Toward Using Alpha and Theta Brain Waves to Quantify Programmer Expertise

Igor Crk and Timothy Kluthe

Abstract— **Empirical studies of programming language learnability and usability have thus far depended on indirect measures of human cognitive performance, attempting to capture what is at its essence a purely cognitive exercise through various indicators of comprehension, such as the correctness of coding tasks or the time spent working out the meaning of code and producing acceptable solutions. Understanding program comprehension is essential to understanding the inherent complexity of programming languages, and ultimately, having a measure of mental effort based on direct observation of the brain at work will illuminate the nature of the work of programming. We provide evidence of direct observation of the cognitive effort associated with programming tasks, through a carefully constructed empirical study using a cross-section of undergraduate computer science students and an inexpensive, off-the-shelf brain-computer interface device. This study presents a link between expertise and programming language comprehension, draws conclusions about the observed indicators of cognitive effort using recent cognitive theories, and proposes directions for future work that is now possible.**

I. Introduction

Evaluations of the human impact in Human-Computer Interaction studies have traditionally relied on performance measures such as task completion timings and completed task accuracies to draw conclusions. While these types of measurements are effective and simple to perform, they have also shown significant individual variation amongst participants. Where factoring out individual differences in such an experiment is desirable, measures of human expertise are typically subjective, relying on self-reporting and subject to several potential sources of bias, e.g., cultural preference, personal tendencies, and overestimation. Here we will examine a direct, non-invasive method of measuring brain activity for the purpose of augmenting or replacing self-reported data.

To the best of our knowledge, we are the first to explore the use of electroencephalogram (EEG)-based brain-computer interface devices in directly observing the cognitive workload associated with programming tasks. Recent studies of brain activity have served to refine the existing models of cognition to where they may provide a basis for useful interpretation of the brain's electrical activity captured by inexpensive devices. This study forms a fundamental proof-of-concept for the field of Human-Computer Interaction by using an EEG device to quantify expertise based on direct brain measurement.

II. Related Work

An EEG reading is accomplished by non-invasive placement of electrodes on the scalp to measure weak electrical potentials generated by brain activity (5-100 μV). Each electrode consists of a gold-plated disk placed close to the scalp and coated with a conductive liquid. Besides the highly invasive electrocorticogram, which requires surgical placement of electrodes within the skull, and magnetoencephalogram, which is prohibitive both in cost and practicality, it is the only method capable of providing a milliseconds-range temporal resolution of direct brain activity readings required for recording oscillatory activity. While the combined use of EEG and functional near-infrared spectroscopy can add the spatial resolution that EEG lacks and is considered an important avenue for our future work, we do not consider it here. Functional magnetic resonance imaging is generally cost-prohibitive, lacks the temporal resolution to capture oscillatory activity, and complicates observations by the subject's impractical setting.

Department of Computer Science, Southern Illinois University Edwardsville, Edwardsville, IL 62026, USA icrk@siue.edu, tkluthe@siue.edu

A. Alpha and Theta Waves

The alpha wave is the most prominent oscillatory activity evident in EEG readings. Alpha has repeatedly been shown to have a peak frequency in the 8-13 Hz range in healthy adults, with highest signal power, or synchronization, occuring during periods of restful wakefulness, i.e., eyes closed, performing no complex cognitive tasks. While the functional significance of alpha is unknown, it is generally considered to be a marker of cognitive inactivity or, more likely, linked to active inhibition of task-irrelevant populations of neurons. For this study, the critical observation is that alpha power increases (populations of neurons contributing to alpha synchronize) in the absence of a task, i.e., cortical inhibition [1], [2], [3], [4], [5], [6], and alpha power decrease (alpha desynchronizes) as increasingly many groups of neurons activate to meet some task demand [1], [5].

