting BCI performance.

There are multiple ways of computing such a chance-performance estimate. In fact, we already have one route for doing so, in the Random Baseline condition described in Section 2.3. This method relied on replaying the recorded EEG signal through the same online BCI software system that the subjects used to play the game in the other controller conditions. 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 often cannot be replicated offline, cannot be performed quickly, 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 so that it can support a reliable replay analysis. By contrast, we wished to develop a general re-usable offline method that is applicable to a wide variety of control scenarios—one that 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.

Our solution was to re-simulate each trial repeatedly using a random-walk method. Although the current study only used one-dimensional control, we will 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 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 defined 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 was simulated S times, and the success rates for all trials within the same scope were 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.

Each simulated trial began with the same initial conditions (cursor position and width) as the corresponding trial of the original data-set. It contained barriers (in our case, only the edges of the screen prevented the cursor from moving) and targets (objects that the cursor must either hit or avoid) which occupied the same positions in space and time that they occupied in the original trial. We then generated a series of normally-distributed random step vectors. We smoothed the time-series of steps so that it had the same auto-correlation (at a lag of one time-step) as the trials in the original scope. We also scaled it so that its variance matched that of the steps in the trials in the original scope (in higher-dimensional tasks, we would match the *co*-variance). In this way, we match the smoothness, distribution of sizes and distribution of directions of the cursor trajectories actually produced by the user. The random steps were then integrated numerically to form a simulated trajectory under the constraint that the cursor may not pass through barriers. Each simulated trajectory was then assessed to determine whether it collided with a target, and the simulation was scored as a success or failure accordingly. Our estimate for P_0 was the proportion of successes during all simulations of a given scope.

We describe and discuss the approach in greater detail in Section S1 of the Supplementary Material, and provide a Python implementation on our website, at http://schalklab.org/ downloads.

3. Results

In this section, we report the results of our experiment in three parts, to address the three respective challenges outlined in the Introduction. Section 3.1 examines the repeatability and consistency of the adaptive staircase procedure we designed to address Challenge I. The staircase procedure's mid-run estimates are also used to examine the extent to which our subjects' performance improved significantly over time. Section 3.2 validates and examines the information gain metric that we developed to address Challenge II, and the random-walk simulation method on which it relies. The information gain results are shown to agree very closely with the mid-run estimates of the staircase procedure, despite the very different origins of these two performance measures. Finally, in answer to Challenge III, Section 3.3 uses the information gain measure to quantify the false performance ceiling imposed by our BCI signal processing pipeline, i.e. the difference in performance between the Direct Controller and the Pseudo-BCI Controller, which reflects the negative impact that the signal processing pipeline has on high-end performance.

3.1. Challenge I: Build an Efficient Adaptive Performance Measurement System

In the Introduction, we described Challenge I as the need to develop a measurement scale, and an efficient measurement procedure, that allow us to measure performance automatically and adaptively both at very high levels (close to the performance of the human motor system, and perhaps beyond) and also at very low levels (random chance performance).

Our adaptive performance measurements consisted of 4 subjects $\times$ 10 sessions $\times$ 4 controller conditions $\times$ 3 repetitions per session. The results are shown in Figure 2. Performance is plotted for each of the four subjects, in each of the four controller conditions introduced in Section 2.3. As explained in Section 2.5, the measure of performance MRE_d is the output of the performance estimation procedure that is built into our adaptive staircase method: specifically, it is the mid-run estimate (MRE) of our unit-less task difficulty variable d. Each data-point is the MRE from one adjustment phase: across all subjects and all active controller conditions, measurement of such a value took an average of 59 seconds, and one such measurement was performed approximately every 4 minutes.

It is clear from Figure 2 that performance in the Direct Controller condition is consistently better than performance in Pseudo-BCI: if we perform an unpaired two-tailed t-test for each subject, we obtain $p < 4 \times 10^{-7}$ for every subject. Pseudo-BCI, in turn, is consistently better than BCI (similarly, $p < 2 \times 10^{-14}$ for every subject). BCI performance is better than chance for only two of the four subjects, namely subjects B ($p = 4 \times 10^{-6}$) and C ($p = 4 \times 10^{-11}$). Note that this is to be expected in cases where subjects are not pre-selected according to their ability to use BCI or according to their resting sensory-motor rhythm amplitude, since it has been been observed that a substantial proportion of people cannot control sensory-motor rhythm BCIs (see for example [26]).

Though the BCI performance of subjects B & C is above chance overall, their performance approaches the random baseline in some individual sessions. Relative to the other controller conditions, there is large session-to-session variability in BCI control. Large session-to-

session variability has been observed in many other BCI studies (see for example Figure 2 of [10]). It seems likely that the observed variation in performance is the result of the intrinsic day-to-day variability of the EEG signals, rather than any property of the measurement procedure (for further analysis, see the Supplementary Material, Section S3).

To assess the extent to which performance improves over time, we computed a Spearman correlation coefficient between session number and MRE_d for each subject in each controller condition. The results are shown in Table 1. All subjects improved their Direct Controller performance significantly over the course of 10 sessions. There was a significant improvement in Pseudo-BCI performance for subjects A and B, but not for subjects C and D. There is a significant increasing trend in BCI performance for subject A, but also for the Random Baseline data of subject D. This illustrates the importance of considering factors other than voluntary control when interpreting performance results that are close to random. In subject D's Random Baseline data, we found a significant increasing trend in the mean squared step size over the course of 10 sessions, which may explain the increase in performance: random movements may be more or less successful in a particular task depending on simple parameters like their amplitude. Furthermore, it is clear that at low performance levels, d may vary according to factors other than the user's degree of voluntary control, as evidenced by the increasing trend in random baseline performance for two of our subjects. Therefore, a performance estimate from an active control condition must always be assessed relative to an estimate from the corresponding random baseline condition. For these reasons, in the random-walk simulations described in Section 2.8, we match parameters of the random walk to those of the original data.

Figure 2 also allows the following observations about the measurement approach itself:

- Our system successfully captured both very high performance (the Direct Controller) as well as chance performance (Random Baseline), while still being able to distinguish beginners' BCI performance from chance (BCI Controller for Subjects B and C). These could all be represented on the same scale, and were measured in a single task without any manual adjustment of parameters by the experimenter.

