Empirical Movement Models for Brain Computer Interfaces

Charles Matlack[1] , Howard Jay Chizeck[1], and Chet T Moritz[2]

***Abstract*—For brain-computer interfaces (BCIs) which provide the user continuous position control, there is little standardization of performance metrics or evaluative tasks. One candidate metric is Fitts's law, which has been used to describe aimed movements across a range of computer interfaces, and has recently been applied to BCI tasks. Reviewing selected studies, we identify two basic problems with Fitts's law: its predictive performance is fragile, and the estimation of 'information transfer rate' from the model is unsupported.**

Our main contribution is the adaptation and validation of an alternative model to Fitts's law in the BCI context. We show that the Shannon-Welford model outperforms Fitts's law, showing robust predictive power when target distance and width have disproportionate effects on difficulty.

Building on a prior study of the Shannon-Welford model, we show that identified model parameters offer a novel approach to quantitatively assess the role of control-display gain in speed/accuracy performance tradeoffs during brain control.

***Index Terms*—Neural engineering, Neural prosthesis, brain-computer interface (BCI), brain-machine interface (BMI), Fitts's law, Shannon-Welford, performance metric**

I. Introduction

Performance metrics are used in BCI studies to discern performance differences between algorithm and experiment design variations. Standardized and straightforward metrics exist for quantifying symbol transmission rates, but not for characterizing the movement control mediating selection tasks. The implementation of metrics for continuous movement control is not consistent, making comparison across studies difficult to impossible. Here, we focus on Fitts's law and similar empirical models predicting the duration of aimed reaching movements, and on within-study comparisons between control modalities. The use of Fitts's law and derived metrics entails an implicit, often untested, assumption that a model developed to describe stereotyped reaching movements applies to a novel and disembodied control task, in some cases performed by a non-human primate. As the state of the art of BCIs evolves, identifying the incremental contributions of individual design choices within studies and the reproducibility of results across studies will be critical for continued performance improvement.

Our goal in this paper is to evaluate the descriptive power of candidate models that can enable comparisons of BCI control across conditions and with muscle-controlled movement. Further, we seek an approach that reveals actual or potential influence of manipulated variables not of primary interest, particularly task geometry and control-display (CD) gain, on results. *Task geometry* here refers to the combination of target distance and width, and could be adapted to include, for example, the direction of required net movement. *CD gain* is defined as the ratio between the user-controlled interface signal (e.g., force) and the rate of pointer movement on the computer display. Critically important for usability, CD gain is often a nonlinear and user-tunable function in consumer computer interfaces. In the BCI context, CD gain may be subsumed in complex algorithms designed to estimate intended movements from neural signals [1], [2]. Consequently, this parameter that undeniably influences performance may be difficult to define.

To this end, we evaluate the ability of Fitts's law and the competing Shannon-Welford model, recently introduced by Shoemaker et al [3], to describe movement times observed in manual and BCI rate-controlled cursor positioning tasks performed by a macaque. Our specific hypotheses are the following:

H1: Fitts's law and the Shannon-Welford model can predict average movement times in a rate-controlled cursor task performed by a non-human primate.

H2: In at least some cases, the Shannon-Welford model will provide significantly better descriptive power than Fitts's law, accounting for its additional parameter.

The rest of this paper is organized as follows: in Section II, we review BCI task performance analyses including Fitts's law before introducing the Shannon-Welford model; we next describe a basic 1D cursor control paradigm designed to isolate differences between

[1] Electrical Engineering Department, University of Washington, Seattle, WA

[2] Departments of Rehabilitation Medicine and Physiology & Biophysics, University of Washington, Seattle, WA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2016.2584101, IEEE Transactions on Neural Systems and Rehabilitation Engineering

manual and brain control in Section III; in Section IV we show modeling results and comparative performance analyses; and finally in Section V we discuss the implications of these findings.

II. Background

A. Performance Metrics in BCI Studies

Several approaches have been taken to quantify the performance of BCI-mediated tasks in addition to Fitts's law. One of the most persistent basic metrics is to report both trials per unit time and percentage of correct trials, with the latter being more relevant during early learning [4]–[6]. Drawbacks to this method are that it does not take into account task difficulty, nor can it accommodate mulitple task geometries or types into the same measurement. More information is provided by reporting trial time histograms [2], [7].

Properties of the trajectory taken to the target provide more nuanced insights into control qualtiy changes. These quantitative properties collectively considered to assess path efficiency can include straightness, direction reversals, movement orthogonal to target vector, and path length [8]. A problem with path efficiency is that it assumes an optimal trajectory without knowledge of the cost functions affecting the subject's control strategy. For example, experimentally-observed trajectories of stereotyped computer mouse movements are neither perfectly aimed, nor without overshoot and corrective actions near the endpoint [9].

A third method is to show average trajectories (after re-orienting many trajectories to match their directions), optionally highlighting the target entry region [2], [10]. Trajectory information provides insight into differences between manual and BCI-mediated control, but averaging obscures corrective feedback strategies near the endpoint. Also, the problem of only evaluating a single task type and geometry exists here as well.

Information transfer rate (ITR) is a metric easily calculated for discrete symbol transmission, and efforts have been made to apply it to target selection tasks by equating possible task outcomes with symbols [11], [12]. More recent studies use Fitts's law to characterize performance and derive putative ITR [1], [13], or to estimate putative ITR directly under the assumption that the model applies [2]. For a survey of recent BCI studies permitting this analysis, see supplement to [2]. We review Fitts's law and its validity in quantifying ITR in the next section.

B. Fitts's Law

Fitts's law is an empirical model developed to describe the average duration of point-to-point reaching movements [14]. The original 1954 model predicts movement time T as a function of movement distance D to a target of width W,

$$T = a + b\log_2\left(\frac{D}{W}\right) = a + b\,(ID) \tag{1}$$

where a and b are free parameters, and the Index of Difficulty term (ID) is meant to capture task difficulty as a function of only the *ratio* between target distance and width.

The relationship, originally verified for 1D reciprocal tapping tasks using a stylus [14], also describes 2D point-and-click tasks using a computer mouse [9], [15] as well as other pointing devices.

The "Shannon" formulation [15] of the ID term changes the model to

$$T = a + b\log_2\left(\frac{D+W}{W}\right) \tag{2}$$

which differs in the ID term from that used in the original study. It is inspired by the Shannon-Hartley Theorem,

$$C = B\log_2\left(\frac{S+N}{N}\right) \tag{3}$$

relating channel capacity C with signal power S, noise power N, and bandwidth B [15]. However, no mathematical equivalency between the two has been established. Nonetheless, the ID term in Fitts's law is often conferred units of $bits$ [2], [13] when interpreting fitting results. Putative ITR, or information throughput, of a pointing device can then be specified in '$bits/sec$', defined either as the inverse of the b parameter ('slope inverse method') or, ignoring a, as the mean of ID/T over subjects and conditions ('mean of means method'); the latter is recommended in [16]. A recent study authored by a previous advocate of Fitts-derived ITR shows that it is inconsistent with Shannon's entropy [17].