Prior work with oscillatory activity discourages the use of broad-band frequency ranges for power measurement. The peak, or gravity, alpha frequency varies between individuals, and displays a normal distribution within age-matched groups, meaning that a fixed 8-13 Hz band may not sufficiently capture all alpha activity [7]. Instead, we compute the Individual Alpha Frequency (IAF) for each participant, and used it as the basis for defining sub-band ranges termed Lower-1 Alpha (L1A, ranging from IAF-4 to IAF-2), Lower-2 Alpha (L2A, ranging from IAF-2 to IAF), Upper Alpha (UA, ranging from IAF to IAF+2), as well as theta (IAF-6 to IAF-4). Decomposition into these sub-bands was shown to improve precision and more accurately reflect the functional differences of oscillations [2]. The same study notes that a healthy individual's relative alpha power is not expected to change prior to 50 or 60 years of age, meaning that we are not measuring age-related changes. Event-related desynchronization (ERD) is a measure of the extent to which neuron populations no longer oscillate in synchrony, as they become activated and process task demands. We calculate ERD simply, as the percentage of band power change between the resting period preceding a trial and the trial itself.

B. Cognitive Load

Cognitive Load Theory (CLT) is based on the premise that working memory has a limited capacity, and proposes that humans are only conscious of information currently held in working memory. In describing the relationship between long-term memory, working memory and schemas [8], CLT provides the basis for interpretation of our findings.

In short, while working memory has limited capacity, able to hold 7 ± 2 items, this capacity is augmented and extended by schema pulled from long-term memory, each of which represents an entity whose complexity is unbounded. Since working memory

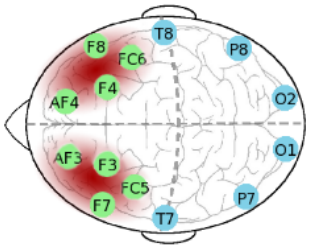

Fig. 1. The 14 data channels provided by the Epoc headset are labeled by their placement based on the International 10-20 system[2]. The shaded area indicates the general location of activity associated with working memory, and the nearby channels are the ones of interest to us.

performance is measured by spectral changes in the alpha frequency sub-bands [2], it follows that for like programming tasks, the observable phasic changes in alpha power should reflect, at least in part, an increased reliance on acquired complex schema for task interpretation and comprehension, observed as a relatively low working memory load, implying expertise, or a high working memory load, implying a lack of expertise.

Direct measures of cognitive load with the use of EEG have more recently seen practical application in studies of visualization effectiveness, dynamic activation of haptic guides, and video game learning rate [9], [10], [11].

C. Program Comprehension and Experience

Program comprehension plays a significant role in experiments that must account for human performance factors, for example, studies which evaluate software development methodologies or programming aptitude [12], [13], and may aid in understanding the role of language design paradigms in language acquisition, adoption, and learnability. A recent survey [14] showed both a lack of consistent measures of expertise in extant program comprehension studies, as well as the shortfalls of self-assessment data in predicting expertise.

Program comprehension studies which fail to account for programming experience may lack reproducibility. Expertise-related implications of EEG-based measures of cognitive load pose a potential paradigm shift towards reproducibility and an immediate step toward a cogent unifying definition of programming expertise. The primary contribution of this study is the direct observation of indicators of cognitive activity as it relates to expertise, as it applies to program comprehension.

III. Methodology

We used an off-the-shelf research version of the Emotiv Epoc headset, a device capable of capturing a raw EEG signal at a sampling rate of 128Hz. The device consists of 14 electrodes, providing a 14-channel subset of the International 10-20 system of electrode placement, seen in Figure 1. An additional benefit of this device is that the raw data are transmitted wirelessly, allowing the participants mobility pre- and post-experiment (the participants were discouraged from excessive movement during the experiment). The use of this low-cost EEG solution has a precedent in cognitive load measurement [9].

A. Procedure

At each session, a single participant provided informed consent, was given a brief tutorial, was fitted with the device and participated

[2]Figure reproduced from [9], with author's permission.

