Running Title: Effects of CI Training on a P300 BCI Spelling Task

The Effects of Early Confidence Interval Training on User Efficacy in a P300 Brain-Computer Interface Spelling Task

By

Adam Starkman

****

Submitted in partial fulfillment of the requirements for Honors in the Department of Neuroscience

UNION COLLEGE June, 2016

ABSTRACT

STARKMAN, ADAM The Effects of Early Confidence Interval Training on User Efficacy in a P300 Brain-Computer Interface Spelling Task. Department of Neuroscience, June 2016.

ADVISOR: Stephen Romero

Brain-computer interface (BCI) technology can provide communication for individuals suffering from degenerative neuromuscular disorders. The present study sought to demonstrate improved BCI performance in healthy individuals using confidence interval training with a P300 BCI spelling program. In this BCI interface, electroencephalographic (EEG) activity was recorded as participants attended to a specific target character within a matrix of flashing letters and numbers presented on a computer screen. The BCI uses the P300 Event Related Potential to select the intended character. In a prior patient case, use of a confidence measure that rejected questionable selections improved that user's spelling efficiency. The present study evaluated the use of this strategy for untrained individuals. Results suggest that confidence interval training resulted in less efficient spelling across four training sessions. This work suggests that early confidence interval training may be counter-productive if used early in training. Further analysis among a larger pool of participants is needed for definitive conclusions.

The Effects of Early Confidence Interval Training on User Efficacy in a P300 Brain-Computer Interface Spelling Task

Brain-computer interface (BCI) technology can provide communication for individuals suffering from degenerative neuromuscular disorders, opening up a fountain of applications determined to improve the quality of life of paralyzed individuals. BCI's transform electrical brain signals into usable outputs providing a non-muscular based form of communication replacing, restoring, enhancing, or supplementing central nervous system function (Wolpaw and Wolpaw, 2012).

Since their inception researchers have explored numerous types of BCI's, including the use of both invasive and noninvasive methods. Invasive BCI's have their roots in animal experimentation, making use of implanted electrodes in the pre-motor cortex and parietal motor control area of monkeys. These BCI's are designed to replicate neural firing patterns associated with movement. Noninvasive BCI's, the subject of the present study, historically operate through training various biofeedback systems (such as heart rate, electroencephalogram, renal blood flow, or dilation and constriction of peripheral arties). The landmark results of Neil E. Miller's work in the 1960's and 1970's suggested that autonomic functions could be controlled voluntarily, without any mediation from the somatic-muscular system (Birbaumer and Cohen, 2007).

Farwell and Donchin (1988) first described a P300 BCI as a reliable method that could be used by able-bodied young adults to input a string of characters to a computer, as a substitute for a typing finger. They demonstrated how a user could concentrate on one character out of 36, in a 6x6 matrix, and elicit a P300 event-related brain potential (ERP) associated with the illumination of that character. As each of the 36 characters flashed randomly, a combination of

letters and numbers on the computer monitor, the user was instructed to concentrate solely on the desired character; the P300 could then reliably be elicited through similarities with the Oddball Paradigm in which this component is usually found (Farwell and Donchin, 1988). This system allowed for the identification of the intended character online and in real time enabling users to spell at a rate of 2.3 characters/minute.

As noted directly above, the P300 response has been identified as an ERP elicited through the presentation of infrequent stimuli in a series of more frequent stimuli. As explained in the BCI2000 User Tutorial (2013), a P300 is usually elicited if four conditions are met: (1) a random sequence of stimulus events must be presented, (2) a classification rule must exist separating the series of events into two categories, (3) the participants' task must make use of this rule, and (4) one category of stimuli must be presented infrequently. These behavioral properties will successfully elicit a P300, which is characterized as a positive deflection in EEG over the parietal cortex about 300ms after the rare stimulus is presented.

For P300 BCI applications, the random sequence of stimulus events is represented by the individual illumination of each character in the matrix. The flash of the desired character represents the rare event in comparison to the rest of flashing characters in the matrix (Fabiani, 1987). The classification rule therefore is whether the desired character flashed, or one of the other 35 characters illuminated. By asking the participant to count the number of times the intended character flashes, they make use of this categorization. Thus, with the intended character flashing infrequently relative to the other 35 characters, the P300 is elicited upon the highlighting of the rare character the user is focusing on.