- Even at the high-performance end of the scale, our system was able to track and quantify improvements in the subjects' performance as a function of time. Thus, even on the scale that successfully registered the very large difference between the subjects' BCI performance and their Direct Controller performance in the early sessions, the measurement still had not hit a ceiling, and still retained a useful degree of precision for measuring further improvements.

- The measurements were repeatable: generally, the within-session spread (of the three MRE_d values per session) was small relative to the differences between controller conditions. (As an illustration of this, suppose that an experimenter had only one session's data available, and wished to establish whether there was a significant impact of the signal-processing chain on control performance. The experimenter might use a single two-tailed two-sample t-test based on the session's 3 one-minute adjustment-phase measurements in the Direct Controller condition

and 3 in the Pesudo-BCI condition. Despite the small data-set sizes, the test would distinguish the two conditions at the 5% significance level on 25 out of the 37 sessions in our data-set.) The within-session variability was also small relative to the session-to-session performance variations we saw in both the BCI Controller condition (likely due to variability in the EEG signal quality) and the Direct Controller condition (largely due to improvement as a result of practice).

Therefore, we conclude that an adaptive staircase approach is an efficient and effective way of measuring control performance in BCI, and of tracking changes in performance over time. The mid-run estimates provided by the staircase method are repeatable and reliable. The key to enabling measurements to capture both high and low performance automatically, without the intervention of the experimenter to change task parameters, is to ensure that task difficulty d is univariate, and to compute mid-run estimates on the scale of d. This allows us to compare performance across subjects, sessions and controller conditions.

However, since the units of d are arbitrary, and unique to the configuration of the other (fixed) task parameters, comparisons become invalid as soon as there is a change in any of the game mechanics or other contextual variables. For this reason, we developed the method explained in Sections 2.7 and 2.8, which we validate in the following section.

3.2. Challenge II: Develop a Transferable Performance Metric

In the Introduction, we described Challenge II as the need to develop a transferable performance metric that could allow comparison of performance between different tasks. In this section, we examine and validate the results of the methods introduced in Sections 2.7 and 2.8 to address this challenge.

Figure 3 shows estimates of the subjects' performance expressed as rates of information gain, RIG_B, computed using equation (1). Each data-point is based on the combined trials from the three adaptive staircases performed in one controller condition during one session. Chance-level performance P_0 was estimated using 1,000 random-walk simulations per trial as described in Section 2.8. Most of the patterns and trends we see in Figure 3 are very similar to those of the MRE_d results in Figure 2. One notable difference is that the Random Baseline, and the BCI performance of subjects A and D, now appears flat and very close to 0.

Figure 4 examines in greater detail the relationship between the original task-specific performance measure MRE_d and the other more general measures: success probabilities P and P_0 in panels (a) and (b), respectively, and information gain rates in bits per trial and bits per minute in panels (c) and (d) respectively. We can see that MRE_d is highly consistent with the information gain measures, with a Spearman rank correlation coefficient of 0.97 between MRE_d and bits per trial, and 0.98 between MRE_d and bits per minute.[§]

[§]The slight difference between bits per trial and bits per minute, and the motivation for examining them separately, arises because the speed of the targets, and hence the rate at which they can be hit, is varied as a way of varying difficulty and keeping pace with the subject's ability level (see Section 2.5). Thus, game difficulty may affect RIG_B in equation (1) through both the numerator (by affecting P_0) and the denominator (t).

The information gain measures agree so well with MRE_d that it is worth pointing out that their similarity was not inevitable *a priori*—they are not merely transformations of each other. MRE_d is the result of a heuristic designed to estimate performance rapidly—specifically, the weighted up-down staircase procedure. The heuristic's output depends not only on the relative proportion of hits and misses and the difficulty levels at which they occur, but also on the serial order in which they occur. It is therefore affected not only by binomial variability, but by the accuracy with which the heuristic converges on the desired success rate of 65%.[||] This is in contrast to the information-gain measures of performance: while they benefit from the fact that the staircase procedure kept the difficulty level away from the performance ceiling—as panel (a) also confirms—they do not rely on the task difficulty variable, nor on the order in which the adaptive steps occurred. They do, however, rely on the estimation of P_0 by random-walk simulation, which MRE_d estimates do not. Due to the differences in their origin, the very high degree of agreement between MRE_d and RIG_B is an encouraging indicator of their validity as performance metrics.

We should note that there are other ways besides equation (1) to express P relative to P_0. We would also expect other such measures to exhibit good validity. A well-known example of such a statistic is Cohen's κ coefficient [15, 25], defined as $\kappa = (P - P_0)/(1 - P_0)$. This statistic also agrees very well with information gain: the Spearman correlation between κ and information gain was 0.98 in our current data-set. The Spearman correlation between MRE_d and κ was 0.95, very close to the value of 0.97 we observed between MRE_d and information gain in bits per trial.

A further desirable property of RIG_B is demonstrated in Figure 5. Ideally, we would like our measure to reflect the capabilities of the BCI user and the BCI system, but in a way that is invariant of the difficulty of the task they are performing.[¶] To test this property, we separated the trials of each session into two groups according to the task difficulty value d at which they were performed. We computed P and RIG_B separately for the easier half and the harder half of the trials of each session. The left panel of Figure 5 confirms that this separation according to d values has the expected effect on the success rate P: the data-points lie predominantly below the diagonal line of equality, indicating that the success rate is lower on trials that were designed to be harder. In the right panel, however, the corresponding RIG_B values are distributed equally on both sides of the line of equality, indicating that the information gain rates measured by our system were similar regardless of whether they were measured on easier or on harder trials. It is interesting to note that our task does not elicit higher information transfer at greater task difficulty levels: this is an encouraging sign that the subjects are unlikely to have been "coasting" or "slacking" during easier trials.

[||]Note also that, when we consider *all* the trials performed during the course of each staircase, the procedure does not succeed perfectly in making the subjects' success rate independent of d, as is clear from panel (a) of Figure 4.

[¶]A bit rate computed by Fitts' Law analysis typically exhibits this property. It is computed from the gradient of a line relating trial duration to task difficulty. A straight line usually fits empirical data very well, indicating that the bit rate is the same whether one looks exclusively at easier or exclusively at harder trials. See the Supplementary Material, Section S4.1 for further discussion of Fitts' Law analysis.