Throughput provides an attractive overall performance measure, but reporting a single number hides vital details relevant to comparing different studies. For example, reporting complete modeling fitting results would enable comparison of predicted performance at equivalent difficulty across studies, in addition to calculation of easily interpretable confidence intervals for predicted movement times.

C. Standardization Efforts

The ISO 9241-9 standard [18] and a follow-on publication by Soukoreff et al [16] define standardized assessment tasks for computer pointing devices and prescribe the use of putative ITR, estimated from the Shannon formulation of Fitts's law, as a performance

metric. Soukoreff et al make specific recommendations for the application of Fitts's law, including using a wide range of ID values, first using linear regression to determine whether the model applies, and finally calculating putative ITR.

However, the standard calls for reporting putative ITR by averaging ID and movement times across both task conditions and subjects, rather than deriving it from fit parameters: studies reporting throughput estimated via regression are considered non-conforming to the ISO standard. This leaves the linear regression out of the metric, with the caveat that the fit intercept should be "small" [16]. Consequently, the relationship between parameter confidence intervals and the range of task conditions and subjects tested is obfuscated if not absent in published standard throughput measures, making it difficult to attribute variability to specific factors. The Soukoreff et al study concluded that throughput measures may have dependence on task design.

Intracortical BCI studies frequently use only one or two subjects, and task design can vary considerably between studies. This suggests significant value in reporting complete and subject-specific model fit parameters to enable better comparisons across studies and conditions, rather than only reporting the standard metric.

D. Success Rates and Dwell Selection

To create a task compatible with Fitts's law and also feasible without a selection command, some studies substitute the *click-to-select* behavior with *dwell-to-select* [19]. This in turn challenges the standard approach to controlling for trial success rates.

Trial success rates must be high for an empirical movement model, such as Fitts's law, to be applied to a data set, since failed trials cannot be included in the model. In tapping and point-and-click tasks, the distribution of trajectory endpoints is assumed to be normally distributed. Post-hoc calculations of effective target size can then be used to capture 96% of all endpoints [15], as originally proposed by Crossman [20]. This approach is shown to improve model accuracy by reducing the effects of subject-to-subject variation in speed/accuracy trade-off [9].

In a *dwell-to-select* task, movement trajectories continue when the cursor crosses the target boundary, but the subject's incentive changes. Consequently, the trajectory endpoint cannot be treated as the selection point in a way that easily admits similar post-hoc target width corrections.

Another important difference in the dwell-to-select paradigm is that aggressive control results in target overshoot rather than trial failure. Thus, more variability in trial times is inevitable, and success rates substantially different from 100% suggest a lack of consistent competence at the task. If a significant number of trials fail due to reaching the trial time limit, the limit can be extended so more trials can be included in analysis, and to avoid skewing the distribution of trial times.

E. Prior BCI Studies Using Fitts's Law

Two cortical BCI studies, by Simeral et al (2011) and by Gilja et al (2012), and one EEG study by Felton et al (2009), employ Fitts's law analyses [1], [2], [13]. The studies by Simeral et al and Felton et al incorporate experimental design and statistical methods to test the predictive power of Fitts's law with human subjects, while Gilja et al report the derived putative ITR metric for non-human primate performance.

Gilja et al. use a center-out dwell-to-select task with a single target distance and width, thus testing performance at a single ID [2]. To then report ITR requires assuming that Fitts's law applies and a zero intercept of the trendline, leaving b as the only free parameter.

In the study by Simeral et al., a human subject with tetraplegia performed a point-and-click task [1]. Their intracortical BCI decoded both cusor velocity and a discrete click signal from a population of recorded neurons. Because targets appeared at random distances from the cursor position and multiple target widths were used, ID's ranging from 1 to greater than 4 were tested. Thus, the data set permits a regression on ID to derive a Fitts's law trendline. Simeral et al report significant ($p < 0.001$) fits to all trial times on each of several experimental days, with R^2 values ranging from 0.31 to 0.67. Note that in this case, the model must predict every trial time, rather than the per-condition average. The data does not permit the repeated-measures statistical analyses commonly used to test for data consistent with Fitts's law in the human-computer interface (HCI) community [3], [21].

Felton et al compared EEG control with a joystick-driven rate-controlled manual task, thus matching system order between control modalities. They used repeated trials with a set of multiple discrete distances and a single width to obtain average trial times at each ID which were then fit using Fitts's law, resulting in R^2 values frequently above 0.9.

F. The Shannon-Welford Model

Although Fitts's law is widely used and shown to have excellent explanatory power over movement data, its robustness to task parameter variations is demonstrably

poor, particularly when CD gain is manipulated [3], [22]. One practice which has emerged in Fitts's law studies of HCIs is grouping data by either target width or distance into sets which can be separately fit with better results [22]. We provide an example of this in Figure 5. This leaves a troubling open question as to why the model fails in some cases.

Exploring this problem further, a recent study by Shoemaker et al [3] compares the performance of Fitts's law with variants of Welford's 2-factor model [23],

$$T = a + b_1 \log_2(D) + b_2 \log_2(W). \tag{4}$$

Welford's model extends Fitts's law under the intuition that separable gross movement and homing phases of motion contribute to the total time. In [3], the authors modify Welford's 1971 model for equivalency with the Shannon form of the index of difficulty using the substitution $D \leftarrow D + W$. (This modification is subsequently justified by empirical fitting results.) They can then algebraically combine terms to match the form of Fitts's law, yielding the Shannon-Welford model,

$$T = a + b \log_2 \left(\frac{D + W}{W^k} \right). \tag{5}$$

Here, $k = b_2/b_1$ captures the effect of W on task difficulty relative to D. Note that when $k = 1$, the model is equivalent to the Shannon form of Fitts's law (Equation 2). However, it is observed in [3] to take on values between 0.2 and 1.8. The fact that k is often close to 1 provides insight into why Fitts's law demonstrates good fit quality in many studies, yet is not robust.

Shoemaker et al [3] found their Shannon-Welford model to better describe manual pointing performance on very large displays, and over a wide range of CD gains, compared to the original and Shannon forms of Fitts's law, and to Welford's original model. They also found that the value of k increases linearly with CD gain.

III. Methods

A. Electrophysiology

In this study, we tested whether the more robust descriptive power of the Shannon-Welford model carries over to isometric manual control (MC) and BCI-mediated (BC) rate-controlled cursor tasks. We implanted a *Macaca nemestrina* with dual 96-channel microelectrode arrays (Blackrock Microsystems, Salt Lake City, UT), bilaterally in motor cortex, and connected these to a 128-channel Cerebus neural signal processor (Blackrock). Manually set time-voltage criteria were used for online spike sorting, which was recorded at 30KHz for offline analysis. Custom LabVIEW software (National Instruments) was used to implement a configurable algorithm for BCI control as well as manual (torque control) of cursor control tasks. The decoding algorithm and cursor task operated at a sampling rate of 60Hz, while the display refresh rate was 30Hz. The combined BCI/manual computer cursor control system input-output latency was measured to be about 50ms. Latency was measured by mapping a cursor state variable to a hardware output, then using an oscilloscope to simultaneously capture input and output signals. The experiments were approved by the University of Washington Institutional Animal Care and Use Committee.