Much of the present day research related to P300 BCI'S explore the effects of various sized matrices, the speed of flashes, or the pattern in which each character illuminates (row-column compared to a checkerboard design). Such studies are designed to increase the rate, accuracy, and ease of spelling for users. Another method for increasing the utility of these applications is to optimize the algorithms used to assess the BCI responses. One such technique is the use of a confidence interval (CI) in the actual character selection. The CI is a measure of confidence a user must achieve prior to the BCI printing the user's selection on the computer monitor. It is aimed at restricting the number of incorrect selections made by the user. It was initially designed to improve the performance of a single patient's ability to use the P300 speller because performance had deteriorated after an illness (Baxter, 2016).

At the end of a sequence of flashes, each character in the matrix receives a score based on the P300 ERP magnitude and latency; the letter with the highest score is printed. Sometimes, the highest scoring letter is still a relatively small score or not much different than others. The confidence interval sets a threshold at which a certain score must be achieved for a letter to print. If the CI score is not achieved no letter is printed, and the user must redo that selection. While five letter words may now require ten selections using a CI, time is saved in the long run by eliminating the added backspace selection that is associated with an incorrect selection (which may result in an error in itself).

In theory, caretakers can adjust the CI threshold, ideally finding a value that prevents errors, but does not require the user to make excessive selection attempts. Nevertheless, the present study seeks to determine if use of a CI can be applied during initial sessions to improve the learning of a P300 BCI spelling application, as compared to users learning the interface without

a confidence interval. Effective CI training would indicate a prominent role for CI's in BCI systems worldwide, eliminating time and frustration associated with correcting mistakes common in BCI spelling during learning.

Method

Participants

Four adults (3 males, age >18, specific ages were not collected) free from neurological disease and no prior P300 BCI experience completed the experiment. All participants completed four sessions of P300-based copy spelling. Upon fifteen blank selections (those under threshold) in any particular confidence interval run, the participant was excluded. One participant was excluded from the CI group when he failed to complete a five-letter copy-spelling task in under 15 selections.

Materials (and Apparatus)

The participant's monitor displayed 36 relevant items (the English alphabet and numbers 0-9) and 36 non-relevant items (dots) arranged in a 9x8 matrix, where all items were light gray and the background was black (Figure 1).

Figure 1. The 9x8 matrix before the start of a run. The yellow letters represent the text to spell, the blank gray line underneath are where the participant's responses are printed, and the gray letters and dots make up the matrix.

EEG was recorded using electrode caps (Electrocap International, Inc.) embedded with 16 tin electrodes covering left, right, and central scalp locations (Fz, Cz, Pz, Oz, Fp1, Fp2, F3, F4, C3, C4, P3, P4, P7, P8, T7, T8) based on the modified 10-20 system of the International Federation (Sharbrough et al., 1991). The recordings were referenced to the right mastoid and grounded to the left mastoid. Signals were amplified using a Guger Technologies 16-channel g. USBamp biosignal amplifier. Signals were sampled a rate of 256 Hz, high-pass filtered at 0.5Hz, and low-pass filtered at 30Hz. Every effort was made to keep electrode impedances as low as possible, with a maximum of 40kΩ.

All aspects of BCI operation and data collection were controlled by the BCI2000 software platform running on a Lenovo T500 laptop (Intel Core2 Duo CPU, 2 Ghz, 1.9GB of RAM, Windows XP SP3). A stepwise linear discriminant function (SWLDA) was used to select and weigh the EEG features (voltages at specific EEG electrode locations in specific time-periods in the 800ms after the matrix flashed) that were used to classify the participant's response to each item and to thereby determine which item was the target (the desired selection), prior to the application of the confidence interval threshold. The SWLDA was used to derive features from the entire 16-electrode montage and from an 8-electrode subset (Krusienski, et al., 2005).