3.3. Challenge III: Measure Limitations in BCI Performance

In the Introduction, we described Challenge III as the need to assess the extent to which BCI system components (such as the BCI signal processing pipeline) not only enable BCI performance, but also potentially restrict the maximum level to which performance might be expected to rise as the user learns to use the BCI system more effectively.

The information gain results are summarized in Table 2. Of particular interest is the false performance ceiling, which is the difference between the Direct Controller and the Pseudo-BCI Controller conditions. This reflects the extent to which the EEG signal processing pipeline restricts the maximum control performance that can be achieved under our chosen constraints. The false ceiling is marked by the yellow shaded region in Figures 2 and 3. The average difference across all four subjects was 21 bits per minute. From this, we conclude that the signal processing pipeline imposed a performance ceiling at least 21 bits per minute below the maximum that could be achieved in this task. This a large and unexpected decrease, as it lowers the maximum bit rate by 33% on average.

4. Discussion

In this study, we developed novel approaches for measuring performance in BCI control tasks. We identified three challenges: Challenge I was to address the need for efficient measurement techniques that could adapt rapidly and reliably to capture a very wide range of performance levels; Challenge II was to express performance results in task-independent units that could allow comparison across a wide range of tasks; Challenge III was to measure the extent to which certain components of a BCI system (for example, the signal processing pipeline) not only enable good performance, but also potentially limit the maximum level we can expect performance to reach.

Our experiments with healthy human subjects confirmed that our approach can provide efficient performance measures on a scale that captured both beginners' performance in a non-invasive EEG BCI and the much higher levels of performance supported by conventional human-computer interface hardware in the same task. (We assume that the latter is much closer to the performance required for real-world tasks.) Our approach consisted of three separate but complementary strategies: the first addressed experimental task design in answer to Challenge I; the second addressed data analysis in answer to Challenge II; and the third took advantage of the combined power of the first two to address Challenge III.

The task-design strategy was to use an adaptive staircase method coupled to a single variable that automatically and monotonically varied the difficulty of the task. Our task was a computer game in which the player had to move a cursor in one dimension to catch falling targets. We used Kaernbach's well-known weighted up-down staircase method [12], and found that it produced reliable and repeatable results efficiently, allowing us to assess differences in performance between subjects, between sessions, and between controller types, as well as improvements in performance due to learning. The staircase procedures themselves took approximately a quarter of the experimental time, indicating that it is feasible to combine this assessment method with other experimental designs. Note that there

are many staircase methods, of varying sophistication, efficiency and robustness—see Leek [27] for a review. We chose Kaernbach's procedure for its simplicity and flexibility, but it is possible that other more sophisticated staircase procedures might produce even better results.

The data analysis strategy was based on success rates (the proportion of successes in a number of discrete trials). This is in contrast to popular approaches based on Fitts' Law analysis (FLA).We avoided FLA due to its limitations—its assumption of negligible rates of failure, its applicability only to tasks in which speed can be traded for accuracy, and its lack of invariance to nuisance parameters (for a more detailed discussion of these points, see Section S4.1 in the Supplementary Material). Instead, we took advantage of the fact that the staircase procedure automatically kept the subject's success rate P below ceiling, even across a very wide range of performance conditions. The subject's success rate was assessed relative to the success rate P_0 that might be expected by chance, under the null hypothesis of no voluntary control. Chance performance was estimated by a random-walk model in which the random steps' size and smoothness in time, as well as the difficulty levels of the trials, were matched to the subject's original input. The two success rates may be combined in a number of ways to obtain performance metrics that should allow comparison between similar but non-identical tasks. One such approach might be to use Cohen's κ [15, 25]. We chose to use a formula for the rate of information gain between two Bernoulli distributions, which we denote RIG_B, yielding a result in bits per minute or bits per trial. This measure is closely related to the information transfer rate ITR_W previously proposed by Wolpaw et al. [16], but we adapted it in two ways: first, we removed the reliance on an integer number of discrete equiprobable task outcomes; and second, we introduced a sign term to make the measure more consistent with the implicit assumptions that distinguish ITR_W from channel capacity measures. For more detailed discussion of this and other aspects of ITR_W, see Section S4.2 in the Supplementary Material.

Since we found it was possible to obtain reliable results from just three one-minute measurement phases per session, it is conceivable that an adaptive assessment system might be incorporated into a BCI system deployed for real-world usage, as a way of monitoring the user's progress. Adaptive assessment would necessarily be carried out as a brief regular exercise separate from ordinary day-to-day BCI usage, since in day-to-day usage there would be no sense in artificially making the user's tasks more difficult than they needed to be. With or without the adaptive staircase, information gain analysis could also be applied to monitor performance in the field. This would also be easier in the context of a structured, perhaps somewhat artificial exercise. It is conceivable that RIG_B could be used to assess performance during actual day-to-day usage, but here its limitations become apparent. First, discrete "trials" must be identified somehow, and each trial must be categorized unequivocally as either successful or unsuccessful. Second, the environment and constraints under which the tasks are performed must be detected and modeled with sufficient accuracy to allow valid random-walk simulations. Third, the results are only comparable across tasks of sufficient similarity: one cannot compare information gain between Bernoulli and non-Bernoulli tasks, for example, nor would it be meaningful to compare tasks with very

different goals (comparing a movement-control bit rate with an item-selection bit rate, for example).

Either the task-design strategy or the data-analysis strategy can be applied alone, but they are particularly powerful in combination, and open the door to exploring some of the important long-term questions for the BCI field. For example, in the current study, we illustrated how the combined approaches can be used in conjunction with a contrastive experimental design to quantify a **false performance ceiling**. By this, we mean the difference between performance under unavoidable constraints (those imposed by the task itself, and those imposed by the capacities of our normal motor output pathways) and performance under the same constraints *when a particular necessary BCI component or algorithm is in use*. This reflects the extent to which the component or algorithm in question restricts the maximum control performance that can be achieved. Note that "maximum" in the context of any one particular experiment is always defined relative to the constraints we choose to accept. In this study we chose to limit ourselves to control methods that contrasted total left-hand activity against total right-hand activity, using either motor imagery or the shaking of two Nintendo Wii remotes. If we had allowed our subjects to use conventional computer game controllers in the same task, their "maximum" performance would presumably have been even higher.