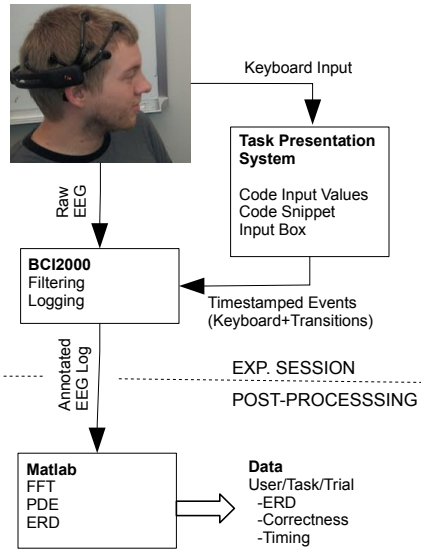

Fig. 2. Major components of the experimental toolchain, along with types of data passed between them, from raw EEG signal capture to SPSS-ready dataset.

in the experimental tasks after a brief period of calibration and adjustment. The details of each significant portion of the experimental procedure are given below. Figure 2 shows the major components of the experimental setup described in detail in the following sections.

1) Pre-Session Tutorials: Each participant was given a brief training session, consisting of instructions for behavior during the experiment, syntax of the programming language used for the presented tasks, format of responses, and interaction with the front-end software.

Since EEG is susceptible to noise from muscle movement, the participants were instructed to refrain from excessive movement by keeping arms at rest on the table in a position that allowed them to reach the keyboard without excessive movement. To further aid in limiting movement, the task presentation system was fully automated and keyboard driven. The participants did not have to intervene manually to progress the rest/task sequence, except by ending their input by depressing the enter key. Participants were also asked to try to minimize any behavior which would cause blinking or jaw clenching.

2) Experimental Sessions: The individual experimental sessions were designed to last no more than 30 minutes after the fitting of the EEG device in order to prevent excessive noise due to the drying of the saline-infused device electrodes. Impedance of each channel is tested during fitting to ensure a good signal-to-noise ratio. Finally, each participant was seated in a typical working environment (office chair, desk, and computer) within a partially shielded room with minimal distractors.

The interactive session is performed using an automated task presentation system, designed and developed specifically for this study. Throughout the course of the session, the participant is presented with one of three screens, in succession and driven only partly by the participant's responses: the rest screen (blank except for a single word "Relax" and displayed for 10 seconds at a time), task screen (no time limit, progressed by user actions) presenting code and an input box, and an answer verification screen stating that the answer was either correct or incorrect.

The task screen contains a set of input values, a snippet of code that uses these input values to produce some result, and at the

Class Level	0		1		2		3		4	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Age	21.1	2.1	23.3	9.8	21.3	1.5	22.4	3.2	24.4	6.1
Estimated Experience (1-10)	1.7	0.9	2.7	1.3	5.0	2.1	7.0	1.7	7.4	1.3
Programming Experience (Years)	0.3	0.7	0.9	0.3	4.8	2.2	2.3	0.9	4.6	1.8
CS Courses Taken	0.3	0.7	1.1	0.3	6.2	1.7	6.6	1.8	12.7	5.6
Java (1-5)	1.3	0.5	2.3	0.7	3.8	0.9	3.6	0.7	3.7	1.3
Experience vs Classmates (1-5)	2.2	1.0	2.9	1.0	3.2	1.5	3.3	0.5	3.6	0.9
Experience vs Experts (1-5)	1.0	0.0	1.1	0.3	1.5	0.5	1.4	0.5	1.6	0.5

TABLE I

PARTICIPANT DETAILS: CLASS LEVELS CORRESPOND TO THE HIGHEST-LEVEL COURSE THAT HAS BEEN COMPLETED OR IS CURRENTLY BEING TAKEN BY THE PARTICIPANT (I.E., 0 – NO PROGRAMMING, 1 – INTRODUCTORY PROGRAMMING COURSES, 2 – SOPHOMORE LEVEL, ETC.)

