An ERP-based Concealed Information Test for Sequential Lineups

Authors

Mario J. Baldassari, mbaldassari@tamusa.edu (corresponding author)¹

D. Stephen Lindsay, slindsay@uvic.ca

Clay B. Holroyd, clay.holroyd@ugent.be²

James W. Tanaka, jtanaka@uvic.ca

Department of Psychology, University of Victoria

PO Box 1700 STN CSC, Victoria, BC V8W2Y2

¹ Current address: Department of Science and Mathematics, Texas A&M University – San Antonio, One University Way, San Antonio, TX 78224

² Current address: Ghent University, St. Pietersnieuwstraat 33, 9000 Gent, Belgium

Abstract

Witnesses to crimes are sometimes reluctant to identify the culprit in a lineup (e.g., for fear of retribution). We introduce an ERP-based Concealed Information Test for sequential lineup identifications, using an oddball paradigm to evoke a larger P300 component when a witness sees a photo of a culprit compared to when they see a photo of an innocent familiar face. At the group level, clear differences were found between P300 amplitudes evoked by the culprit's face and the innocent face. At the individual level, the participants' waveforms were less diagnostic. This method of eyewitness assessment may prove useful if the procedure can be improved in ways that predict P300 amplitudes for individual participants. Success in this endeavor would be best applied with witnesses who recognize the culprit easily but are compelled to claim falsely that they do not.

Keywords: Memory, Eyewitness, Face recognition, Lineup, Event-related potential, P300

General Audience Summary

Witnesses to crimes are sometimes reluctant to identify the culprit in a lineup for fear of retribution, though they do recognize his face. One way to solve this real-world problem could be through analysis of brain-based responses that have previously served as indices of lying. Most of previous studies showed participants slideshows that contain infrequent “oddball” stimuli that produce a robust neurological change that is measurable through EEG. The critical element of these designs is that these oddball stimuli are only recognizable as such if the participant is “guilty” of the simulated crime. The current study builds on this work by combining this oddball method with sequential lineups (Lefebvre, Marchand, Smith, & Connolly, 2007, 2009). Our version of the method was effective at detecting lies at the group level but not at the level of individual witnesses. Future versions of this test could take steps to clarify the findings on an individual witness-by-witness basis.

Introduction

Eyewitness identification evidence is often pivotal in police investigations (Wells, Memon, & Penrod, 2006). Investigators might miss out on this pivotal piece if a witness is motivated to deny that they recognize the culprit in a lineup due to the witness's fear of reprisal, or because they wish to avoid entanglement in the prosecution of a case (Anderson, 2007). We describe a method designed to solve this problem of an uncooperative but knowledgeable eyewitness.

When in doubt about the trustworthiness of a witness's overt responses, police sometimes measure autonomic bodily responses to infer answers to critical questions. The lie detection method based on Event-Related Potentials (ERP) utilizes the P300, the component considered to be the index of the unexpectedness of a stimulus, to create a test most often called the Concealed Information Test (CIT; Allen, Iacono, & Danielson, 1992; Farwell & Donchin, 1991).

The typical CIT method (in which a suspect is shown a slideshow of crime-relevant items)³ would need to be only slightly altered to address a different real-world scenario, namely that of an uncooperative eyewitness. The most likely scenario for application of such a paradigm is a case in which the witness recognizes the culprit but has some reason to wish that the culprit not be caught (e.g. the witness is threatened by associates of the culprit, or the witness is a co-conspirator). The CIT in this case would seek to reveal recognition of a suspect in such a witness, thus the procedure would be filled with photographs of faces of people who matched the description of the suspect.

³ See Appendix A for a hypothetical implementation of the CIT, intended for newcomers to the method.

