“What Was He *Thinking*?”: Using EEG Data to Facilitate the Interpretation of Performance Patterns

Gwendolyn E. Campbell[1], Christine L. Belz[1], and Phan Luu[2]

[1] Naval Air Warfare Center Training Systems Division, 12350 Research Parkway,
Orlando, FL 32826
[2] Electrical Geodesics, Inc., Riverfront Research Park, 1600 Millrace Drive, Suite 307,
Eugene, OR 97403
{Gwendolyn.Campbell,Christine.Belz}@navy.mil, pluu@egi.com

Abstract. Previous research has demonstrated that EEG data can be used to identify and remove unintentional responses from a data set (guesses and slips). This study sought to determine if removing this error variance has a significant impact on the interpretation of a trainee's performance. Participants were taught to recognize tank silhouettes. Multiple linear regression models were built for each participant based on three sets of their data: 1) all trials of their performance data, 2) only trials that were learned according to a state space analysis, and 3) their intentional data as identified by EEG. When compared to an expert model, each of the three models for every participant yielded a different diagnosis, indicating that filtering performance data with EEG data changes the interpretation of a participant's competence.

Keywords: electroencephalography, training, and student modeling.

1 Introduction

There is a growing movement to incorporate measures of physiological and neurological data collected from warfighters into military systems. The premise is that the systems could make intelligent adaptations on the basis of those measurements, in order to increase the overall effectiveness of the warfighter [1]. Training seems to be a particularly promising field for the incorporation of these data [2]. Many researchers in this area have proposed monitoring physiological data to ensure that the trainee is being kept at an optimal level of alertness and engagement during the exercise – neither bored nor overwhelmed [3].

Of course, alternative applications of neurophysiological data in training systems have been suggested as well. This work follows up on the proposal that electroencephalography data (EEG) could be used to support the process of diagnosing a trainee's underlying competence [4]. Currently, trainers make inferences about a trainee's competence based on the pattern of correct and incorrect actions that he or she takes during an exercise. While these data are obviously highly relevant, it has long been known that performance is not a perfect reflection of competence. Some actions, for example, represent guesses (lucky or unlucky) and slips (unintentional

D.D. Schmorrow et al. (Eds.): Augmented Cognition, HCII 2009, LNAI 5638, pp. 339–347, 2009.

actions, which are more likely to occur when a person is working quickly) and thus are not representative of stable cognitive patterns.

In fact, it has been shown that EEG data can reliably discriminate between intentional and unintentional responses [5]. While these results are suggestive of the potential for neurophysiological data to support the accomplishment of training goals, many questions still remain. Of particular interest is the question of whether or not distinguishing between intentional and unintentional responses has any practical impact on the actual diagnosis of trainee performance. In other words, it has yet to be demonstrated that diagnosing only the intentional behaviors will lead a trainer to draw different conclusions about a trainee's underlying competence from what he or she would have concluded based on a diagnosis of the entire set of performance data.

We address this issue in the current study, using the Brunswik Lens Model [6] as our paradigm for diagnosing underlying trainee competence in a decision-making task. According to this paradigm, a mathematical model is derived relating trainee decisions to the characteristics of the environment or stimulus being processed. This model is interpreted as indicating which characteristics the trainee is using, and to what degree, when making an identification decision. An analogous model is built on either expert performance data or perfect performance data and the two models are compared. Discrepancies between the two models are interpreted as weaknesses in the student's strategy of using characteristics or cues to make decisions.

Our hypothesis is that using EEG data to remove guesses and slips (unintentional responses) from a trainee's performance data set will result in a different interpretation of that trainee's competence than would have been derived if the entire set of performance data had been modeled. As an additional control condition, we used a statistical technique to try to identify and remove guesses from each trainee's performance data set, to see if the EEG data had any impact over and above that which could be achieved by a simpler and cheaper methodology.

2 Method

2.1 Participants

Ten right-handed volunteers, 7 women and 3 men, over the age of 18 were recruited for this study. The mean age of the participants was 28 years (SD = 11; range: 18-48). Each received financial compensation for their participation.

2.2 Apparatus

A 256-channel HydroCel Geodesic Sensor Net (Electrical Geodesics, Inc., Eugene, OR) was used to acquire the EEG data. All recordings were referenced to Cz and all of the electrodes were kept below 70 KΩ. The EEG was bandpass filtered (0.1- to 100-Hz) and sampled with a 16-bit analog-to-digital converter at 250 s/s. Eprime© (Psychology Software Tools, Pittsburgh, PA) was used for stimulus control.