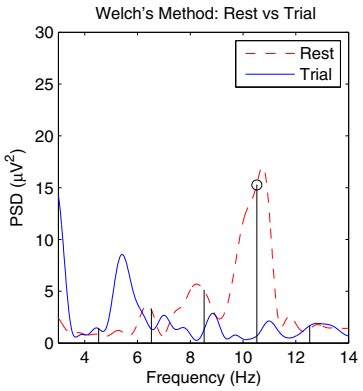

Fig. 3. An example comparison of the Power Spectral Density between the Rest and Trial states. The IAF is marked by a circle, and the resulting sub-band ranges are indicated as Theta, L1A, L2A, UA from left to right.

bottom of the screen, a text-box where the participant was to place the expected output of the code. The output text-box is always in focus, ready for their typed input, and depressing the enter key records their answer and progresses the interface to the next screen. All keyboard input was recorded and timestamped. The participants were also informed that they have the option of skipping any task by pressing the escape key. The task screen was not timed, so as to relieve any pressure that may affect a participants performance.

The automated task presentation system is integrated with the Emotiv EEG headset and BCI2000. BCI2000 records the raw signal from the Emotiv headset and the presentation system, tagging the signal with the relevant state changes (i.e., Rest to Trial state transition).

Each participant completed an expertise survey following the completion of the experimental session, similar to the one used in Siegmund et al. [14].

B. Experimental Tasks

The participants performed three tasks encoding simple concepts accessible to those with no programming experience. The programming tasks are: Task 1 – iterates across a string, in reverse, printing each character; Task 2 – computes and prints the mean of five numbers; Task 3 – doubles the values of each input integer in-place, printing comma-separated values on completion. Each task is associated with five trials, each corresponding to a unique set of input values. Trials corresponding to the same task were never presented in succession, but were rotated round-robin, until all trials were exhausted. All tasks were presented in the same order to all participants. Presenting the experimental tasks in this manner serves to reduce several potential sources of bias.

C. Signal Analysis

In post-analysis, the raw EEG signal is divided into epochs of rest and trial periods according to our timestamped annotations. A rest period is bookended by markers showing the start and end of the rest screen presentation. A trial period is marked when the trial screen is first presented and ends when the first keystroke is encountered following the onset of the trial screen.

In post-analysis we focus on four of the available 14 channels: AF3, F7, F3, and FC5. These electrode locations are clustered near Brodmann areas 8 and 46, shown in a prior study to correspond to working memory activity [15], and roughly indicated in Figure 1 by the shaded areas.

The signal was first adjusted by a band-pass filter on the range of 2Hz to 15Hz, based on expected ranges of the Theta to Alpha bands. The mean of the signals was found and subtracted from the total signal for baseline adjustment. Power Spectral Density (PSD) estimates were obtained using Welch's method. The gravity frequency, IAF, for the range 7Hz to 13Hz was found for each participant and used to calculate L1A, L2A, and UA sub-band ranges, as well as the theta range. Figure 3 shows an example Welch's PSD estimate of the Rest state versus the Trial state.

Next, the ERD values were obtained using rest and trial periods split into 125ms windows. Averaged over the windows, the rest period power is used as reference for calculating ERD of the subsequent trial period. Readings outside the bounds of $\pm 200 \mu V$ are considered errors; if errors comprise more than 20% of an epoch, the rest/trial pair is discarded.

D. Participants

Data were collected from 34 computer science undergraduates at the authors' institution (IRB 13-0218-2). Each of the individuals was enrolled in at least one computer science course and grouped according to class level corresponding to the highest numbered computer science course completed or enrolled in at the time of the experiment (see Table I for the makeup of each class level). Due to the expected familiarity with programming language constructs used in the tasks, we regard participants in class levels 0 and 1 as novices, and 2 and above as expert. Accuracies achieved by individuals from each group are shown in Table II.

