Brain Activity Based Assessment (BABA)

Roy Stripling and Grace Chang

National Center for Research on Evaluation, Standards, and Student Testing (CRESST),
University of California, Los Angeles (UCLA)
10945 Le Conte Ave
Los Angeles, CA 90095, United States
stripling@cse.ucla.edu, gychang@ucla.edu

Abstract. Event-Related Potentials (ERP) are changes in brain activity detected using electroencephalographic (EEG) methods. One well-studied ERP is the P3b, which is generally elicited by asking participants to press a key when presented a target stimulus (e.g., "T") that is intermixed with a much more commonly presented non-target stimulus (e.g., "S"). We hypothesized that we could assess knowledge by asking participants to solve a problem then press a key when they see the correct answer in a series of (mostly wrong) answers. Early pilot testing (four participants) suggests that the P3b shows promise in this regard. In a math test, P3b responses were produced when shown correct, but not incorrect answers. In a foreign-language vocabulary test (matching picture to foreign word), P3b responses were not produced when shown correct answers prior to studying the words, but did produce P3b responses after studying. Some notable deviations in individual participants are discussed.

Keywords: Evoked Potential, Electroencephalogram, EEG, Knowledge Assessment.

1 Introduction

Assessing knowledge in learners, whether pencil-and-paper or though computer-based methods, typically involves an explicit question and answer process, where the answers are taken as evidence (for or against) learner knowledge. The scoring of such assessments may be straight forward (e.g., percent correct), may include individual test items that are weighted to account for differences in their a priori assessed difficulty, and/or may see item or overall scores adjusted based on statistical arguments that the learner was guessing on a given item (for full discussion of this approach, see the field of Item Response Theory [1],[2]). However, none of these approaches provides direct evidence that can distinguish correct answers that reflect true knowledge possessed by the learner from her guesses, or that can distinguish true misconceptions (wrong answers the learner believes to be right) from a simple lack of knowledge (wrong answers that are the result of guessing and/or that the learner knows she did not know). Empirical data suggests that the learner may, at least in some cases, lack introspective awareness of these differences, or at least, is not a reliable source of clarification[1],[2].

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 390–398, 2013.

Event-Related Potential (ERP) responses have been used as an additional source of direct evidence for possession of knowledge. ERPs are changes in brain activity that can be seen following presentations of stimuli to a person. They are measured using electroencephalographic (EEG) sensors, which detect small changes in the voltage potential on an individual's scalp. One particularly interesting ERP for this purpose is the P3b (also called the P3 and the P300). It is a transient positive shift in voltage observed from central EEG sensors, reaching its peak amplitude between 300-600 msec following presentations of "oddball" stimuli[3], [4]. It is independent of stimulus modality, but is typically stronger when the individual is consciously searching for the rare stimulus. Most P3b eliciting protocols expose participants to serial presentations of a non-target stimulus (e.g., the letter T), and ask the participant to perform a key press when they see the (much less common) target stimulus (e.g., the letter S). More complex presentations of non-targets (or distractors) also work, as long as the target is known and is relatively rare (10-20% of total stimuli presentations).

Most efforts exploiting ERP analysis in studies of learning and memory focus on gaining insight into the process of learning itself – i.e., what are the cognitive mechanisms of learning? However, a few efforts have sought the use of P3b detection as a method for knowledge assessment. These efforts have used the ERPs as evidence for word recognition, recognition of deviations of musical expectancy in experts versus novices, and for detecting "guilty knowledge" in criminal suspects.

Johnson, et al. [5] had participants study word lists, and then tested these participants for P3b elicitation during subsequent presentation of those words mixed with distractor words. They observed greater P3b amplitudes during presentation of studied words, which increased with the extent of studying permitted. Words that were correctly recognized elicited stronger P3b responses than study words that were recognized less consistently. Besson and Faita [6] studied musicians and non-musicians listening to musical phrases that were either selected from the classical repertoire or composed for the experiments. The musical phrases ended either congruously or with a musical violation. Musicians performed better than non-musicians in recognizing familiar musical phrases and classifying terminal violations. The ERPs (in this case an N400 ERP) to the end notes differed both in terms of amplitude and latency between musicians and nonmusicians, and as a function of participants' familiarity with the melodies and type of violation.