A P300 graphical user interface (GUI) was used to analyze data offline (online accuracy was determined in real time based on the results on the computer monitor). The offline analysis tool operated in two steps. First, recorded data represented by EEG time courses was transformed and represented by a specific set of features. A single feature corresponded to raw EEG amplitude at a certain time offset after the stimulus, in a certain channel. After this transformation, data was sorted into two groups according to the conditions specified in the

GUI. One condition was the presentation of an unattended stimulus, while the other condition is the presentation of an attended stimulus. Now, each feature possessed a number of sampled values taken from two conditions. These values were used to compute a number representing how much a feature's value told about the condition under which it was recorded. That measure is called the determination coefficient(i.e.,r-squared). The larger a feature's r-squared, the more correlation existed between a feature's value and the condition under which it was recorded. Simply put, offline accuracy is a measure of accuracy calculated using sparse data methods that demonstrates what an optimized system would look like (User Reference: BCI2000 Offline Analysis, 2012

Procedure

Participants were split into two groups, an experimental group which utilized the CI feature while making selections and a control group taught to use the BCI without the CI. The entire session consisted of a consenting process, electrode cap application, task instructions, eight runs, and cap removal. Each session took about 60 minutes. This study was reviewed and approved by the New York State Department of Health Institutional Review Board, and each participant gave informed consent.

Participants sat in a chair about 1 meter away from a 20” computer monitor for the duration of each session. Each session consisted of eight runs. A run was defined as a block of letters. The first run consisted of fifteen characters (i.e., WADSWORTHCENTER) and was used for calibration. Using the data from the first run of fifteen characters, the SWLDA classifier weights were developed with the 8-channel subset. The weights were then applied online for the following seven runs of data collection to determine online accuracy for each participant.

The remaining runs were each comprised of one five-letter word, for a total of 50 characters (or trials) per session. For each character selection (referred to as a trial as noted above) the participant was asked to pay attention to the target character and to count the number of times it flashed. The word to be spelled during a given run was displayed in the text-to-spell bar (TTSB) above the matrix on the computer screen. At the beginning of each run, the words waiting to start were displayed over the matrix, and the target item (e.g., the first letter of the word to be spelled) was shown in parenthesis at the end of the word in the TTSB. Each trial consisted of six sequences. For each sequence, the columns and rows of the matrix flashed twice in a random order (at a rate of 8Hz), i.e., six sequences of 17 flashes, or 102 stimuli in all. Each trial was followed by a brief pause during which the matrix items did not flash and the next letter in the word to be spelled was displayed in parenthesis in the TTSB. Once each letter in the word had served as the target item, the phrase 'Time Out' was displayed and the run was over. After several minutes, the next run began (described in McCane et al., 2014).

Results

The impact of the confidence interval on BCI accuracy was measured by online and offline accuracy, as well as a third measure of accuracy created for this study. Differences were compared between individual performances across four sessions, as well as by comparing performances by group (control versus experimental). Statistical analyses of these findings were not possible because there were only 4 participants. As described in Table 1, the results of each trial for the CI group was classified as either a false negative (FN), false positive (FP), true negative (TN), or true positive (TP). A negative designation indicated the user's selection was below threshold, while positive selections were above the CI value. All unprinted (or blank selections below threshold) were examined after the sessions to determine if they were either a TN or FN, while those that did print were classified as TP or FP. Positive or negative indicated if the user's selection printed or not, and true of false indicated if letter selection was correct or not. Selections from participants in the control group were either classified TP or FP, as none of their selections were be rejected by a CI.

Table 1.	**True**	**False**
Positive	*Printed Correct Letter*	*Printed Incorrect Letter*
Negative	*Correctly Rejected Selection Below Threshold*	*Incorrectly Rejected Selection Below Threshold*

Table 1. Each trial by control group participants were marked either as true positive or false positive; accuracy was defined as the number of true positives divided by total user selections. In order to account for blank selections, experimental group participants received one designation of the possible four shown above. Accuracies of experimental group participants was determined by the number of true positive and true negatives divided by total user selections.

Confidence Interval Results

The performances of the two confidence interval participants (CI1 and CI2) are characterized in Table 2. Out of the four possible outcomes per trial (TN, TP, FN, FP), the only clear trend across the four sessions for Subject CI1 was a decrease in FN's. Considering only the blank selections however, there was an increase in the percentage of TN's and decrease in FN's across each session. Subject CI2 experienced no clear trends in performance, although the number of FP's somewhat increased across the sessions.