	Correct Trials	Incorrect Trials	Skipped Trials
Novices	40	167	3
Experts	213	72	0

TABLE II

ACCURACY PER GROUP, SHOWING THE NUMBER OF CORRECT AND INCORRECT TASK RESPONSES FOR EACH, AS WELL AS THE NUMBER OF TASKS SKIPPED.

IV. Results

We performed an analysis across all trials for each participant, using only those trials remaining after those with high EEG error rates were discarded (more than 20% of the trial is outside of the expected -200 to 200 μV range). Our intuition tells us that over the course of 15 trials, the expert group will more quickly recognize the concept encoded by the task code, arriving at ERD levels corresponding to the task's intrinsic load, i.e., the cognitive load experienced irrespective of task encoding, more quickly than the novice group.

Collapsing the five self-reported class levels into two groups, novices and experts, we consider performance over time for each task, and measure the overall inter-trial performance by computing a regression line over each participant's task-specific ERD values. Recall that there are three tasks, with five related trials each. Once regression lines are obtained, they are clustered by slope using several thousand iterations of k-means clustering. We thereby minimize the classification distance, defined as the absolute difference of the participants' class-level and cluster assignment (class-level indicates the extent of the participants' progression through a CS curriculum, and so is ostensibly not subject to reporting bias).

Considering all trials containing both correct and incorrect responses, a random cluster assignment of the participants results in an average distance of 0.5. Perfect, or 100% accurate classification would yield an average distance of 0. Clustering the performance on Task 1 by Upper Alpha and Theta ERD results in average distances of 0.45 for both. This means that the class levels of our participants were correctly identified only 55% of the time. Clustering the performance on Task 2 by Upper Alpha and Theta ERD results in average distances of 0.41 and 0.37, meaning that in this case accuracy of novice/expert classification was 59% and 63%. Lastly, clustering the participant performance on Task 3 first by Upper Alpha then Theta ERD results in average distances of 0.41 and 0.37, respectively, meaning that class levels of our participants were accurately identified 59% of the time with Upper Alpha ERD and 63% of the time with Theta ERD.

Now, considering only those trials where we recorded a correct response from the participant, the random cluster assignment of the participants remains 0.5 and perfect assignment 0. In this case, prediction distances using Upper Alpha and Theta ERD and considering Task 1 are 0.44 and 0.37, respectively, meaning classification accuracies of 56% and 63%. For Task 2, the average distances are 0.44 for Upper Alpha and 0.26 for Theta, resulting in 56% and 74% accuracies. Finally, for Task 3, the average prediction distance using Upper Alpha ERD is 0.33, and using Theta ERD 0.37, resulting in 67% and 63% prediction accuracy.

Granted, assigning class levels 0 and 1 to the novice group and 2, 3, and 4 to the expert group is somewhat arbitrary, assuming only that the language constructs used in the tasks will be most familiar to class levels 3 and above. Adding level 2, or both levels 2 and 3, to the novice group yields similar prediction accuracies, which may be improved by more sophisticated exploratory techniques.

V. Future Directions and Conclusion

Besides the expertise measures seen here, one obvious direction for near-term future work is a study to examine the role of academic performance (GPA) in the development of expert performance. In general, we see implications for computer science education, where insight into the perceived complexity of concepts, as evidenced by cognitive load, would aid educators in evaluating the effectiveness of concept learning, tailoring concepts to particular expertise levels, and in presenting sequences of complex concepts.

Our study also opens the door to further testing in other areas of computer science. The methodology which we have applied to how a programmer perceives a programming task could be tested on studies examining expert and novice user approaches to interactions with general interfaces, like development environments, similar to extant exploratory studies [9], [16], [10].

To the best of our knowledge, ours is the first study using direct measures of cognitive load to quantify expertise from programming task performance, showing that cognitive demands differ across expertise levels. This is an important contribution to a growing body of EEG-based cognitive load studies whose successes imply that direct, objective measures of human performance with interactive tasks hold potential for augmenting future HCI studies.