Our study examined the false ceiling imposed by the artificial signal processing pipeline that is necessary to extract BCI control signals from non-invasive EEG measurements of sensory-motor rhythm modulation. We demonstrated that this can be assessed by using our combined task-design and data-analysis approach to measure the performance difference between a Direct Controller (i.e. a non-BCI input system that is engineered to maximize performance) and the corresponding Pseudo-BCI Controller (i.e. the same input device that is used by the Direct Controller, but interfaced with the same signal processing pipeline that is used in BCI). In our one-dimensional control task, the average size of the false ceiling imposed by the signal processing pipeline was 21 bits per minute (0.3 bits per trial), a 33% reduction in bit rate. Furthermore, while all four subjects showed a significant improvement in Direct-Controller performance over the course of the study, two of the four subjects did not significantly improve their Pseudo-BCI performance. This raises the question of whether the signal processing pipeline imposed an absolute limit that BCI performance could never be expected to exceed, even with an arbitrarily large amount of practice.

We believe that in future, such techniques will be vital for evaluating components of a BCI system, whether they are hardware or (like the signal processing algorithms we examined) software. Each algorithm, component or set of components should be evaluated not only in terms of what it enables us to do, but also in terms of the limits it may impose on performance. We believe that this approach will be critical for achieving the substantial performance improvements that will be necessary if neuroprosthetic devices are to meet the demands of real-world tasks.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. Clinical Neurophysiology. 2002; 113(6):767–791. [PubMed: 12048038]

2. Wolpaw, JR.; Wolpaw, EW. Brain-Computer Interfaces: Principles and Practice. 1st edition. Oxford, UK: Oxford University Press; 2012.

3. Hochberg LR, Serruya MD, Friehs GM, Mukand Ja, Saleh M, Caplan AH, Branner A, Chen D, Penn RD, Donoghue JP. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature. 2006; 442(7099):164–171. [PubMed: 16838014]

4. Velliste M, Perel S, Spalding MC, Whitford AS, Schwartz AB. Cortical control of a prosthetic arm for self-feeding. Nature. 2008; 453(7198):1098–1101. [PubMed: 18509337]

5. Hochberg LR, Bacher D, Jarosiewicz B, Masse NY, Simeral JD, Vogel J, Haddadin S, Liu J, Cash SS, van der Smagt P, Donoghue JP. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature. 2012; 485(7398):372–375. [PubMed: 22596161]

6. Schalk G, Miller KJ, Anderson NR, Wilson JA, Smyth MD, Ojemann JG, Moran DW, Wolpaw JR, Leuthardt EC. Two-dimensional movement control using electrocorticographic signals in humans. Journal of Neural Engineering. 2008; 5(1):75–84. [PubMed: 18310813]

7. Wang W, Collinger JL, Degenhart AD, Tyler-Kabara EC, Schwartz AB, Moran DW, Weber DJ, Wodlinger B, Vinjamuri RK, Ashmore RC, Kelly JW, Boninger ML. An Electrocorticographic Brain Interface in an Individual with Tetraplegia. PLoS ONE. 2013; 8(2):e55344. [PubMed: 23405137]

8. Wolpaw JR, McFarland DJ. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(51):17849–17854. [PubMed: 15585584]

9. Galán F, Nuttin M, Lew E, Ferrez PW, Vanacker G, Philips J, Millán JDR. A brain-actuated wheelchair: asynchronous and non-invasive Brain- computer interfaces for continuous control of robots. Clinical Neurophysiology. 2008; 119(9):2159–2169. [PubMed: 18621580]

10. McFarland DJ, Sarnacki WA, Wolpaw JR. Electroencephalographic (EEG) control of three-dimensional movement. Journal of Neural Engineering. 2010; 7(3):036007. [PubMed: 20460690]

11. Doud AJ, Lucas JP, Pisansky MT, He B. Continuous Three- Dimensional Control of a Virtual Helicopter Using a Motor Imagery Based Brain- Computer Interface. PLoS ONE. 2011; 6(10):e26322. [PubMed: 22046274]

12. Kaernbach C. Simple adaptive testing with the weighted up-down method. Perception and Psychophysics. 1991; 49(3):227–229. [PubMed: 2011460]

13. Thomas E, Dyson M, Clerc M. An analysis of performance evaluation for motor-imagery based BCI. Journal of Neural Engineering. 2013; 10:031001. [PubMed: 23639955]

14. Schlögl, A.; Kronegg, J.; Huggins, JE.; Mason, SG. Evaluation Criteria for BCI Research. In: Dornhege, G.; Millán, JdR; Hinterberger, T.; McFarland, DJ.; Müller, K-R., editors. Toward Brain-Computer Interfacing. Cambridge, MA: MIT Press; 2007. p. 327-342.

15. Billinger, M.; Daly, I.; Kaiser, V.; Jin, J.; Allison, BZ.; Müller-Putz, GR.; Brunner, C. Is it significant? Guidelines for reporting BCI performance. In: Allison, BZ.; Dunne, S.; Leeb, R.; Millán, JdR; Nijholt, A., editors. Towards Practical Brain- Computer Interfaces. Springer; 2013. p. 333-354.

16. Wolpaw JR, Ramoser H, McFarland DJ, Pfurtscheller G. EEG-based communication: improved accuracy by response verification. IEEE Transactions on Rehabilitation Engineering. 1998; 6(3): 326–333. [PubMed: 9749910]

17. Leeb R, Friedman D, Müller-Putz GR, Scherer R, Slater M, Pfurtscheller G. Self-Paced (Asynchronous) BCI Control of a Wheelchair in Virtual Environments : A Case Study with a Tetraplegic. Computational Intelligence and Neuroscience. 2007; 2007:79642. [PubMed: 18368142]

18. King CE, Wang PT, Chui LA, Do AH, Nenadic Z. Operation of a brain-computer interface walking simulator for individuals with spinal cord injury. Journal of Neuroengineering and Rehabilitation. 2013; 10:77. [PubMed: 23866985]

19. Sharbrough F, Chatrian GE, Lesser PR, Lüders H, Nuwer M, Picton TW. American Electroencephalographic Society guidelines for standard electrode position nomenclature. Journal of Clinical Neurophysiology. 1991; 8(2):200–202. [PubMed: 2050819]

20. Schalk G, McFarland D, Hinterberger T, Birbaumer N, Wolpaw J. BCI2000: a general-purpose brain-computer interface (BCI) system. IEEE Transactions on Biomedical Engineering. 2004; 51(6):1034–1043. [PubMed: 15188875]

21. Schalk, G.; Mellinger, J. A Practical Guide to Brain-Computer Interfacing with BCI2000. Springer-Verlag London Limited; 2010.