Table 2	**Session 1**	**Session 2**	**Session 3**	**Session 4**
Subject CI1				
Total percentage of selections FN	20%	0%	14%	5%
Total percentage of selections TN	40%	0%	38%	63%
Total percentage of selections FP	19%	9%	11%	18%
Total percentage of selections TP	21%	91%	37%	14%
Total Selections	80	35	63	56
Number of blank selections	48	0	33	38
Percentage of selections blank	60%	0%	52%	68%
Percentage of blank selections TN	67%	N/A	73%	92%
Percentage of blank selections FN	33%	N/A	27%	8%
Subject CI2				
Total percentage of selections FN	11%	0%	3%	7%
Total percentage of selections TN	64%	0%	6%	16%
Total percentage of selections FP	18%	29%	24%	42%
Total percentage of selections TP	7%	71%	67%	36%
Total Selections	101	35	33	45
Number of blank selections	76	0	3	10
Percentage of selections blank	75%	0%	9%	22%
Percentage of blank selections TN	86%	N/A	66%	70%
Percentage of blank selections FN	14%	N/A	33%	30%

Table 2. A breakdown of the trial categorizations for both experimental group participants

Copy-Spelling Results

Copy-Spelling results consider what the users actually spelled on the monitor. Figure 3 shows the online accuracy of each participant after each session. It is referred to as online because it can be calculated in real time based on the spelled words achieved by the participant. Online accuracy for the control group was calculated by dividing the number of true positives by total user selections. The experimental group online accuracy was defined as the sum of true positives *and true negatives* divided by the total number of user selections (to account for the blank selections when user selections were below CI threshold). Including the rejected selections, despite the fact that such selections were not printed, demonstrates what the results would look like had the BCI printed everything. Therefore, this measure most accurately reflects how CI participants would perform without the CI.

The online performance of each participant fluctuated from session to session. The average online accuracy for the control group was 57%, however there was a good deal of variation across participants as CG1 achieved an average accuracy of 85% and CG2 achieved an average accuracy of 29%. CI1 finished with an average online accuracy of 76%, and CI2 finished with 67% accuracy, combining for a group average of 72%. Figure 4 depicts the participants' change in online accuracy across each session. Relative to their online accuracies, the control group achieved higher offline accuracies (94% and 38% for subjects CG1 and CG2, respectively, for an average of 66%), while the experimental group declined to an average of 64%. CI1 averaged 65% and CI2 averaged 62% (see Figure 5). Offline accuracies were determined through the P300 GUI described in the method section above.

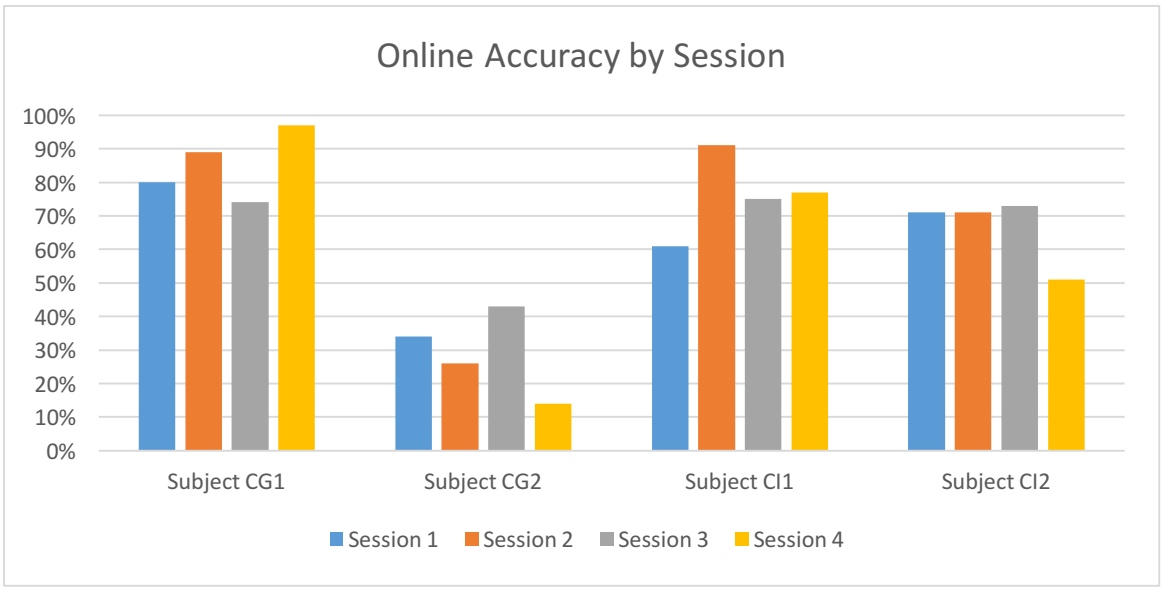

Figure 3. The online accuracy, a percentage calculated by the sum of TN's and TP's divided by the total number of selections (note for the control group TN was always 0).