References

[1] W. Klimesch, P. Sauseng, and S. Hanslmayr, "EEG alpha oscillations: the inhibition-timing hypothesis." *Brain research reviews*, vol. 53, no. 1, pp. 63–88, Jan. 2007.

[2] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis." *Brain Research*, vol. 29, no. 2-3, pp. 169–95, Apr. 1999.

[3] W. J. Ray and H. W. Cole, "EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes." *Science (New York, N.Y.)*, vol. 228, no. 4700, pp. 750–2, May 1985.

[4] O. Jensen, "Oscillations in the Alpha Band (9-12 Hz) Increase with Memory Load during Retention in a Short-term Memory Task," *Cerebral Cortex*, vol. 12, no. 8, pp. 877–882, Aug. 2002.

[5] T. A. Rihs, C. M. Michel, and G. Thut, "Mechanisms of selective inhibition in visual spatial attention are indexed by alpha-band EEG synchronization." *The European journal of neuroscience*, vol. 25, no. 2, pp. 603–10, Jan. 2007.

[6] N. R. Cooper, A. P. Burgess, R. J. Croft, and J. H. Gruzelier, "Investigating evoked and induced electroencephalogram activity in task-related alpha power increases during an internally directed attention task." *Neuroreport*, vol. 17, no. 2, pp. 205–8, Feb. 2006.

[7] W. Klimesch, "Memory processes, brain oscillations and EEG synchronization." *International journal of psychophysiology*, vol. 24, no. 1-2, pp. 61–100, Nov. 1996.

[8] J. Sweller, "Implications of cognitive load theory for multimedia learning," in *The Cambridge handbook of multimedia learning*, 2005, pp. 19–30.

[9] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. a. Preston, and C. T. Silva, "A User Study of Visualization Effectiveness Using EEG and Cognitive Load," *Computer Graphics Forum*, vol. 30, no. 3, pp. 791–800, Jun. 2011.

[10] L. George, M. Marchal, L. Glondu, and L. Anatole, "Combining Brain-Computer Interfaces and Haptics : Detecting Mental Workload to Adapt Haptic Assistance," *EuroHaptics'12 Proceedings of the 2012 international conference on Haptics: perception, devices, mobility, and communication*, 2012.

[11] K. E. Mathewson, C. Basak, E. L. Maclin, K. a. Low, W. R. Boot, A. F. Kramer, M. Fabiani, and G. Gratton, "Different slopes for different folks: alpha and delta EEG power predict subsequent video game learning rate and improvements in cognitive control tasks." *Psychophysiology*, vol. 49, no. 12, pp. 1558–70, Dec. 2012.

[12] E. Arisholm, H. Gallis, T. Dyba, and D. Sjoberg, "Evaluating Pair Programming with Respect to System Complexity and Programmer Expertise," *IEEE Transactions on Software Engineering*, vol. 33, no. 2, pp. 65–86, Feb. 2007.

[13] R. Bornat, S. Dehnadi, and Simon, "Mental models, consistency and programming aptitude," *ACE '08 Proceedings of the tenth conference on Australasian computing education*, pp. 53–61, Jan. 2008.

[14] J. Siegmund, C. Kästner, J. Liebig, S. Apel, and S. Hanenberg, "Measuring and modeling programming experience," *Empirical Software Engineering*, Dec. 2013.

[15] J. B. Rowe, I. Toni, O. Josephs, R. S. Frackowiak, and R. E. Passingham, "The prefrontal cortex: response selection or maintenance within working memory?" *Science (New York, N.Y.)*, vol. 288, no. 5471, pp. 1656–60, Jun. 2000.

[16] J. Klingner, B. Tversky, and P. Hanrahan, "Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks." *Psychophysiology*, vol. 48, no. 3, pp. 323–32, Mar. 2011.