22. Hill NJ, Schreiner T, Puzicha C, Farquhar J. BCPy2000. 2007 http://bci2000.org/downloads/ BCPy2000.

23. Burnham, KP.; Anderson, DR. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. New York, NY USA: Springer; 2002.

24. Mackay, DJC. Information Theory , Inference , and Learning Algorithms. Cambridge, UK: Cambridge University Press; 2003.

25. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20(1):37–46.

26. Blankertz B, Sannelli C, Halder S, Hammer EM, Kübler A, Müller K-R, Curio G, Dickhaus T. Neurophysiological predictor of SMR-based BCI performance. NeuroImage. 2010; 51(4):1303– 1309. [PubMed: 20303409]

27. Leek MR. Adaptive procedures in psychophysical research. Perception & psychophysics. 2001; 63(8):1279–1292. [PubMed: 11800457]
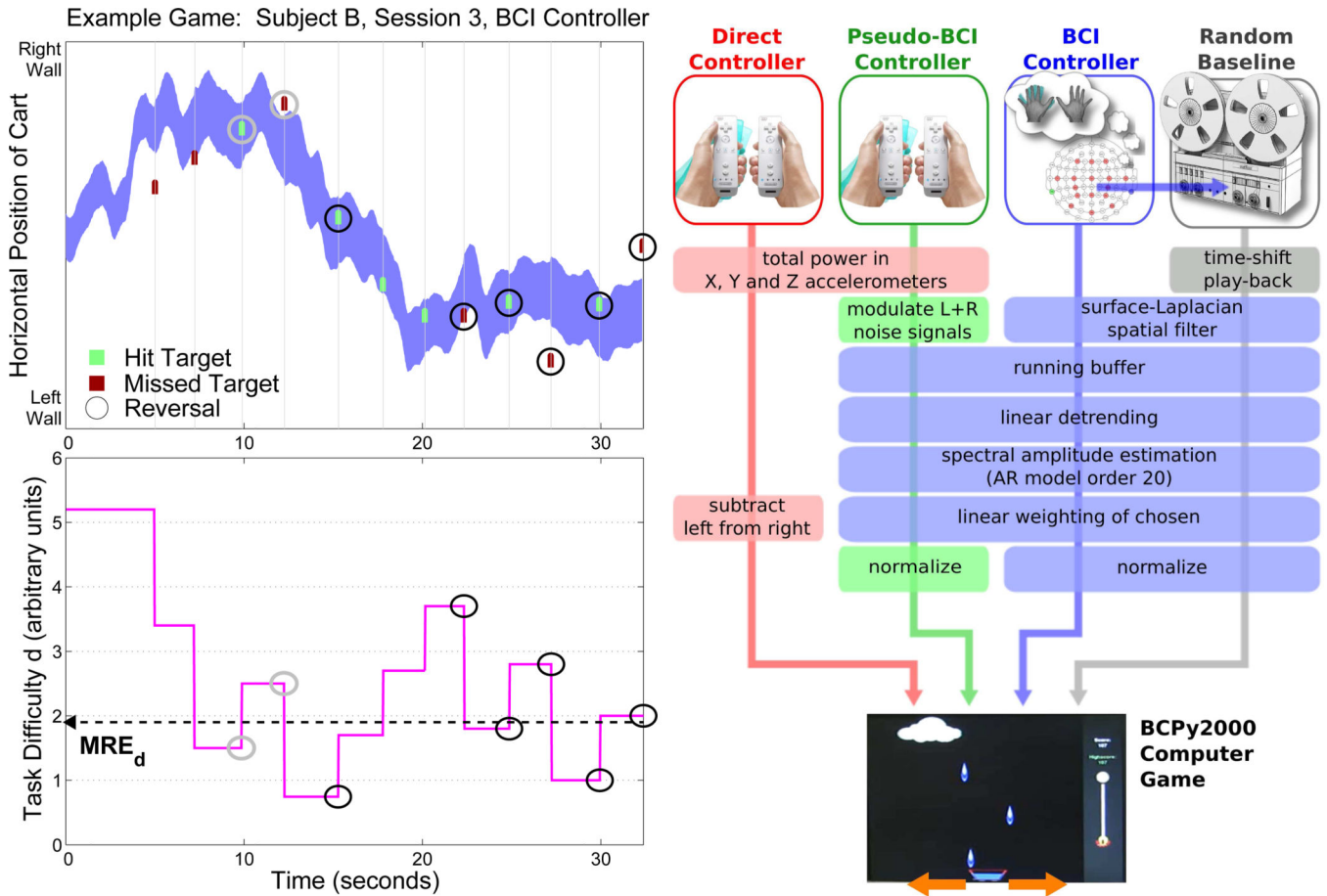
Figure 1.

The panels on the left show the time course of the adjustment phase of an example game, plotting cursor and target positions (upper panel) and the task difficulty variable d (lower panel) as a function of time. In the upper panel, the thick wavy blue line indicates the portion of the screen occupied by the cursor at each time point. The vertical gray lines indicate the times at which trials end. The small rectangular patches indicate the spatial extent of the targets (raindrops) and the temporal window during which the cursor (cart) had to catch them. Light green patches indicate targets that were hit, whereas dark red patches indicate targets that were missed. In both panels, circles highlight the reversals, i.e. hits that followed misses and misses that followed hits. The first two reversals (gray circles) were discarded, and the subsequent six (black circles) were used to compute the mid-run estimate of task difficulty (MRE_d) that was recorded as a measure of performance. Step changes in the task difficulty variable d, triggered by the hits and misses, are shown in pink in the lower panel. The MRE_d is indicated by the triangle and dotted line: it is equal to the median of the d values at the last six reversals.