B. BCI Architecture

We implemented a simple decoder based on population vector mapping [24]. For each unit, spike counts in $1/60s$ bins were first filtered with a truncated Gaussian kernel, $\sigma = 0.05s$, before a baseline firing rate estimate was subtracted and the resultant instantaneous firing rate modulation estimate contributed to the population vector sum $\mathbf{u}[t]$ according to

$$\mathbf{u}[t] = \sum_{i=1}^{N} \alpha_i (f_i[t] - \hat{b}_i[t]) \tag{6}$$

$$\hat{b}_i[t] = \hat{b}_i[t-1]\gamma + f_i[t](1-\gamma). \tag{7}$$

The baseline rate $\hat{b}_i[t]$ for each filtered firing rate $f_i[t]$ was continuously updated using an exponential moving average filter with decay constant $\gamma = 0.99999$ resulting in a decay time constant of 28 minutes (we set initial values using a time constant of 2.8 minutes for several minutes at the beginning of each experimental session). This basic per-channel algorithm is illustrated in Figure 1. Modulation from each of 2-4 cells, typically 4, were summed with weights $\alpha_i = \pm 0.01$ to determine cursor velocity. Selection of cortical units, and the sign of their corresponding α_i, were chosen on the basis of observed correlations with torque production during manual control, or on the basis of modulation depth. The same algorithm was used to implement manual control tasks, except $f_i[t]$ was replaced with the flexion/extension torque signal recorded from the manipulandum and α_i were generally fixed across days and chosen by the experimenter.

C. Behavioral Tasks

We conducted experiments in a primate behavior booth outfitted with a computer monitor, buzzers, and a computer-controlled feeder containing apple sauce. The animal's arm, contralateral to the implanted array,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2016.2584101, IEEE Transactions on Neural Systems and Rehabilitation Engineering

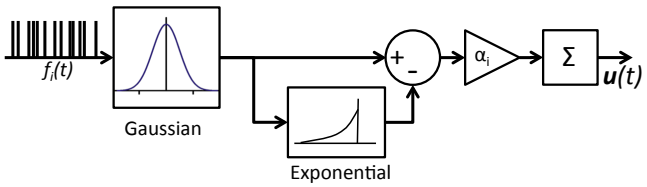

Fig. 1. Block diagram of basic BCI algorithm used for each single-unit or torque input channel. The summation block combines signals from several copies of the basic system for multiple single units.

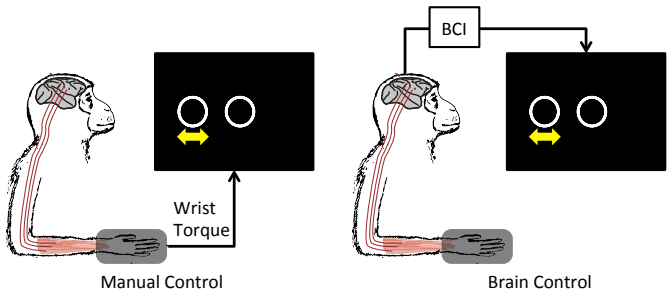

Fig. 2. The two 1D cursor positioning tasks compared in this study. Cursor velocity was controlled either by wrist torque (left) or neural modulation (right) of up to 4 single units. The primate's forearm was enclosed in a nearly-isometric manipulandum, minimizing proprioception of limb state relevant to the task. The cursor center had to be held within the target radius for 1 second for a successful trial, and a 0.5s break separated trials, during which the screen was blank.

was situated in a custom 2-DOF near-isometric manipulandum. The computer monitor was 30 cm x 23 cm, and located 28 cm in front of the animal's head. The manipulandum was used for task training and to measure motor correlates of neural activity. It also eliminated the possibility that neural activity is accounted for by limb kinematics or dynamics, since there were minimal postural changes.

We presented the macaque with 1D cursor positioning tasks with either of two control modalities: during manual control (MC), isometric wrist torque determined cursor velocity, and during brain control (BC), velocity was determined by aggregate cortical single-unit activity, both via Equation 6.

The animal was tasked with moving the center of a circular cursor to within the boundary of a circular target, then holding the cursor inside the target for 1s to receive an applesauce reward, as shown in Figure 2. Thus, a target of width W centered a distance D from the starting position required a minimum movement of $D - W/2$ to reach the boundary. The cursor and target were always shown with the same radii. Targets were randomly drawn from a discrete set with position referenced to cursor position at the beginning of each trial, and only one target was shown and selectable during a given trial. A $0.5s$ break was provided between trials, and trials timed out if the target was not acquired within $40s$. The cursor was always under the primate's control, even when not visible on the screen, with the partial exception of movement limits at screen edges. Candidate target locations were constrained to be greater than $W/2$ from screen edges so that it was always possible to overshoot the target. The cursor was initialized to the center of screen at the beginning of trial blocks.

In a typical daily experiment session, we first selected putative single units correlated with manipulandum torque and assigned them to the BCI mapping during several minutes of a 2D warm-up task. Then, we conducted alternating 10-minute blocks of MC and BC with 1-minute breaks in between. Target widths were typically held constant during a single block, and randomly-selected distance was biased or constrained to help balance the distribution of trials at each (D, W) condition.

For the purpose of performance analysis in our dwell-to-select context, we subtract the constant 1s dwell from each trial, and parameterize distance using center-to-boundary distance as illustrated in Figure 3. This distance measure is consistent with the ID definition used by Gilja et al [2], and subtracting dwell time is standard in the application of Fitts's law [16]. This approach is also similar to previous successful modeling of a crossing-based interface using Fitts's law [25], and justified because the task requirement becomes simply to avoid leaving the target area after the boundary is crossed.