Detection of the P3b has been used (with some controversy) to determine if a criminal suspect posses knowledge of a crime that only the criminal or an investigator could know [7]. These suspects are typically shown a sequence of crime scene images. Most of the images in this sequence are not from the crime in question, but a few are. Detection of P3b ERPs in response to the images from the crime in question are taken as indicators of specific knowledge of the crime. If the suspect does not have a suitable explanation (e.g., they witnessed the crime, they investigated the crime, etc.), then these results are taken to connect them to the crime. The controversy with this approach is not whether it provides some useful information relevant to guilt or innocence; rather the controversy is related to the perfect accuracy rate claimed by its proponents [7].

These previous studies provide limited evidence that ERPs can be used to assess acquisition or possession of knowledge in some respect, but none provide a systematic exploration of the potential of ERPs in neuro-based assessments. What types of knowledge can be assessed? What form must the testing take to provide reliable valid, evidence of specific knowledge? What parameters can be manipulated without invalidating the approach? This paper provides a qualitative description of ongoing/preliminary work that is exploring whether ERPs, and in particular the P3b can be reliably used to assess possession of explicitly learned procedural and/or declarative knowledge. In the most common form of P3b eliciting experimental paradigms, P3b responses are elicited by rare target (visual or auditory) stimuli, presented as part of a series of non-target stimuli. Instead of instructing participants on what target stimulus they should search for, as is commonly done in P3 studies, we adapted this approach by presenting them with a problem and asking them to search for the correct solution in the set of answers that we presented to them serially. Our hypothesis was that by embedding the correct answer in a series of wrong answers, the correct answer (if recognized as such) would elicit a P3b response. Further, incorrect answers that the participant believes to be correct (reflecting misconceptions) will also elicit P3b responses, but both incorrect answers and correct answers that are not recognized by the participant will fail to elicit P3b responses.

2 Methods

All methods involving participants were approved by the University of California, Los Angeles (UCLA) Institutional Review Board. At the time of writing, four individuals (3 female, 1 male), age range 28-33, all fluent in English, have participated in this study.

2.1 Tasks

Each participant was asked to complete a series of 5 tasks. In each case, the participant was presented on screen instructions and told to press the space bar when they were ready to begin. They were also instructed to press the space bar when they saw the target stimulus (tasks 1 and 2) or the correct answer to the problem (tasks 3-5). Stimuli in all tasks were presented on screen for 500 msec. A single dot was displayed in the same location on screen for 2000 msec between each stimulus presentation. Participants were instructed that reaction time was not critical, but that they needed to press the space bar before the next stimulus appeared on the screen. They were also instructed to try not to blink while the stimulus was on the screen, but to blink a second or so after it went off screen. Their blinking pattern was surreptitiously observed during task 1 and feedback a reminder was provided if necessary.

Tasks 1 and 2 were replications of common P3 inducing protocols. In task one, a non-target stimulus (the letter “T”) was presented 90 times and a target stimulus (the letter “S”) was presented 10 times, randomly interspersed within the non-target

sequence, but not appearing within the first 5 presentations. Prior to beginning this task, participants were instructed to press the space bar when they saw the letter "S". In task two, participants were again instructed to press the space bar when they saw a new target stimulus (the letter "U"), which was presented a total of 10 times. But this time non-target stimuli (90 presentations total) were selected randomly from all of the other letters of the alphabet.

Tasks 3 through 5 were tests of explicit procedural or declarative knowledge. Task 3 asked participants to solve or simplify math equations. Twenty-two different problems were presented. When a problem was presented on the screen, participants were given as long as they needed to solve the problem, and then asked to press the space bar to initiate the sequential presentation of possible answers. To ensure participants focused on searching only for the correct answer, they were instructed that the answers might appear more than once, and that they should press the space bar every time they saw the correct answer. For the sequence of possible answers to each problem, one correct answer was presented (never in the first three presentations), and nine unique wrong answers were presented. Five of the wrong answers were repeated again (at random), for a total of 14 wrong answer presentations and only one correct answer presentation.