The panel on the right schematically illustrates the four controller conditions used in our experiments—see Section 2.3 for a full description.
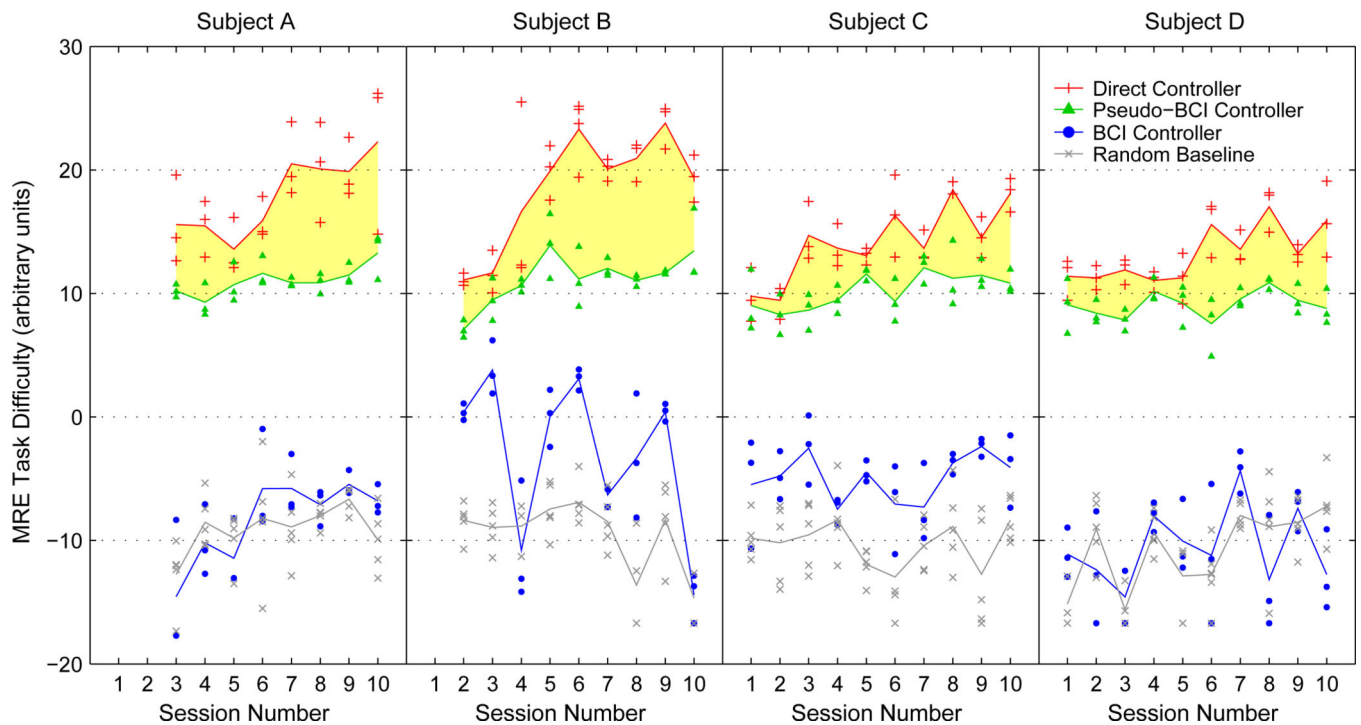
Figure 2.
Performance levels are plotted as a function of number of sessions, for each subject (panels left to right), in each of the four controller conditions (different symbol shapes/colors). Each point marks the mid-run estimate of task difficulty (MRE_d) obtained during the adjustment phase at the end of one game cycle. Solid lines show the session means. The yellow shaded area marks the "false performance ceiling". We discarded Subject A's first two sessions and Subject B's first session, because they were recorded before slight changes to the game framework introduced a change to the difficulty scale.
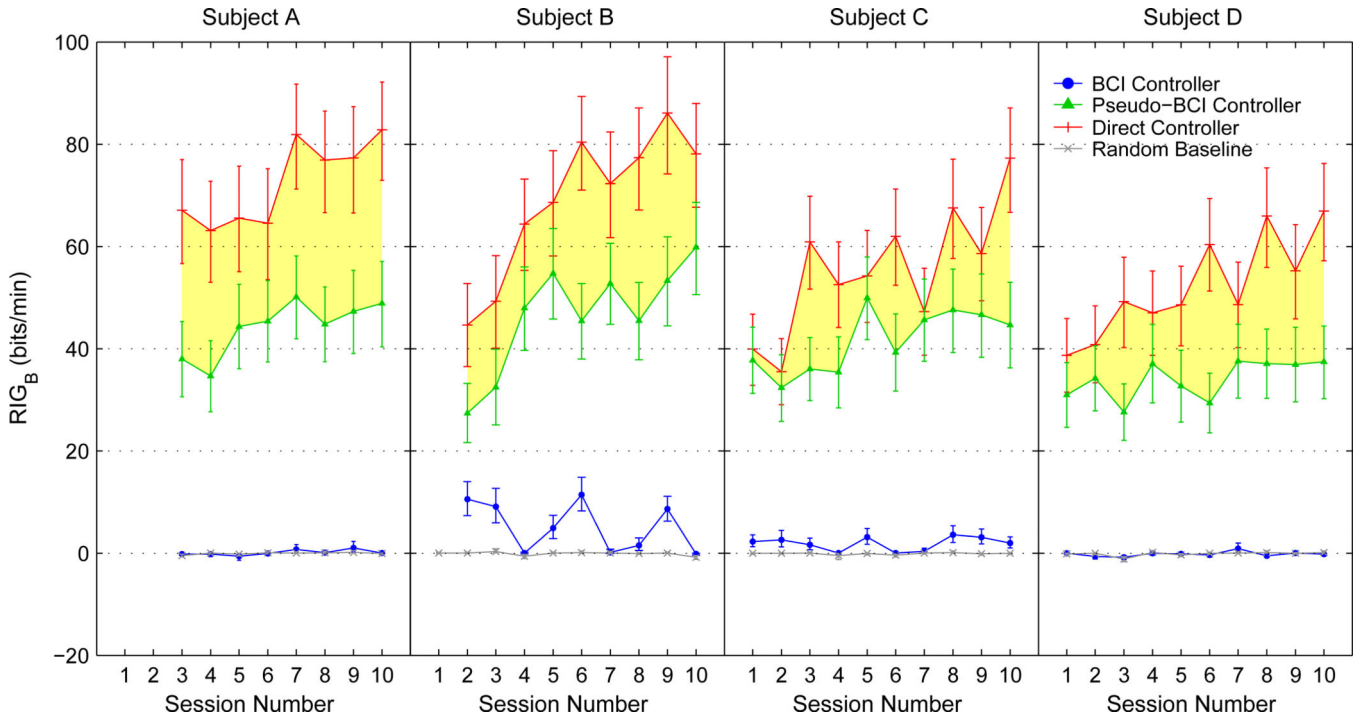
Figure 3.
Performance levels are plotted as a function of number of sessions, for each subject (panels left to right), in each of the four controller conditions (different symbol shapes/colors). Each point marks the rate of information gain (RIG$_B$) in bits per minute, computed over the all trials in the 3 adjustment phases performed during a given session. Error bars are equal-tailed 68.3% confidence intervals (the same coverage as the mean ±1 standard error for normal variables) computed as described in the Supplementary Material, Section S2. The yellow shaded area marks the "false performance ceiling."
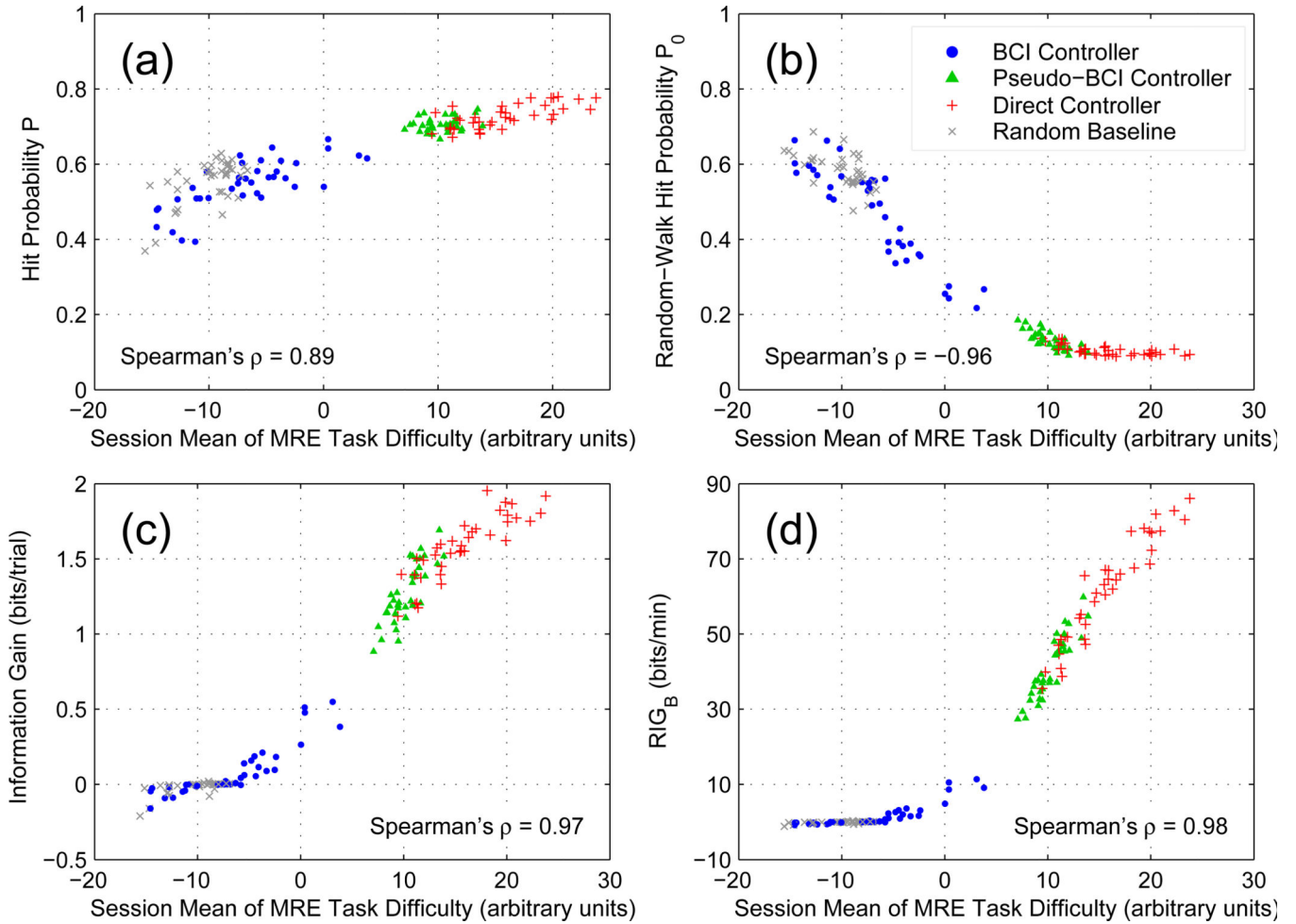
Figure 4.
This figure shows the relationship between our initial measurement of performance, MRE_d expressed in arbitrary units on a highly task-specific scale, and the re-computed, more general metrics based on information gain. Each panel shows a scatter-plot of performance values for all four subjects in all four controller conditions, each point denoting one session. The four controller