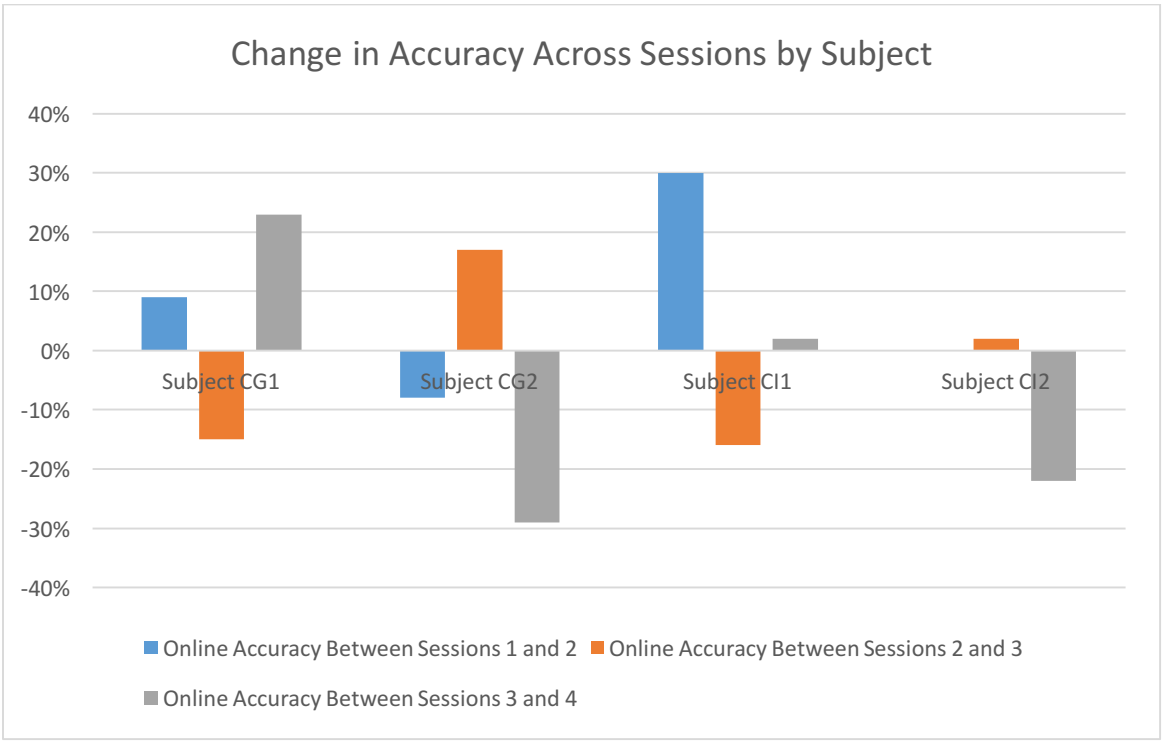

Figure 4. The change in accuracy across each session for the four participants.