Task 4 and 5 tested participant recognition of ten common words in Pinyin (Chinese characters into Latin script). Task 4 tested their recognition of these words prior to being given the opportunity to study them, and task 5 tested them after studying them with provided flash cards. The words chosen were the Pinyin names of common animals (cat, dog, horse, pig, etc.). Each task used the same 10 words, but prompted the participant to identify them with different pictures of those animals. Likewise the flash cards included different pictures of the same animals used in tasks 4 and 5. Pre and post written tests were also given using different pictures to provide further evidence of whether the participant had prior knowledge of these words and/or had successfully learned them using the flash cards. As with task 3, the participant was presented a picture of the animal, asked to recall the Pinyin name of the animal, and to press the space bar to initiate the sequential presentation of possible answers. In this case, wrong answers were the names of the other nine animals. Answers were presented in random order. Five wrong answers were repeated again (at random), but the correct answer was not presented in the first three presentations and was presented only once.

2.2 Event Related Potential (ERP) Data Collection and Processing

Electroencephalographic (EEG) data were collected from each participant during all five tasks, using a B-Alert X10 EEG system (Advanced Brain Monitoring). The B-Alert X10 system records activity through nine sites (F3, F4, Fz, C3, C4, Cz, P3, P4, and POz) digitizing each at 256 samples per second. Event synching was achieved by processing bin files generated by the task presentation software. These files were processed in MATLAB in order to obtain the event (stimulus and response) and it's corresponding epoch and data-point. The epoch and data-point of each event was then stored in a common log file (CLF) that is processed with the .ebs file in the B-Alert

batch software. The resulting ERP outputs are time locked to the start of each stimulus presentation and is presented for 1 second (256 data-points).

Data were processed for all sites by staff at ABM who were blind to the conditions of the study, using standard methods for artifact detection and removal. Briefly, ERP waveforms that included artifact such as eyeblinks or excessive muscle activity were removed on a trial by trial basis using the B-Alert automated software. Additionally, trials with data points exceeding ±50μV were manually removed.

2.3 Quantitative/Qualitative Analysis

Quantitative analyses have not performed on the current preliminary dataset. Data collection from additional participants are ongoing and quantitative/statistical analyses will take place once the dataset is complete.

3 Results

Data were processed for the nine EEG channels recorded. However, P3b responses are attributed to central/posterior sources. For this reason, and because the data are preliminary, we report only descriptive results for POz. Cz and Fz displayed similar patterns across all subjects.

Figure 1 depicts the global average response across all participants as recorded from the POz location. Prominent P3b responses to target, but not non-target stimuli, are evident in trials from Task 1 and Task 2. Here we see peak amplitudes of approximately 20 μV ~450 msec after target stimulus presentation. Non-target stimuli peak amplitudes do not exceed 10 msec, and tend to peak closer to 300 msec post stimulus onset. Task 3 exhibits a weaker, but still evident P3b response to target stimuli. Target stimuli elicit an average response peaking at approximately 15 ~450 msec after stimulus onset. Non-target stimuli generate an average wave that is qualitatively similar to that observed in Task 2. In task 4, a P3b response is not evident to either target or non-target stimuli. Peak amplitude for either stimulus type is ~10 μV or lower and occurs ~350 msec after stimulus onset. Task 5 target stimuli, may exhibit a modest P3b response to target stimuli, but not to non-target stimuli. Target stimuli are associated with a peak amplitude of ~13 μV between 450 and 500 msec after presentation onset. Non-target stimuli are associated with a peak amplitude of less than 10 mV, with the peak occurring between 300 and 350 msec after presentation onset – qualitatively similar to non-target responses in tasks 2 and 3.

Participant variation from these averages are illustrated in figure 2. As can be seen in this figure, the first and third participants exhibit prominent P3b responses to target stimuli in task 3, while the second and fourth participant show no apparent P3b responses at all in this task. In task 5, participants three and four exhibit moderate to strong P3b responses to target stimuli. Participant 2 does not appear to produce a P3b response, but does show a very prominent negative response beginning around 500 msec after presentation onset that is selective for target stimuli. This participant also showed similar, but less intense pattern of response in Task 4 (which tested the same

stimuli but before the participants were allowed to study the words; data not shown). The first participant does not appear to produce a P3b response to target or non-target stimuli in this task.