D. Statistical Analyses of Model Performance

Our use of target geometries randomly chosen from a discrete set for each trial led to a data set with repeated measures unevenly distributed among conditions. We also observed large variability in trial times during brain control. We therefore chose to fit only to conditions with enough trials to estimate typical performance, according to the criteria below (III-D2), rather than fitting models to every trial in a given data set. This is consistent with models predicting typical performance, and the resultant R^2 values will reflect a graded measure of model performance. In contrast, fitting to every trial depresses R^2 values (approaching perfect prediction of average movement times does not drive R^2 to 1), diminishing their interpretability as a graded model performance measure, and significance tests provide only binary outcomes. Non-normal trial time distributions motivated the substition of median for average trial times in our analyses. We selected trial sets

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2016.2584101, IEEE Transactions on Neural Systems and Rehabilitation Engineering

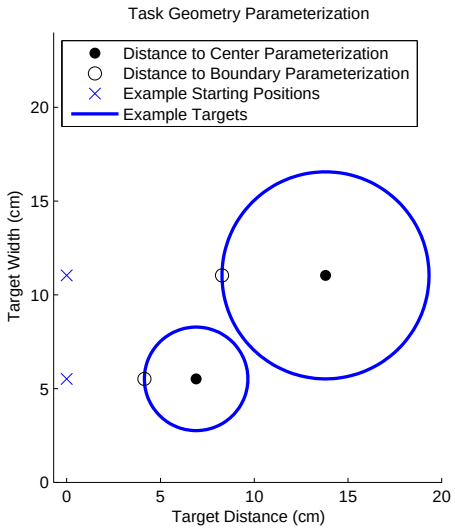

Fig. 3. Plot of example task geometries, including starting positions ($\times$'s) and visual target (large circles), showing two choices of task parameterization (D, W) for model fitting: defining distance as to target boundary (open circles) represents a significant correction in modeled task difficulty compared to traditional Fitts's law distance measured to target centers (solid dots).

well-described by at least one candidate model to support comparisons of model performance and discussion of the estimated model parameters.

1) Parameterizing Trial Time Distributions: The convention is for Fitts's law to predict average trial time at each (D, W) condition, under the assumption of normally-distributed trajectory endpoints in click-to-select tasks. An alternative possibility, suggested by the fact that our data represent task completion times, each of which could represent multiple attempts at holding the target, is that trial times have a log-normal distribution. To assess normality, we can combine trial times across (D, W) conditions into a single distribution by subtracting the (D, W) mean from each. We can then plot this distribution histogram, and compare it, and its log transform, to the normal distribution using a quartile-quartile (Q-Q) plot. An example of this analysis is shown Figure 4 for a set of trial times during brain control on one experimental day. We chose to use the median trial time to represent each (D, W) condition, because this heuristic analysis of trial time distributions suggested that they have non-normal, skewed distributions that resemble log-normal distributions.

2) Trial Set Definition and Screening Criteria: A trial set is drawn from trials during a single daily experimental session, and defined as all trials from a single control condition (MC or BC, at a single gain in the case of MC). Within each set, we require a minimum of 30 successful trials at each (D, W) condition to have confidence in the sample median, and so trials at a (D, W) condition not meeting this criteria are discarded. After this culling, each trial set is required to have 4 unique (D, W) conditions so that the 3-parameter Shannon-Welford model can be applied (we do not require a fully crossed or balanced set, and consequently include degenerate cases where 3 or 4 of only 4 tested conditions share the same D). Finally, we discard trial sets without at least a 95% success rate, following the precedents of 90% and 80% in the Felton et al study [13], and of the standard of capturing 96% of trajectory endpoints in Fitts's law analysis of click-to-select tasks (see II-D).

3) Modeling and Strength-of-Fit Screening: We then apply the Fitts's law and Shannon-Welford models to each eligible trial set, using a least-squares regression against median trial times, and weighted by the number of trials at each (D, W) condition. We calculate R^2 values for each model fit, and perform a final screening requiring at least one model to have a degree-of-freedom adjusted $R^2 > 0.9$, consistent with the criteria proposed by MacKenzie in 1992 for the application of Fitts's law [15]. We use the trial sets for which at least one of our two models showed strong predictive power as test cases for comparative analyses and for visualization of parameter estimates and trial time prediction surfaces in Section IV.

4) Testing Comparative Model Performance: Our goal is to test the predictive power of two different empirical models, where one model is nested with the other, the latter having an additional parameter. For this, we desire a statistical hypothesis test, not simply a heuristic that facilitates model selection, such as comparing R^2 values or Akakie Information Criterion (AIC) or Bayesian Information Criterion (BIC) metrics. We do not know what margin of difference one of these measures must show between models to have confidence that the difference is significant. We therefore report degree-of-freedom adjusted R^2, as a standard measure of model power, but do not use it to determine which model performs significantly better.

We can instead use the F-test to perform one-way statistical hypothesis tests of relative model performance. Specifically, the F-test of nested models [26] can be used to test whether the Shannon-Welford model provides significantly improved explanatory power over Fitts's law for a given trial set. The F-test is the only available analysis that provides a statistical hypothesis test of whether the more complex model is significantly better. To apply it here, we use the F-test $F(p_2 - p_1, n - p_2)$ with significance level $p \leq 0.05$, where $p_2 = 3$ is the number of parameters in the Shannon-Welford model;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2016.2584101, IEEE Transactions on Neural Systems and Rehabilitation Engineering

$p_1 = 2$ is the number of parameters in Fitts's law; and n is the number of median trial times in the analysis. This closely follows the comparative analysis of these same models in the recent study by Shoemaker et al [3].

IV. Results

Our screening criteria for 30 successful trials at each of at least 4 (D, W) conditions per trial set yielded 20 MC sets and 51 BC sets, however 11 BC sets were removed due to success rates below 95%. We next found that 12 MC trial sets and 35 BC trial sets failed to result in adjusted $R^2 > 0.9$ for either model, leaving us with 8 MC (40%) and 5 BC (10%) trial sets out of the original pools. Our nested F-test results show that the Shannon-Welford model is significantly better in 4 of the 13 comparisons we performed. Although these results are mixed, they do not reject any of our hypotheses. We show below that the mixed comparative results are consistent with the results of Shoemaker et al when viewed in context with the suitability of the trial sets for model fitting and the estimated values of the model parameter k.

We show an example of our heuristic analysis of trial time distributions in Figure 4. The histogram reveals an asymmetrical distribution with very long tails (truncated in the plot), which are not fully eliminated by a log transform, as shown in the second Q-Q plot. Consequently, we elected to perform all model fitting using median trial times.

For insight on the fragility of Fitts's law compared to the Shannon-Welford model, we show in Figure 5 examples of both models fit to trial set #1. We chose a set where $k = 0.56$ with the expectation that Fitts's law would over-estimate the influence of target width on task difficultly, which is reflected in the separation of the two W groupings despite similar trial times when plotted against ID. Fitts's law is able to model either grouping with an excellent coefficient of determination, which shows that if this particular experiment did not include multiple target widths, there would be no indication that the results do not generalize to different target widths. However, when all six conditions are fit, the coefficient of determination for Fitts's law is poor while the Shannon-Welford model performs well. We confirm the significantly better performance of the Shannon-Welford using an F-test, included in Table I.