2.3 Materials

Participants were trained on identification of military vehicles in a computer-based learning program created in EPrime©. The images used were bitmap files scanned from images selected from the United States Marine Corps unclassified anti-armor training materials. They consisted of eight tank silhouettes: the ASU85, Centurion, Chieftain, Leopard, M60A1, T62, T72, and ZSU23-4. The program presented a bitmap file of the tank silhouette, allowed a fixed amount of time for a response, provided feedback, and kept a record of the stimuli presented as well as participant responses including reaction times. Trials were presented in a block randomized order so that for each consecutive eight trials each stimulus was presented once but in random sequence.

2.4 Procedure

After completing the informed consent paperwork, each participant was fitted with a 256-channel sensor array, and then began the computer-based learning task, which was divided into four stages. The first stage consisted of 120 trials divided into 15 randomized blocks of the eight tanks. The goal of this stage was to familiarize participants of the association between tank names and response keys. A target tank name was displayed in the center of the screen. Near the bottom of the screen a representation of the response keys with the tank names indicated on each key was displayed. The participant's task was to press the corresponding key as quickly as possible. Pressing the correct response key initiated the next trial in the program.

The goal of the second stage was to oblige participants to remember which response keys corresponded to each tank name. The task was identical to the task in the first stage except that the labeled keyboard display was removed. Participants engaged in seven randomized blocks of the eight tank names for a total of 56 trials. Feedback was given after each trial, and the next trial was initiated by a correct key press by the participant.

The third stage consisted of the primary learning task in which participants were asked to learn to identify the tank silhouettes. Each trial began with presentation of a silhouette in the center of the computer screen. Participants had 2000 milliseconds to identify the silhouette by pressing the appropriate response key. Immediately following the response, or if the response time ended before a key press was made, the participant was given feedback including information about the correctness of their response and the correct name of the tank. The feedback remained on the display for 2000 milliseconds, or until the participant pressed a key, upon which the computer screen went blank for 100 milliseconds and then the next trial began. This stage proceeded through 400 trials divided into 50 randomized blocks of the eight tank silhouettes.

Finally, the fourth stage consisted of the testing stage. Similar to the previous task, participants were asked to identify the tank silhouettes in a brief period of time (1000 milliseconds) by pressing the appropriate key. No feedback was provided, however, at any time during this stage. There were 32 test trials divided into four randomized blocks of the eight tank silhouettes.

Following these four stages participants filled out a standard questionnaire regarding the comfort of the 256-channel sensor array net and a debriefing questionnaire regarding the learning task. This was comprised of Likert scale ratings of both the difficulty of the learning task and usefulness of the feedback as well as questions in which participants were asked to describe the features that they used to identify the tank silhouettes and any learning strategies they may have used during the experiment.

3 Results

In preparation for modeling, the eight tank images were decomposed into a set of 7 features, for example, the ratio of the length of the gun barrel to the length of the vehicle body. This was accomplished partly through visual inspection by the authors and partly using subjective reports from pilot subjects. These features were sufficient to uniquely discriminate each of the images. The following paragraphs describe the steps used to build models. It should be noted that separate models were built for each participant and data were never combined across participants.

Each participant's data from stage three of the procedure (400 trials) were assembled into a table that contained one row per trial, and detailed the actual stimulus, the participant's response, the participant's reaction time and the values that those seven features took on for that stimulus. The analysis tool pack from Microsoft Excel ® was used to conduct a multiple linear regression on each participant's complete data set from stage three, with the constant set to zero, resulting in the first of the three models for each participant.

While logistic regression would technically have been more appropriate given the nature of the response (a vehicle name), a logical ordering of the vehicles (based on similarity) was imposed and a comparison of the two regression techniques indicated that they derived representations that were equally predictive of the participants' responses. The linear regression representation was used for this study because it provided cue weights that were easier to interpret within the context of the Brunswik Lens Model.

Next, following [7], state space analyses were applied separately to each participant's performance on each of the eight stimulus images, in an attempt to estimate the trial (if any) at which each image was reliably learned. Responses made before these learning points were discarded (as guesses) and the subset of performance data remaining was again analyzed by multiple linear regression.