conditions are distinguished by different symbol shapes and colors, according to the legend at the top right of panel (b). In all panels, the horizontal axis measures MRE_d, averaged across the 3 separate measurements of each session. The four panels illustrate the relationship between this measure and **(a)** the subject's average success rate P, across all trials of a given session, the number of which varied from 45 to 129; **(b)** the probability P_0 of succeeding according to the null hypothesis of no voluntary control, estimated using 1000 random-walk simulations of each trial in a given session; **(c)** information gain in bits per trial, obtained using the estimates P and P_0 in equation (1), with $\bar{t} = 1$ trial; **(d)** the rate of information gain in bits per minute, obtained using the estimates P and P_0 in equation (1) with $\bar{t}$ equal to the average trial duration in minutes for each session. At the bottom of each panel, we give the Spearman rank correlation coefficient ρ between each respective measure and MRE_d.
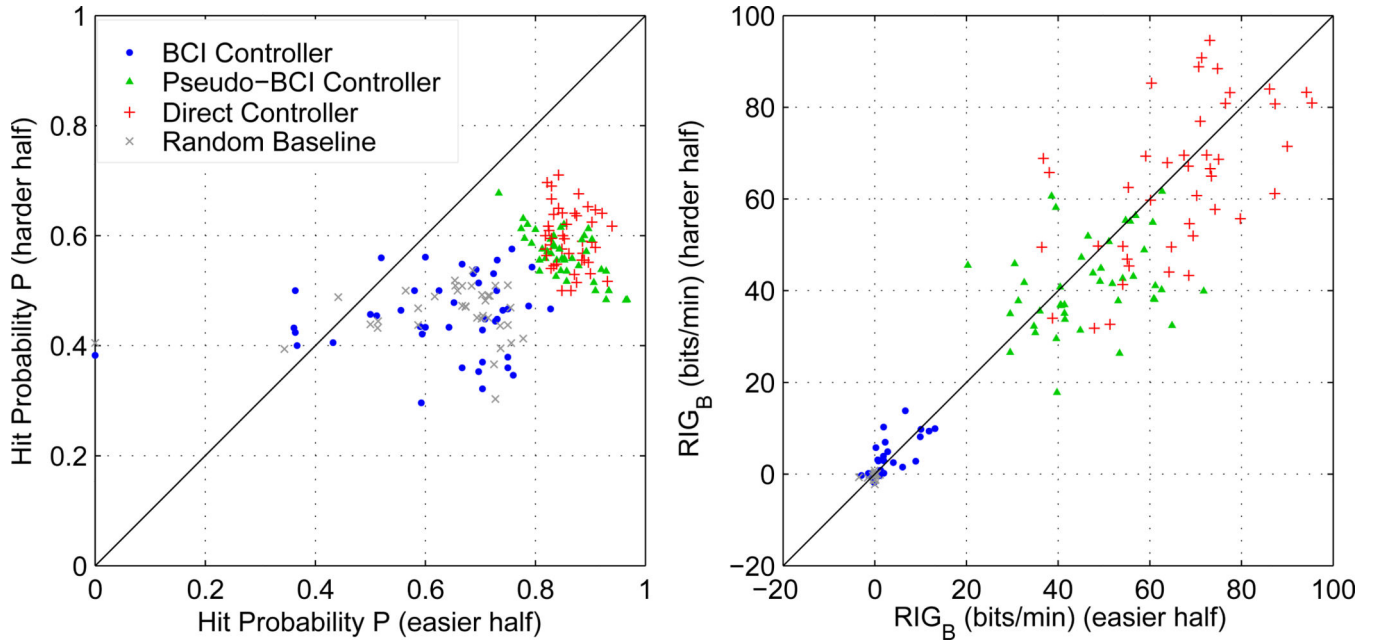
Figure 5.
This figure illustrates that the rate of information gain RIG_B was to some extent invariant of the difficulty level of the task. Success rates P (left panel) and information gain rates RIG_B (right panel) were computed separately for the easier half and harder half of the trials in each session. Each panel is a scatterplot of all four subjects' data in all four controller conditions and all sessions: one point represents the result from one subject in one session in one condition. Different symbol shapes/colors represent the different controller conditions, as indicated in the legend on the left. Performance on the easier half is plotted on the horizontal axes, and performance on the more difficult half is plotted on the vertical axes. The diagonal lines are lines of equality.