We provide a table of model fitting results for all trial sets where at least one model performed well in Table I. We sorted the enumerated sets by control condition (MC/BC), then by gain for MC sets, and finally by descending p of the F-test results. D, W, and $D \times W$ refer

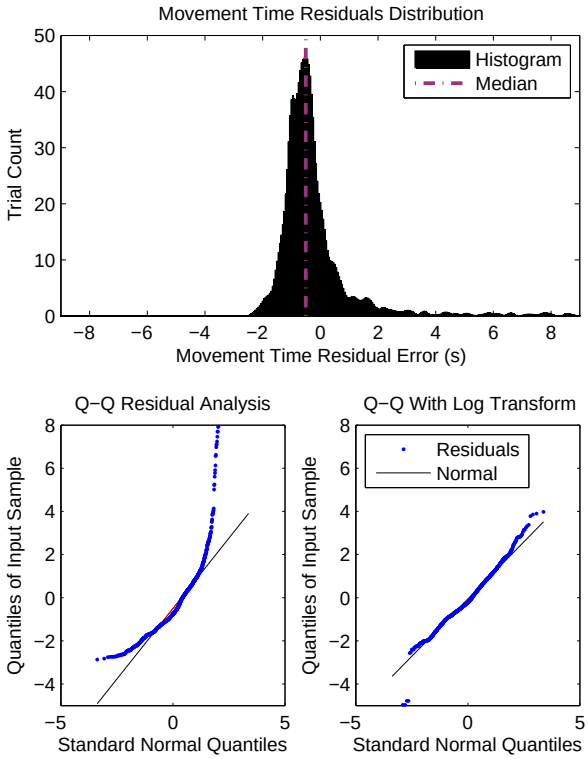

Fig. 4. Heuristic analyses of trial-time distributions reveal an asymmetrical distribution with long tails (truncated in the plots), which are not eliminated by the log transform. Top, histogram of trial times with (D, W) condition means subtracted, for the set of BC trials from trial set 13, with median shown. Bottom are quartile-quartile (Q-Q) plots comparing this histogram and its log-transform, respectively, to the standard normal distribution.

to the number of distinct target distances, widths, and unique combinations thereof included in each trial set. The number of $D \times W$ levels is not equal to the product of D and W levels because not every combination was tested with enough trials. Nested models F-test results in the rightmost column show in which cases the Shannon-Welford model is significantly better than Fitts's law ($p < 0.05$).

An examination of R^2 (not adjusted for degrees of freedom) shows that without exception, the Shannon-Welford model produces higher R^2 values, as it mathematically must. Further, in most cases where Fitts's law showed a degree-of-freedom adjusted coefficient of determination ($\bar{R}^2$) above 0.9, the Shannon-Welford model did as well. The exception is one degenerate case (3) in which all task conditions are at the same W. Note that we have four such degenerate trial sets (3, 9, 10, 11), in which three or four of only four total (D, W) conditions share the same W value. For a quantitative categorization of suitability for model fitting,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2016.2584101, IEEE Transactions on Neural Systems and Rehabilitation Engineering

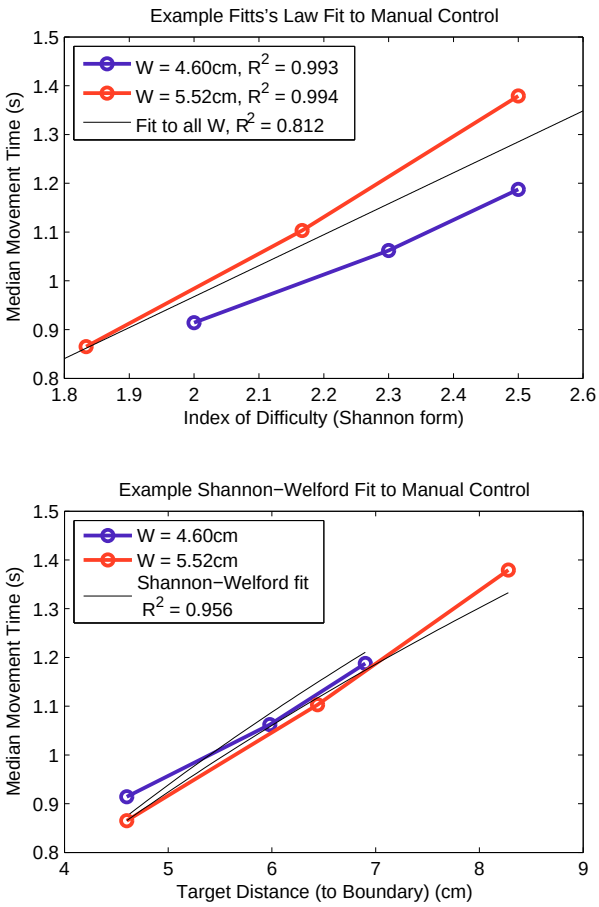

Fig. 5. Example of model fitting with Fitts's law (top) and the Shannon-Welford model (bottom), with MC trial set 1. Fitts's law is applied to conditions grouped by W (trendlines not shown, see R^2 in legend) as well as to all six conditions (black trendline). Note that x-axis is different in each, and the bottom plot represents two super-imposed slices of a surface plot (the two black curves are the same model) in D-W space.

we examined the condition number of the Jacobian matrix in the last iteration of model regression. We found that for most trial sets the condition number was $< 10^2$, but for 3, 9, and 10 it was $> 10^8$, also corresponding to cases where the F-test yielded a p-value of 1.0. The high condition numbers indicate that good parameter estimates for our two-factor model cannot be derived from those ill-conditioned trial sets, so their confidence intervals are omitted in Figure 6.

For four trial sets, Fitts's law shows $\bar{R}^2 < 0.9$ while the Shannon-Welford model retains predictive power, shown in bold in Table I. The F-test of nested models shows that the Shannon-Welford model is significantly better ($p < 0.05$) at describing these same 4, out of the 10 trial sets with well-conditioned model fits.

Manual Control

Set	D	W	D×W	Trials	$\bar{R}^2$ Fitts	$\bar{R}^2$ SW	k	F-test $p =$
1	**5**	**2**	**6**	**511**	**0.811**	**0.956**	**0.6**	**0.033**
2	**4**	**2**	**5**	**719**	**0.710**	**0.968**	**0.5**	**0.037**
3	*4*	*1*	*4*	*621*	*0.928*	*0.856*	*0.1*	*1.000*
4	4	2	7	751	0.943	0.930	0.9	0.819
5	7	2	8	1143	0.975	0.971	0.9	0.661
6	6	2	7	886	0.947	0.946	1.1	0.404
7	5	2	9	929	0.969	0.978	0.8	0.099
8	**5**	**2**	**10**	**982**	**0.844**	**0.916**	**0.5**	**0.026**

Brain Control

Set	D	W	D×W	Trials	$\bar{R}^2$ Fitts	$\bar{R}^2$ SW	k	F-test $p =$
9	*4*	*1*	*4*	*397*	*0.986*	*0.971*	*1.4*	*1.000*
10	*4*	*1*	*4*	*209*	*0.967*	*0.935*	*0.5*	*1.000*
11	4	2	4	412	0.996	0.993	1.0	0.683
12	11	4	15	1368	0.940	0.942	0.7	0.243
13	**5**	**5**	**11**	**1155**	**0.693**	**0.927**	**1.4**	**0.001**

TABLE I
SUMMARY OF FITTING RESULTS. TOTAL NUMBER OF UNIQUE DISTANCES AND WIDTHS ARE SHOWN, AS WELL AS UNIQUE TARGET GEOMETRIES ($D \times W$), FOLLOWED BY FITTING RESULTS. ITALICIZED ROWS INDICATED TRIAL SETS WITH ILL-CONDITIONED MODEL FITS FOR SHANNON-WELFORD MODEL, WHILE BOLD ROWS INDICATE SIGNIFICANTLY BETTER PERFORMANCE OF THE SHANNON-WELFORD MODEL.