The results of the state space analyses were also used as inputs to support the single trial analyses of the EEG data that was collected during stage three of the procedure. EEG-based indices were developed to discriminate between intentional (or learned) responses, guesses and slips. More details on the single trial analysis procedure that was applied can be found in [5]. Once the single trial analyses were complete, the results were used to identify and discard all of the responses that were not flagged as intentional and learned by the EEG signal. The remaining subset of data was used to generate the third model for each participant, following the procedure described above. Beta weights for each of the three models built for each participant can be found in Table 1.

In order to apply the Brunswik Lens Model paradigm, we needed to create one last model using hypothetical data from a "perfect participant." This was accomplished by using the correct stimulus images as the criterion in a regression analysis, instead of the responses given by a real participant. The beta weights and the 95% confidence intervals around those beta weights are presented in Table 2.

Table 1. Cue weights from Regression Equations Built from Three Different Subsets of Each Participant's Data

CUES	1	2	3	4	5	6	7
Participant #1							
All	0.16	3.07	0.77	-0.79	1.22	0.00	1.71
SSA	0.05	3.40	0.91	-0.93	1.21	0.00	1.89
EEG	-0.28	5.20	1.32	-1.40	1.74	-0.01	2.18
Participant #2							
All	-0.23	7.02	1.65	-1.55	3.02	-0.02	1.64
SSA	-0.37	7.17	1.70	-1.62	2.91	-0.02	1.83
EEG	-0.26	5.22	1.34	-1.39	1.75	-0.01	2.18
Participant #3							
All	0.60	2.76	0.36	-1.97	2.17	0.01	0.40
SSA	-1.02	0.00	-0.26	-0.38	-1.11	0.03	3.25
EEG	-0.21	4.98	1.26	-1.37	1.64	-0.01	2.17
Participant #4							
All	0.71	-0.67	0.28	-1.23	0.46	0.02	0.74
SSA	-0.11	-4.14	0.50	0.00	1.38	0.05	0.92
EEG	-0.27	4.95	1.29	-1.37	1.71	-0.01	2.15
Participant #5							
All	0.88	4.81	-0.57	-2.06	1.16	0.00	0.94
SSA	0.71	5.10	0.13	-1.64	0.59	-0.01	1.81
EEG	-0.30	5.09	1.32	-1.39	1.76	-0.01	2.17
Participant #6							
All	0.46	-1.11	-0.67	-1.46	-0.30	0.03	0.96
SSA	0.72	0.00	0.36	-0.73	0.00	0.01	1.64
EEG	-0.18	5.05	1.31	-1.31	1.59	-0.01	2.22
Participant #7							
All	-0.14	4.58	1.13	-1.33	1.75	-0.01	1.97
SSA	-0.11	3.92	1.18	-1.09	1.44	0.00	2.09
EEG	-0.22	4.80	1.37	-1.24	1.64	-0.01	2.19
Participant #8							
All	0.44	5.60	-0.38	-2.89	1.56	0.00	1.14
SSA	-0.07	4.92	1.04	-1.72	2.00	-0.01	1.71
EEG	-0.36	5.26	1.42	-1.38	1.88	-0.01	2.18
Participant #9							
All	0.06	-0.46	-1.18	-0.98	-0.19	0.04	1.10
SSA	0.24	0.17	-0.45	-1.07	-0.19	0.02	1.59
EEG	-0.27	4.94	1.32	-1.33	1.69	-0.01	2.18
Participant #10							
All	0.88	3.32	-0.05	-1.38	0.31	0.00	1.55
SSA	0.85	0.00	0.14	-0.47	-0.46	0.01	1.87
EEG	-0.38	5.08	1.43	-1.34	1.89	-0.01	2.16

Table 2. Beta weights and 95% confidence intervals in hypothetical perfect participant model

CUES	1	2	3	4	5	6	7
Perfect Performance Model							
Lower 95th confidence interval	-0.37	4.43	1.20	-1.44	1.43	-0.01	2.14
Beta weights	-0.25	4.93	1.38	-1.27	1.70	-0.01	2.19
Upper 95th confidence interval	-0.12	5.43	1.56	-1.11	1.97	-0.01	2.24

Table 3. Summary of beta weight comparisons between the perfect participant's model and each of the three models built per participant