Table 1

For each controller condition and each subject, this table shows the degree of improvement in performance, as measured by the Spearman correlation coefficient ρ between session number and MRE_d. Significance probabilities are given in parentheses after each correlation coefficient.

	Subject A	Subject B	Subject C	Subject D
Direct	0.58 ($p = 3.1 \times 10^{-3}$) *	0.55 ($p = 2.6 \times 10^{-3}$) *	0.71 ($p = 9.1 \times 10^{-6}$) *	0.71 ($p = 9.4 \times 10^{-6}$) *
Pseudo-BCI	0.64 ($p = 8.4 \times 10^{-4}$) *	0.65 ($p = 2.4 \times 10^{-4}$) *	0.52 ($p = 3.5 \times 10^{-3}$)	0.18 ($p = 0.34$)
BCI	0.66 ($p = 4.6 \times 10^{-4}$) *	-0.40 ($p = 0.039$)	0.17 ($p = 0.37$)	0.14 ($p = 0.47$)
Baseline	0.35 ($p = 0.027$)	-0.22 ($p = 0.17$)	-0.02 ($p = 0.90$)	0.49 ($p = 6.0 \times 10^{-4}$) *

A star indicates significant correlations at the significance threshold of 0.05, with Bonferroni correction for the fact that we are testing 16 hypotheses simultaneously (so, p 3.125×10^{-3}).

Table 2

For each controller condition (first four rows) and each subject (first four columns), this table shows the mean and standard error across sessions of the estimated bit-rates RIGB, in bits per minute (black rows) and bits per trial (red italic rows). The last column shows the mean and standard error of each row (i.e. the mean and standard error across subjects of the per-subject means). The bottom row is similar, but the statistic of interest is the **false performance ceiling**, i.e. the session-by-session difference in bit rate between the Direct Controller and Pseudo-BCI Controller conditions.

	Subject A	Subject B	Subject C	Subject D	Average
Direct (bits per minute)	72.4 ± 2.89	69.0 ± 4.70	55.6 ± 3.97	52.1 ± 3.08	62.3 ± 4.96
(bits per trial)	*1.70 ± 0.050*	*1.66 ± 0.078*	*1.52 ± 0.071*	*1.47 ± 0.058*	*1.59 ± 0.054*
Pseudo-BCI (bits per minute)	44.2 ± 1.88	46.6 ± 3.53	41.5 ± 1.92	34.1 ± 1.18	41.6 ± 2.71
(bits per trial)	*1.30 ± 0.064*	*1.38 ± 0.093*	*1.32 ± 0.043*	*1.14 ± 0.028*	*1.28 ± 0.051*
BCI (bits per minute)	0.1 ± 0.18	5.1 ± 1.62	1.9 ± 0.42	−0.2 ± 0.15	1.7 ± 1.22
(bits per trial)	*0.00 ± 0.014*	*0.25 ± 0.079*	*0.11 ± 0.025*	*−0.04 ± 0.020*	*0.08 ± 0.064*
Baseline (bits per minute)	−0.1 ± 0.07	−0.1 ± 0.11	−0.1 ± 0.06	−0.1 ± 0.12	−0.1 ± 0.02
(bits per trial)	*−0.01 ± 0.010*	*−0.02 ± 0.018*	*−0.01 ± 0.006*	*−0.03 ± 0.021*	*−0.02 ± 0.004*
False Ceiling (bits per minute)	28.2 ± 1.87	22.4 ± 2.75	14.1 ± 3.49	18.1 ± 2.95	20.7 ± 3.04
(bits per trial)	*0.40 ± 0.027*	*0.28 ± 0.042*	*0.21 ± 0.068*	*0.33 ± 0.056*	*0.30 ± 0.041*