We note that excepting ill-conditioned cases, the trial sets for which the Shannon-Welford model performs significantly better are those with k values farthest from one. We highlight this pattern in a plot of k estimates versus trial set in Figure 6, with significantly better Shannon-Welford performance denoted by larger markers. This is consistent with the fact that the two models are equivalent when $k = 1$.

We found that k values in the identified models fell within the range $[0.1, 1.8]$ reported in a previous study of multiple human-computer interfaces [3]. We did not find, and did not expect, a significant trend in the relationship between k and CD gain during manual control. We varied gain by only a factor of 2, whereas previous results varied gain by a factor of 10 and observed a total change in k of less than 1 [3].

Our estimates of k for well-conditioned cases averaged 0.78 for MC and 1.05 for BC. This suggests that width played a marginally dominant role in determining difficulty during BC, whereas distance played a dominant role during MC. Finally, we note that for the three sets of BC trials with well-conditioned fits, the Shannon-Welford model had good predictive power and provided estimates of k with confidence intervals on par with those identified during MC. We explore the implications of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2016.2584101, IEEE Transactions on Neural Systems and Rehabilitation Engineering

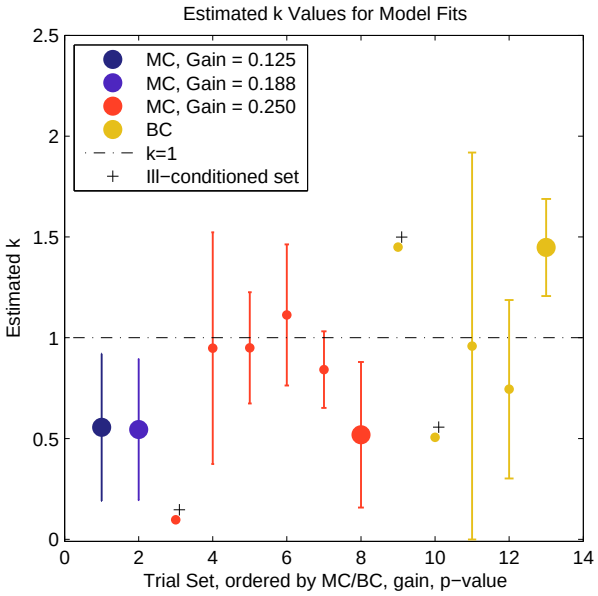

Fig. 6. In this plot of identified k values for 13 trial sets, large markers indicate instances where the Shannon-Welford model significantly outperformed Fitts's law, and vertical bars indicate 95% confidence intervals for k estimates. MC trial sets 1-8 are ordered by control-display gain, but no significant k vs. gain trend is indicated. The dashed $k = 1$ line indicates where the Fitts and Shannon-Welford models are equivalent.

these results in the next section.

We chose the models from MC and BC with the smallest confidence interval on estimated k to compare predicted movement time surfaces using contour plots in Figure 7. Note that the prediction surface for MC only partially overlaps the BC prediction surface, because the surface boundaries in (D, W) space are determined by the range of experimental task geometries to avoid extrapolation. The gradient direction (orthogonal to the rendered contour lines) at each point in (D, W) space shows the relative marginal difficulty change if D or W is changed. Thus, we can see that where the prediction surfaces meet, BC difficulty is more sensitive to changes in W while MC is more sensitive to changes in D.

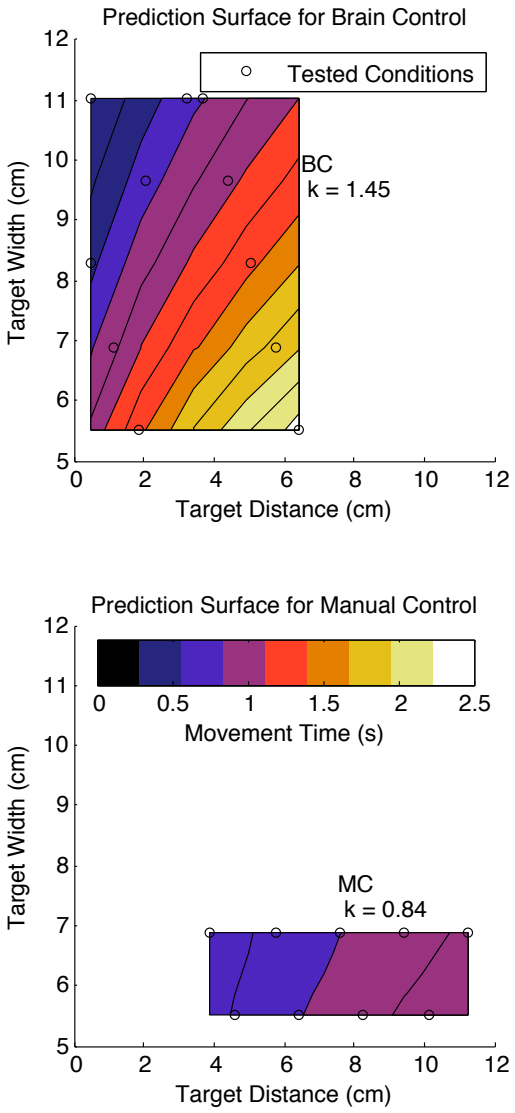

Fig. 7. These contour plots compare movement time predictions for manual and brain control, using trial sets 7 and 13, chosen because they had the smallest confidence intervals for the k estimate. We match prediction surface boundaries to the range of tested task geometries, shown by open black circles; for this reason they are only partially overlapping.