	# of Beta Weights	
	β > Upper Bound OR β < Lower Bound on P.P. Model	Lower < β < Upper Bounds of P.P. Model
Participant 1		
All	7	0
SSA	7	0
EEG	0	7
Participant 2		
All	6	1
SSA	6	1
EEG	0	7
Participant 3		
All	7	0
SSA	7	0
EEG	0	7
Participant 4		
All	6	1
SSA	7	0
EEG	0	7
Participant 5		
All	6	1
SSA	6	1
EEG	0	7
Participant 6		
All	7	0
SSA	7	0
EEG	0	7
Participant 7		
All	3	4
SSA	6	1
EEG	0	7
Participant 8		
All	6	1
SSA	6	1
EEG	0	7
Participant 9		
All	7	0
SSA	7	0
EEG	0	7
Participant 10		
All	6	1
SSA	7	0
EEG	1	6

The final step was to conduct the diagnosis of each participant model. This was accomplished by comparing each of the beta weights in the participant's model to the 95% confidence intervals around the beta weights in the perfect model. If a beta weight fell outside of a confidence interval, the qualitative interpretation would be that the participant did not use information available in that cue appropriately.

For example, consider the regression equation built for participant #1 using all of his or her data. The beta weight on the second cue is 3.07. The perfect participant model shows a beta weight of 4.93, and the lower bound on the 95% confidence interval around this beta weight is 4.43. This could be interpreted as saying that, according to the participant's data, the participant is under-utilizing the information available in this cue or characteristic of the vehicle images.

Next, consider the same cue for participant #2. This participant has a beta weight of 7.02 on the second cue, which is above the upper bound of 5.43 on the 95% confidence interval around the beta weight in the perfect participant's model. It would appear that participant #2 over-relies upon information contained in this feature of the vehicle images when making his or her identification decision.

The results of this comparison are summarized in Table 3. The comparison of interest is, for each participant, the interpretation of the accuracy of his or her cue usage to make identification decisions within each of the three models. More specifically, an examination of the table reveals a high degree of overlap in the first two rows of each participant's section of the table, and a large deviation in the third row of each participant's section of the table. In other words, the model based on EEG information led to a different diagnosis of competence for every single one of the 10 participants.

4 Discussion

Currently, the measurement and analysis of electroencephalographic (EEG) signals can be a complicated, cumbersome and costly procedure. While there has been some scientific work suggesting that single trial analysis of EEG data can distinguish between intentional and unintentional responses given in a training context [5], that is only a first step towards determining the potential practical value added of using neurophysiological data in a real training setting. In this paper, we investigated the question of whether or not making this discrimination at a response-by-response level has the potential to influence a more global diagnosis of a trainee's cognitive strategy. If discriminating intentional from unintentional responses doesn't change the ultimate diagnosis of a trainee's strengths and weaknesses, then it is unlikely that this extra step is worth the required resources.

We used multiple linear regression equations to estimate the extent to which a trainee was over-, under- or appropriately using the various features of tanks to help identify them. We evaluated the appropriateness of their cue usage by comparing cue weights from their strategy regression equations to cue weights from the strategy regression equation of a hypothetical perfect participant. More specifically, we concluded that a participant was appropriately using a particular vehicle feature if his cue weight on that feature fell within the 95% confidence intervals around the perfect participant's cue weight for that feature.

In total, we built three equations for each participant. We built the first equation using all of that participant's data. Next, we applied a statistical technique to try to discriminate guesses from learned responses. The second equation for each subject was based on only the subset of responses that appeared to be learned according to this state-space analysis. Finally, we used single trial analyses of EEG data to discriminate between guesses, slips and intentional responses. The third equation for each subject was based on only the subset of responses that appeared to be intentional according to this neurophysiological analysis.

As our results clearly demonstrate, when compared to a diagnosis based on the entire set of responses given by a single participant, the statistical technique of identifying learned responses had little impact on the conclusions that a trainer would draw about the trainee's mental strategy. However, using the EEG-based filter to identify intentional responses had a dramatic impact on the conclusions that a trainer would draw for every single one of our ten participants. In each case, the diagnosis would flip from indicating that the participant was using few, if any, cues appropriately to indicating that the participant was using most, if not all, cues appropriately. Needless to say, these two sets of diagnoses would lead to very different instructional "next steps" for these trainees.