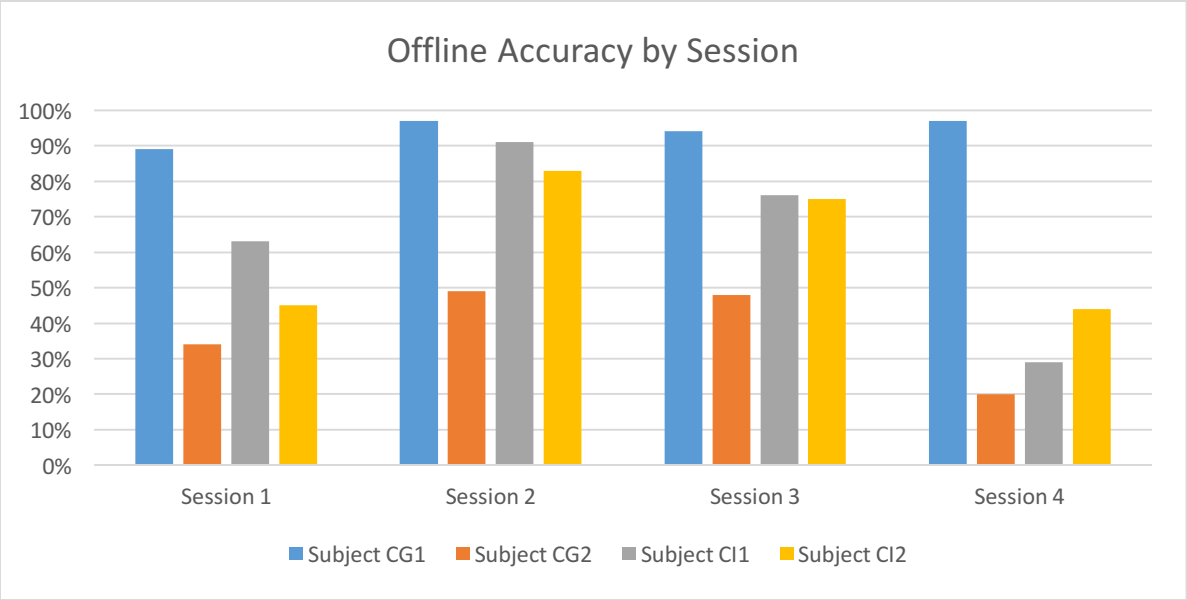

Figure 5. The offline accuracies for each participant at the end of each session. This accuracy was calculated offline (using sparse data methods) so we can guess what an optimized system would look like.

A third measure of accuracy created for this study reflects the efficiency of the selected CI threshold. This value defined accuracy as the number of true positives divided by total selections made above the CI threshold (a value that increases the influence of CI on accuracy, by only including confident selections). Therefore, this metric only considers TP's and FP's. While this does not change the accuracies of the control group, the experimental group saw a decline in performance. A summary of all the average accuracies according to this third measure can be found below in Table 3.

Table 3	**Online Accuracy**	**Offline Accuracy**	**Accuracy above CI Threshold**
CG1	85%	94%	N/A
CG2	29%	38%	N/A
CI1	76%	65%	66%
CI2	67%	62%	55%

Table 3. A comparison of average accuracies across the four sessions by participant. The online accuracy is calculated by dividing the TP's and TN's by the sum of selections, while the accuracy above CI threshold is solely the TP's divided by the sum of TP's and FP's. Offline accuracy was calculated using sparse data methods.

Discussion

The primary question of the current study was whether use of a confidence interval improves performance for novice P300 BCI users. Specifically, the use of the confidence interval during initial BCI use was tested to determine if there would be a reduction in the percentage of false negatives and false positives and an increase in the percentage of true positives and true negatives. The results do not support these expected trends. Furthermore, at the same CI threshold, and with training, the number of total selections per run should decrease across the four sessions; this was also not the case.

Looking at just the rejected selections in isolation however (those selections with negative designations), indicates if the CI threshold value was effective. By evaluating if the majority of rejected selections would have been accurate or not (would have been TP or FP) indicates if the specific CI threshold value was effective in doing its job. Out of just the blank selections, or those below threshold, the increase in TN's across each session for CI1 indicates that the CI was in fact operating at an effective value. This shows that each session the percentage of rejected selections was increasingly selections that should have been rejected, confirming an effective CI. This was not the case for CI2 however, who experienced a fluctuation (and somewhat decrease) in selections that were correctly rejected (TN's). Perhaps the CI threshold could have been raised in order to reduce the number of incorrectly rejected selections, although that may significantly increase the total number of selections. It is also important to recognize that although the trend is unclear for CI2, on each of the 4 sessions over 60% the rejected selections were appropriately rejected. While not necessarily reducing the total amount of incorrect selections, those that were rejected were rejected appropriately for

CI1. This indicates that perhaps there is a learning period, and CI2 may have benefited if further training sessions were provided.

The average online and offline accuracy, compared to evaluating accuracy from individual sessions, reflects how the CI influences accuracy over a longer period of time, relative to the control group. While the experimental group did have an overall higher average accuracy than the control group, suggesting the confidence interval feature increases accuracy, the results are only from two participants with a larger variance in this performance measure (good performance for CG1 and much poorer performance for Subject CG2). Similarly, the results for the control group were equally variable, consisting of one participant who performed very well and one participant who struggled with the system, therefore producing mediocre average accuracies. Because of this drastic difference in performance between the two control and CI participants, along with an insufficient number of participants, it is hard to draw strong conclusions from this study.