V. Discussion

A. Summary of Results

We showed that the Shannon-Welford model more robustly describes movement times in reaching tasks controlled manually as well as directly by cortical activity in a non-human primate, outperforming the incumbent Fitts's law model. In particular, it captures performance when the relative influence of distance and width are disproportionate with the assumption of Fitts's law. This result is consistent with prior work on the use of empirical movement models as performance metrics for pointing tasks, including recent works probing the fragility of the Fitts's law model [22], and establishing superiority of the Shannon-Welford model [3]. At the same time, less than half of the experimental sessions tested were well-described by either model, stressing the danger of assuming a movement model is valid. The variability of our identified parameters for manual control across well-modeled days provides context for interpreting brain control parameter estimates, and the examples of ill-conditioned fits show the importance of balanced experiment design.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSRE.2016.2584101, IEEE Transactions on Neural Systems and Rehabilitation Engineering

B. Insights About CD Gain from the k Parameter

By construction, and supported by experimental results, the k parameter in the Shannon-Welford model captures the relative difficulty contributions of target width and distance. Therefore, the model offers a quantitative way to refine a common intuitive method of gain assessment: that it is too high if it is hard to stay on-target, and too low if it takes too long to get to the the target. Additionally, the Shoemaker et al study [3] identified a linear relationship between the k parameter and gain, which held for mouse as well as gesture-based pointing with a wall-sized display. This information motivates new research questions: is there a correlation between k and subjective user preference of a gain setting? If so, does this gain setting have correlates in the model-predicted trial times?

If we assume this relationship holds for interfaces including BCIs, which is consistent with although not shown by our data, then the k parameter provides a measure of the effect of gain on interface performance, i.e. an indirect unit-free measure of CD gain. Two opportunities for comparison then arise. First, we can compare the effect of BCI gain across experimental sessions with different decoder weights and neural signals, where no well-defined direct measure of CD gain exists. Second, we can compare the effect of CD gain during BC with the effect of gain during MC.

For example, during each session with an otherwise fixed BCI decoder, an experimenter can collect enough trial data to estimate k for two or three levels of gain. The experiment can be repeated with different decoders or with re-fitted decoders using the same algorithm. If k correlates with gain within the same decoders, then it is a validated method to indirectly measure gain for BCIs, e.g. across days with re-fitted decoders and potentially with different decoding algorithms.

C. Evaluating Performance Across Studies

The Shannon-Welford model introduces richer details into the conditions where one interface enables lower movement times than another, and motivates a new perspective on cross-study comparison. While Fitts's law can only predict a point of intersection in predicted movement times along the ID axis, intended as an abstract measure of both information and difficulty, the Shannon-Welford model offers prediction surfaces in the physical (D, W) space. This creates the possibility of curved intersections between two prediction surfaces, partitioning the task geometry space according to which interface will enable faster movements.

There are two consequences of this: first, experimenter choice of task geometry can bias an an experiment testing two control interfaces; second, the only sufficient test of interface performance for a particular task is experiments designed to include the same range of task geometries. The use of the Shannon-Welford model, and reporting of tested geometries and CD gains, therefore provides a check as to whether experimenter choice of task geometry may have biased an experiment in favor of one control mode. There is no gold standard of appropriate task geometries to test; this choice is ultimately grounded in the degree to which the experiment is representative of a real-world task.

We hope to have convinced the reader that there is much to gain from switching to this richer and more descriptive empirical model of reaching movements, and that the non-conformity of experimental data with Fitts's law emphasizes the importance of reporting complete experimental details. Although the Shannon-Welford model does not resolve the challenges of cross-study comparison, we submit that it represents an important step in the right direction. Along these lines, we endorse the consensus opinion from the 2013 International BCI Meeting at Asilomar [27], with the important exception of their endorsement of Fitts's law as defined in the ISO standard. We submit that the Shannon-Welford model should be used instead, because it offers a more robust standard model and the intuitively meaningful k parameter; results from modeling with Fitts's law can also be reported for completeness.

D. Trade-offs Unique to BCI Experiments

Here we examined 1-dimensional movement tasks, the simplest test case for this movement model. The model can be applied to 2- and 3- or even higher dimensional tasks [28], as well as over-actuated experimental paradigms where the controlled degrees of freedom (DOF) can exceed the task degrees of freedom (e.g., [29]). We expect the model to work independent of dimensionality, allowing a distance metric to be substituted for D and W in the one-dimensional case addressed here, but it may be necessary to design the distance metric to ensure that speed/accuracy tradeoffs across dimensions are equivalent. A movement model based on task DOF may seem inadequate for overactuated systems, but this is exactly the context in which a task-based model can reveal similar performance tradeoffs between BCI and native motor control. In a native motor task, task difficulty, and consequently movement times, marginally increase with task DOF [9], despite the *control* DOF (i.e. the arm) remaining exactly the same -

and much larger than the task DOF. We expect to see this dissociation between task and control degrees of freedom in proficient BCI-mediated control. Movement times well-explained by an experimentally validated movement model are a necessary condition to claim performance tradeoffs on par with the native motor system in pointing tasks.

VI. Conclusion

In developing summary recommendations for BCI task design, we are keenly aware that each trial performed with a BCI consumes precious experimental time and subject motivation. While we are sensitive to the limited experimental time available to test both animal and human BCI performance, we recommend the use of a balanced design with at least 2, and ideally 3 distinct distances and widths of target – with the largest possible separation – to efficiently determine the Shannon-Welford model fit. We acknowledge the frustration at the loss of a single quantitative performance measure, but emphasize that the failure to inform nuanced attribution of variability, in addition to the logical inconsistency of information claims, compel the community to find a better standard measure than Fitts's law.

Acknowledgment

The authors wish to thank Robert Robinson for his excellent animal care and assistance during these experiments, and the Moritz and UW Biorobotics Lab research groups for feedback on the material. We would also like to thank the editor and reviewers for helpful feedback.

This work was supported by American Heart & Stroke Association Scientist Development Grant (NCRP 09SDG2230091), a DARPA Young Faculty Award (D12AP00251) and the Center for Sensorimotor Neural Engineering (CNSE), a National Science Foundation Engineering Research Center (EEC-1028725). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, National Science Foundation, or other funding agencies.

References

[1] J. D. Simeral, S.-P. Kim, M. J. Black, J. P. Donoghue, and L. R. Hochberg, "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*, vol. 8, no. 2, p. 025027, apr 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21436513

[2] V. Gilja, P. Nuyujukian, C. a. Chestek, J. P. Cunningham, B. M. Yu, J. M. Fan, M. M. Churchland, M. T. Kaufman, J. C. Kao, S. I. Ryu, and K. V. Shenoy, "A high-performance neural prosthesis enabled by control algorithm design," *Nature Neuroscience*, vol. 15, no. November, pp. 1752–1757, nov 2012. [Online]. Available: http://www.nature.com/doifinder/10.1038/nn.3265

[3] G. Shoemaker, T. Tsukitani, Y. Kitamura, and K. S. Booth, "Two-Part Models Capture the Impact of Gain on Pointing Performance," *ACM Transactions on Computer-Human Interaction*, vol. 19, no. 4, pp. 1–34, dec 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2395131.2395135

[4] J. M. Carmena, M. A. Lebedev, R. E. Crist, J. E. O'Doherty, D. M. Santucci, D. F. Dimitrov, P. G. Patil, C. S. Henriquez, and M. A. L. Nicolelis, "Learning to control a brain-machine interface for reaching and grasping by primates." *PLoS Biology*, vol. 1, no. 2, p. E42, nov 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14624244