The fact that the use of the EEG filter led to the conclusion that most of the trainees were using all of the available information appropriately to make their identifications is not really surprising, given our training methodology. Remember that trainees were given the correct identification of each vehicle after every response. What this suggests is that this particular training paradigm led to accurate learning and that the trainees may have been further along that learning path than their performance data alone would lead us to believe.

While this work moves us one step closer to addressing the practical question of whether or not the incorporation of EEG-based measurement in a training system has value added, it is still not a final answer. We have demonstrated that the use of an EEG-based filter may lead to a different diagnosis of a trainee's underlying competence, however these data do not tell us if that diagnosis is, in fact, more accurate. The next step, which we are currently working on, is to see if using the EEG-based diagnosis to control the instructional response is either more effective or more efficient than relying on the trainee's entire data set.

It should be noted that there are also technical challenges that must be overcome, even if the EEG-based diagnosis does turn out to be more accurate. The method we used to conduct the single trial analysis of EEG data was largely data-driven and reasonably time consuming. To truly have practical application in a military training system, the EEG analyses would need to be automated and able to run in very-close to real-time.

Finally, of course, the fact that the EEG-based diagnoses differed substantially from the full data diagnoses for this particular training context does not guarantee that it will always have an impact. There could easily be many environments in which the use of this technology does not confer any advantage. For example, in slow moving domains that allow for a lot of deliberation before taking a single action, we would not expect to see a lot of unintentional responses that needed to be filtered out with EEG data. Similarly, in a domain that allowed operators to "undo" an accidental action, the use of EEG data to identify these slips would be overkill. Also, statistical techniques to distinguish intentional (or reliable) from unintentional (or unreliable) data are likely to be more cost effective then neurological data when there is the opportunity to collect a large enough sample of performance data from a trainee.

Despite the limitations of this study and the possible limitations on the use of this technology, we think that this work represents an important step forward towards the goal of effectively incorporating neurophysiological measurement into the assessment and diagnosis of trainee performance patterns. Our data have shown that, at least under some circumstances, the use of EEG data to filter the corresponding set of performance data can have a substantial impact on the conclusions that a trainer would draw about the trainee's underlying knowledge and competence.

References

1. Berka, C., Levendowski, D., Ramsey, C., Davis, G., Lumicao, M., Stanney, K., Reeves, L., Regli, S., Tremoulet, P., Stibler, K.: Biomonitoring for Physiological and Cognitive Performance during Military Operations. In: Caldwell, J.A., Wesensten, N.J. (eds.) Proceedings of the SPIE, vol. 5797, pp. 90–99 (2005)
2. Dickson, B., Belyavin, A.: The use of electrophysiological markers of expertise to configure adaptive training systems. In: Schmorrow, D., Nicholson, D., Drexler, J., Reeves, L. (eds.) Foundations of Augmented Cognition, 4th edn., pp. 138–144. Strategic Analysis, Inc. and the Augmented Cognition International Society, Arlington (2007)
3. DuRousseau, D.R., Mannucci, M.A., Stanley, J.P.: Will augmented cognition improve training results? In: Schmorrow, D. (ed.) Foundations of Augmented Cognition, vol. 2, pp. 956–963. Lawrence Erlbaum & Associates Inc., Mahwah (2005)
4. Luu, P., Campbell, G.E.: "Oops, I did it again": Using neurophysiological indicators to distinguish slips from mistakes in simulation-based training systems. In: Schmorrow, D. (ed.) Foundations of Augmented Cognition, vol. 2, pp. 941–945. Lawrence Erlbaum Associates Publishers, Mahwah (2005)
5. Campbell, G.E., Luu, P.: A preliminary comparison of statistical and neurophysiological techniques to assess the reliability of performance data. In: Schmorrow, D., Nicholson, D., Drexler, J., Reeves, L. (eds.) Foundations of Augmented Cognition, 4th edn., pp. 119–127. Strategic Analysis, Inc., Arlington (2007)
6. Campbell, G.E., Buff, W.L., Bolton, A.E.: Viewing training through a fuzzy lens. In: Kirlik, A. (ed.) Adaptation in Human-Technology Interaction: Methods, Models and Measures, pp. 149–162. Oxford University Press, Oxford (2006)
7. Smith, A.C., Frank, L.M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A.M., Suzuki, W.A., Brown, E.N.: Dynamic analysis of learning in behavioral experiments. The Journal of Neuroscience 24(2), 447–461 (2004)