Alternatively, the measure of change in online accuracy from session 1 to session 4 demonstrates if there was a learning curve to be accounted for, and how experience affects accuracy for both CI and non-CI users. Dissecting the apparent trends in results across groups, CI1 and CG1 both experienced an increase, decrease, then increase in online accuracy across the sessions while CI2 and CG2 performed with opposite trends. Additionally, CI1 and CG1 achieved a net gain in accuracy while CI2 and CG2 experienced a net decrease. Therefore, based on this data, there is no conclusion as to which group learns more quickly. There is a caveat however, as it is important to recognize results on any given day can be influenced by

motivation, fatigue, and background noise in the lab; a poor performance on session 4 could result from various external factors.

The third measure of accuracy, which measured the participant's accuracy while above CI threshold tests if the CI would be successful in elevating accuracies in application. Because the user only sees what is printed online (in real time), *and* when above CI threshold, this value indicates the perceived success of the CI to the user. Both participants achieved levels that seemed to be well above chance, however still low enough to be insufficient for day to day communication. Furthermore, these newly calculated accuracies were lower than the participant's online accuracies. This suggests that the CI was more accurate in rejecting incorrect responses than printing correct responses. In other words, it was more likely that a character under threshold was properly rejected than a letter that was above threshold being accurate.

Some limitations of the present study may explain the somewhat inconclusive results. As noted above, on any given day it is possible a participant had an off day, and with only four sessions this could drastically impact the performance statistics. Additionally, not every participant made an equal number of selections, as this depended on the confidence value achieved each run. Experimental group participants often made a greater number of selections than the control group, contributing to fatigue, frustration, distraction, and perhaps a loss of motivation.

Future studies should continue to test the potential benefits of a confidence threshold, or a way to limit the number of mistakes made by users of the P300-BCI. A larger subject pool

as well as a greater number of sessions could provide better insight into the mechanisms of P300 training.

References

Baxter, W., Vaughan, TM., Townsend, G., Zeitlin, D., Wolpaw, JR. Confidence measures may improve brain-computer interface (BCI) reliability for home use by people with ALS. Program No. 594.17/OO17. Neuroscience Meeting Planner, Washington, DC: *Society for Neuroscience*, 2011. Online.

Birbaumer, N., & Cohen, L. G. (2007). Brain-computer interfaces: Communication and restoration of movement in paralysis. *The Journal of Physiology, 579*(3), 621-636.

Fabiani, M., Gratton, G., Karis, D., Donchin, E., 1987. Definition, identification, and reliability of measurement of the P300 component of the event-related brain potential. *Advances in Psychophysiology* 2, 1–78.

Farwell, L., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology, 70*(6), 510-523.

Krusienski, D., Sellers, E., Vaughan, T.M., McFarland, D.J., Wolpaw, J.R., 2005. P300 matrix speller classification via stepwise linear discriminant analysis. In: Proceedings of the Third International Meeting of Poster Presentation at the Brain-Computer Interface Technology, Rensselaerville, New York.

Mccane, LM., Sellers, EW., Mcfarland, DJ., Mak, JN., Carmack, CS., Zeitlin, D., Vaughan, TM. (2014). Brain-computer interface (BCI) evaluation in people with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 15*(3-4), 207-215.

Sharbrough, F., Chatrian, GE., Lesser, RP., Luders, H., Nuwer, M., Picton, TW., 1991. American electroencephalographic society guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology* 8, 200–202.

User Reference:BCI2000 Offline Analysis. (n.d.). Retrieved June 01, 2016, from http://www.bci2000.org/wiki/index.php/User_Reference:BCI2000_Offline_Analysis

User Tutorial:Introduction to the P300 Response. (n.d.). Retrieved June 01, 2016, from http://www.bci2000.org/wiki/index.php/User_Tutorial:Introduction_to_the_P300_Response

Wolpaw J.R. and Wolpaw E. W. (2012) Brain-Computer Interfaces: Something New Under the Sun. *in* Brain-Computer Interfaces Eds. J.R. Wolpaw and E.W. Wolpaw (New York, NY: Oxford University Press; pp 3-12.