Some researchers have applied the ERP-CIT⁴ to detect concealed face recognition in classic study-test paradigms (Sun, Chan, & Lee, 2012; Treese, Johansson, & Lindgren, 2010) and in lineups (Friesen, 2010; Lefebvre, Marchand, Smith, & Connolly, 2007, 2009). A model study testing the ERP-CIT with lineups used four crime videos depicting four different perpetrators committing the same crime (each with a different victim) and experimented with varying time delays between video and lineup presentation (Lefebvre et al., 2007). The lineup consisted of six faces repeated 40 times, one of whom was the victim (whose inclusion was not important to the procedure, but functioned as an attention check). Each participant completed one *culprit-present* (CP) and one *culprit-absent* (CA) lineup followed by an immediate test, and a CP lineup followed by a test after a 1 hour delay and after a 1 week delay.

The data mostly showed small effects (and some only when collapsed across electrodes and time delay conditions), likely due to the difficulty of the task as reflected in low identification rates overall. Additionally, ERP's were not much more informative than participants' confidence ratings in differentiating correct ID's from false rejections for lineups in which the actual culprit was present (CP lineups). This is possibly because a participant must have a strong memory of the critical item for it to register as an oddball and elicit a P300. In 2009, the same authors conducted a replication study with an additional condition in which participants were instructed to lie by indicating that they did not recognize the culprit (Lefebvre et al., 2009). They employed two competing classification algorithms at the individual witness level, both of which enabled

⁴ As the reaction time based CIT is referred to as the RT-CIT, we will identify the ERP based version here as the ERP-CIT.

identification of culprit photographs from every participant in the truth group and most in the lie group.

These studies by Lefebvre and colleagues presented evidence that eyewitness ID accuracy can be assessed using a repeated sequential lineup method to elicit a P300 to the culprit. In their studies, participants who correctly identified the culprit with confidence tended to also show an increased P300 to the culprit. While groundbreaking, their work had a few limitations. First, Lefebvre et al. (2007, 2009) showed participants two videos, which was efficient but less ecologically valid than showing participants a single crime video. Second, witnesses saw the culprit for only 15s, which likely led to weak memory and smaller P300 amplitudes to the culprit. Third, the other members of the lineup included an image of the victim, which could confuse the participant-witness, especially if the victim matched the description of the culprit. Fourth, it could also be possible that splitting the identification decisions into three types (culprit, victim, new face) enabled lying participants to ignore the culprit, as participants never used the response key to which the culprit was assigned.

Current Experiment

Recall that the basic goal of an ERP-CIT lineup procedure is to uncover evidence of memory of the culprit's face in the participant/witness. In the real world, detecting such evidence would only be useful to the criminal justice community if it were applicable in situations in which the culprit would not otherwise be identified by the witness's overt decision or confidence level. Since the CIT only produces effects with strong evidence of oldness in memory, the most likely scenario for application of this paradigm is with witnesses who recognize the culprit easily but are motivated to falsely

claim that they do not. We aimed to recreate these conditions through the following application of a CIT design.

This CIT lineup contained three types of faces: the culprit, the known-innocent, and the fillers. The culprit item was the face of the culprit who committed the crime. The known-innocent item was a face the participant attempted to memorize separately from the crime video (in a front-facing photograph). This face matched the general description of the culprit, and served as a control condition against which to compare the ERP evoked by the culprit. The filler faces all match the description of the culprit, analogous to the slideshow of knives discussed earlier. When a participant sees either the culprit or the known-innocent face, these faces should stand out from the otherwise-homogeneous set of similar looking people (as oddballs) and elicit a P300.

Our study addressed the limitations of the work by Lefebvre and colleagues as follows. First, it more closely mirrored real-world witnessing situations by showing each participant only one video. Second, in order to evoke a larger P300, the video afforded participants ample opportunity to view the culprit, and participants were warned before seeing the video that they would later be asked to identify the culprit from a lineup. Further, in order to make the culprit a less frequently appearing member of the lineup, the current procedure included more filler faces than the Lefebvre procedure did. Third, as a control stimulus, rather than using an image of the victim from the crime video, we instead familiarized participants with a known (innocent) face after the video. Fourth, to avoid a scenario in which a witness would never use the response key for the culprit, we instructed participants to respond to both faces using the same response key.