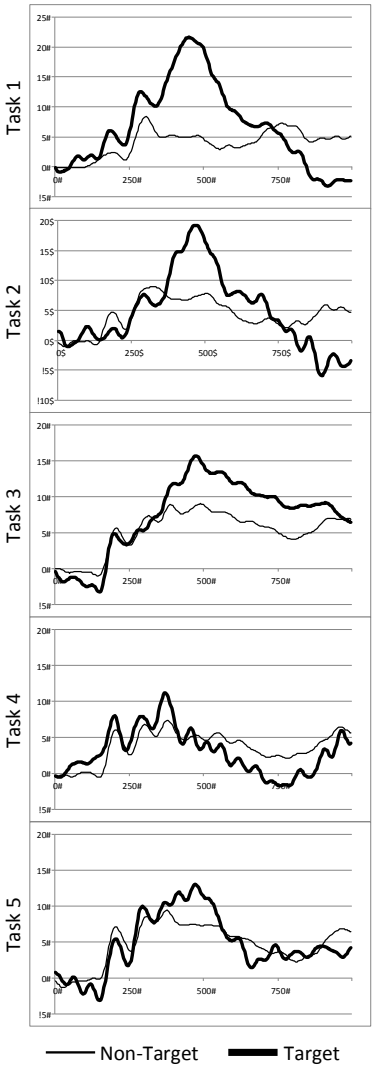

Fig. 1. Population ERP responses from each task following exposure to target and non-target stimuli. Tasks 1 and 2 replicate previous methods for inducing P3b responses. In task 3, participants were asked to solve math problems to determine the target stimuli. In tasks 4 and 5, participants were shown pictures of animals and told that the correct name for the animal in Pinyin (Chinese written using Latin characters) was their target. Participants were given the correct answers to study after task 4, but before task 5.

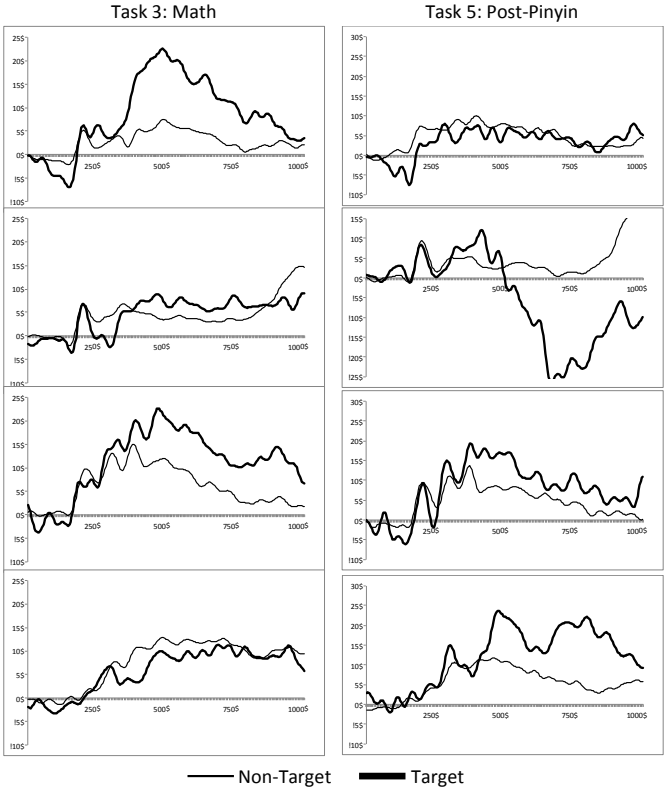

Fig. 2. Average ERP responses from individuals in Task 3 (Math) and Task 5 (Pinyin – after studying). Each row in this figure presents data from the same individual participant – averaged across their own target or non-target trials.

The first participant had no prior knowledge of Chinese, but was fluent in Korean. In discussing the tasks after the experiment, this participant indicated that while the written languages of Pinyin and Korean are very different, the spoken forms of some of the words are similar. This participant correctly answered 6 out of 10 of the words on the written pre-test. The second participant revealed some confusion about the anticipated solution format in the math task. In particular, this participant when confronted with problems that could be simplified but not solved (e.g., 2x+3x+ 5=__) assumed that the blank held a value of zero and solved the equation, rather than simplifying it. Despite this confusion, the participant selected correct answers in all but 3 of 22 problems. In addition, this participant revealed at the end of the experiment that they had prior exposure to Chinese, having taught English in rural China for a year. This participant correctly identified 4 of the 10 Chinese words in the pre-test. The third and fourth participants were given more explicit instructions with regard to solving versus simplifying problems. The third participant selected the correct answer in all but 4 problems, however the fourth participant selected 11 incorrect answers out of 22.