[5] J. M. Carmena and K. Ganguly, "Emergence of a Stable Cortical Map for Neuroprosthetic Control," *PLoS Biology*, vol. 7, no. 7, p. e1000153, 2009. [Online]. Available: http://dx.plos.org/10.1371/journal.pbio.1000153

[6] C. T. Moritz and E. E. Fetz, "Volitional control of single cortical neurons in a brain-machine interface." *Journal of Neural Engineering*, vol. 8, no. 2, p. 025017, apr 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21436531

[7] A. L. Orsborn, S. Dangi, H. G. Moorman, and J. M. Carmena, "Exploring time-scales of closed-loop decoder adaptation in brain-machine interfaces." *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, vol. 2011, pp. 5436–9, jan 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22255567

[8] A. J. Suminski, D. C. Tkach, A. H. Fagg, and N. G. Hatsopoulos, "Incorporating feedback from multiple sensory modalities enhances brain-machine interface control." *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, no. 50, pp. 16 777–87, dec 2010. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3046069{&}tool=pmcentrez{&}rendertype=abstract

[9] J. O. Wobbrock and K. Shinohara, "The effects of task dimensionality, endpoint deviation, throughput calculation, and experiment design on pointing measures and models," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11)*. Vancouver, British Columbia: New York: ACM Press, 2011, pp. 1639–1648. [Online]. Available: http://portal.acm.org/citation.cfm?id=1979181

[10] J. Dethier, P. Nuyujukian, S. I. Ryu, K. V. Shenoy, and K. Boahen, "Design and validation of a real-time spiking-neural-network decoder for brain-machine interfaces." *Journal of neural engineering*, vol. 10, no. 3, p. 036008, jun 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23574919

[11] D. M. Taylor, S. I. H. Tillery, and A. B. Schwartz, "Information conveyed through brain-control: cursor versus robot." *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 11, no. 2, pp. 195–9, jul 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12899273

[12] G. Santhanam, S. I. Ryu, B. M. Yu, A. Afshar, and K. V. Shenoy, "A high-performance brain-computer interface." *Nature*, vol. 442, no. 7099, pp. 195–8, jul 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16838020

[13] E. a. Felton, R. G. Radwin, J. a. Wilson, and J. C. Williams, "Evaluation of a modified Fitts law brain-computer interface target acquisition task in able and motor disabled individuals." *Journal of neural engineering*, vol. 6, no. 5, p. 056002, oct 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19700814

[14] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement." *Journal of Experimental Psychology*, vol. 47, no. 6, pp. 381–391, sep 1954. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1402698

[15] I. S. MacKenzie, "Fitts' law as a research and design tool in human-computer interaction," *HumanComputer Interaction*, vol. 7, no. 1, pp. 91–139, 1992. [Online]. Available: http://www.informaworld.com/index/784766241.pdf

[16] R. W. Soukoreff and I. S. MacKenzie, "Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts law research in HCI," *International Journal of Human-Computer Studies*, vol. 61, no. 6, pp. 751–789, dec 2004. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1071581904001016

[17] R. W. Soukoreff, J. Zhao, and X. Ren, "The Entropy of a Rapid Aimed Movement: Fitts Index of Difficulty versus Shannons Entropy," *Human-Computer InteractionINTERACT ...*, pp. 222–239, 2011. [Online]. Available: http://www.springerlink.com/index/B6J468N381X82X81.pdf

[18] ISO, "Ergonomic requirements for office work with visual display terminals (VDTs)Part 9Requirements for non-keyboard input devices (ISO 9241-9)," 2002.

[19] X. Zhang and I. S. MacKenzie, "Evaluating Eye Tracking with ISO 9241 - Part 9," pp. 779–788, 2007.

[20] E. R. F. W. Crossman, "The measurement of perceptual load in manual operations," Doctoral dissertation, Birmingham University, 1956.

[21] G. Casiez, D. Vogel, R. Balakrishnan, and A. Cockburn, "The Impact of Control-Display Gain on User Performance in Pointing Tasks," *Human-Computer Interaction*, vol. 23, no. 3, pp. 215–250, jul 2008. [Online]. Available: http://www.informaworld.com/openurl?genre=article{&}doi=10.1080/07370020802278163{&}magic=crossref||D404A21C5BB053405B1A640AFFD44AE3

[22] E. Graham, "Pointing on a computer display," 1996. [Online]. Available: http://dl.acm.org/citation.cfm?id=223691

[23] A. T. Welford, *Fundamentals of Skill*, Methuen, London, 1971.

[24] A. P. Georgopoulos, A. B. Schwartz, and R. Kettner, "Neuronal population coding of movement direction," *Science*, vol. 233, no. 4771, pp. 1416–1419, sep 1986. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.3749885

[25] J. Accot and S. Zhai, "More than dotting the i's — foundations for crossing-based interfaces," *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02*, no. 1, p. 73, 2002. [Online]. Available: http://portal.acm.org/citation.cfm?doid=503376.503390

[26] N. R. Draper and H. Smith, *Applied Regression Analysis, 3rd Edition*, 1998.

[27] D. E. Thompson, L. R. Quitadamo, L. Mainardi, K. U. R. Laghari, S. Gao, P.-J. Kindermans, J. D. Simeral, R. Fazel-Rezai, M. Matteucci, T. H. Falk, L. Bianchi, C. a. Chestek, and J. E. Huggins, "Performance measurement for brain-computer or brain-machine interfaces: a tutorial." *Journal of neural engineering*, vol. 11, no. 3, p. 035001, jun 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24838070

[28] P. J. Ifft, S. Shokur, Z. Li, M. a. Lebedev, and M. a. L. Nicolelis, "A brain-machine interface enables bimanual arm movements in monkeys." *Science translational medicine*, vol. 5, no. 210, p. 210ra154, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24197735

[29] J. L. Collinger, B. Wodlinger, J. E. Downey, W. Wang, E. C. Tyler-Kabara, D. J. Weber, A. J. McMorland, M. Velliste, M. L. Boninger, and A. B. Schwartz, "High-performance neuroprosthetic control by an individual with tetraplegia," *The Lancet*, vol. 6736, no. 12, pp. 1–8, dec 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0140673612618169

Charles Matlack received his BS in engineering from Harvey Mudd College, followed by his PhD in electrical engineering from the University of Washington in 2014. His dissertation work investigated adaptation for brain-computer interfaces in non-human primate models.

Howard Jay Chizeck received his Sc.D. from MIT in 1982. He is a Professor of Electrical Engineering and Adjunct Professor of Bioengineering at the University of Washington, and a member of the faculty in the Neuroscience graduate program. He became a Fellow of the IEEE in 1999 and the AIMBE in 2011.

Chet Moritz received his PhD from the University of California, Berkeley (2003), followed by post-doctoral training at the University of Colorado and the University of Washington. He is currently an Associate Professor in the departments of Rehabilitation Medicine and Physiology & Biophysics at the University of Washington.