4 Discussion

Prior studies provide limited evidence that ERPs can be used to assess acquisition or possession of knowledge in some respect, but none provide a systematic exploration of the potential of ERPs in neuro-based assessments. This paper provides a qualitative description of ongoing/preliminary work that is exploring whether ERPs, and in particular the P3b can be reliably used to assess possession of explicitly learned procedural and/or declarative knowledge. In the most common form of P3b eliciting experimental paradigms, P3b responses are elicited by rare target (visual or auditory) stimuli, presented as part of a series of non-target stimuli. Instead of instructing participants on what target stimulus they should search for, we adapted this approach by presenting them with a problem and asking them to search for the correct solution in the set of answers that we presented to them serially. Our hypothesis was that by embedding the correct answer in a series of wrong answers, the correct answer (if recognized as such) would elicit a P3b response.

We began by establishing a baseline P3b response for each participant through tasks 1 and 2. The results of these tasks replicate prior results using similar if not the same paradigms. In addition, task 2 may have prepared the participant for our problem-solution variation by challenging them to find a specific target in a complex set of non-target stimuli. All four participants tested to date were able to discriminate the target from non-targets in tasks 1 and 2, and all generated robust P3b responses to targets, and not to non-targets in these tasks.

In tasks 3-5, we test our hypothesis by asking participants to determine what the target stimulus should be based on their knowledge of math procedures, or based on their knowledge of Chinese (written in Pinyin). We had assumed going in that our math problems were solvable by the population from which we would be recruiting, and that none of our population would have prior knowledge of Chinese. Instead, as described in the results section, we discovered that our math problems were sometimes confusing as written and our instructions on how to handle them too vague for some. This may have lead to a reduced P3b response to target stimuli, particularly for the second participant. Participant four did not produced an apparent P3b response to correct solutions in the math task, but their performance in that task suggests that this may be due to identifying incorrect solutions to the problem, in which case this lack of P3b would be consistent with our hypothesis.

We had also assumed that knowledge of Chinese language would be sparse in our participant population, but post-experiment discussions with our participants suggests otherwise. Participants one and two had partial knowledge of the words used in our tests, and both failed to generate a P3b responses in tests run before and after allowing them to study the words. The lack of any ERP response from participant one in this task does not support the hypothesis. The second participant did produce a late, negative ERP that was present in task 4 and more prominent in task 5 (post-studying). This is not consistent with our specific hypothesis that a P3b ERP should be elicited by recognized stimuli, but does suggests that other ERPs may also be a source of knowledge assessment. The particular ERP that reveals knowledge may differ based on cognitive strategies employed by the participant and/or may reflect some natural

individual variation that will have to be accounted for if ERPs are to be put to practical use in this regard.

Collectively, analysis of the data collected to date suggests that there may be potential for using ERPs, including the P3b, as a basis for knowledge assessment. The data also indicate that clear test items and unambiguous instructions are critical. In order to improve the quality and clarity of our test items, we plan to utilize math items taken from the National Assessment of Educational Progress (NAEP) database of test items. In addition, we will test alternative foreign languages to ensure that each participant is fully naïve in that task. And we will begin exploring history/civics test items also taken from the NAEP database to broaden the types of test items with which we test our hypothesis.

Acknowledgments. We thank Chris Berka and Veasna Tan from ABM for their assistance processing data for this project. This work was supported by funding from the U.S. Department of Education, award R305C080015.

References

1. de Ayala, R.J.: The theory and practice of item response theory, vol. xv. Guilford Press, New York (2009)
2. Lord, F.M.: Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Inc., Hillsdale (1980)
3. Comerchero, M.D., Polich, J.: P3a and P3b from typical auditory and visual stimuli. Clinical Neurophysiology 110(1), 24–30 (1999)
4. Polich, J.: Updating P300: An integrative theory of P3a and P3b. Clinical Neurophysiology 118(10), 2128–2148 (2007)
5. Johnson, R., Pfefferbaum, A., Kopell, B.S.: P300 and Long-Term Memory: Latency Predicts Recognition Performance. Psychophysiology 22(5), 497–507 (1985)
6. Besson, M., Faïta, F.: An event-related potential (ERP) study of musical expectancy: Comparison of musicians with nonmusicians. Journal of Experimental Psychology: Human Perception and Performance 21(6), 1278–1296 (1995)
7. Farwell, L.A., Donchin, E.: The Truth Will Out: Interrogative Polygraphy ('Lie Detection') With Event-Related Brain Potentials. Psychophysiology 28(5), 531–547 (1991)