Method

Participants. Participants ($N = 51$) were recruited through the University of Victoria psychology participation pool of undergraduate students who were compensated with optional bonus points. Participants reported an average age of 23.34 years, with a range from 18 to 43. Twelve self-reported as male, 36 as female, 4 as left-handed and 44 as right-handed. None reported regular seizures or recent brain injury.

Materials. We made two versions of a 41 s video depicting a simulated crime. The two versions differed only in the culprit. Both culprits were young, Caucasian men with light-brown hair and medium build. In both videos, a man enters an office and takes keys from a desk and is then shown walking in a parking lot and finding and unlocking a car with the stolen key. The video then shows him entering the car and starting it. In total, the man's face is directly facing the camera for about 17.5 s in one video and 16 s in the other. The video was recorded and displayed in high definition [<https://osf.io/b8tk9/>].

Photographs for the lineup were taken from a database of photos collected using a single camera for a previous study. Each student's face was captured looking directly into the camera in identical lighting conditions while wearing a black smock over their clothing⁵. This database included a photo of one of the culprits in the two videos; a photograph of the second culprit was taken and edited until it closely matched the characteristics of those already in the database. Another 15 faces of men who matched the description of the two culprits were pulled from the database. These were used for a

⁵ We do not have consent to share photos in publication form but have shared them via the Open Science Framework [<https://osf.io/b8tk9/>].

set of simultaneous lineups as a manipulation checks to ensure that the participants remembered the culprit.

Procedure. Participants were seated in front of a 19-inch computer monitor in an electromagnetically shielded booth. They read about the specifics of the procedure and completed a consent form. Participants watched one of two crime videos passively, and then were asked to narrate the same video aloud while watching it a second time. Each video depicted a young male of the same physical description commit a car theft in almost-identical scenes. Before the first showing of the video, participants were instructed to pay attention to the culprit "Because we will ask you to identify him later."

To create a *known-innocent* member in the forthcoming lineup, participants then studied a photograph of the face of a man named "Chris" for as long as they liked on the computer screen. For each participant, the known-innocent was the culprit from the video they did not see; thus he matched the same description. Then, in 24 yes/no recognition memory practice trials, they attempted to identify Chris. Chris appeared 4 times among 5 description-matched filler faces (which also appeared 4 times each). If a participant did not achieve 80% correct decisions, the study/test cycle was repeated.

In the next phase, the lineup, a set of 10 new faces matching the description of the culprit (filler faces) were combined with the known-innocent (Chris) and the culprit into a block of 12 trials. Participants were exposed to 30 blocks while EEG data were recorded. The blocks contained lists of 12 trials in 6 different pseudo-random orders, thus each participant was exposed to the culprit 30 times, the known-innocent 30 times, and a set of 10 fillers 30 times each. Face order was pseudo-randomized to mitigate the risk of muting EEG amplitude for a rapid second presentation of the same image

(Schweinberger, Pickering, Jentzsch, Burton, & Kaufmann (2002); Trenner, Schweinberger, Jentzsch, & Sommer, 2004). Although faces did not repeat within a set of 12 trials, the 6 random orders were edited to prevent against the culprit or the known-innocent faces appearing as both the last face in a set of 12 and the first face in the next set of 12. Thus, familiar faces were not repeated without at least 2 intervening unfamiliar faces. Participants were told that there were two faces to recognize, the known-innocent (who they knew as Chris) and the culprit, and to press a response key every time they saw either known face (the same response key was used for both familiar faces). Response keys were the 'm' and 'z' keys on the keyboard, participants used the other response key to mark a face as unfamiliar, and the category of each response key was counterbalanced between participants. Each trial began with a fixation cross, with a varying inter-stimulus interval between the cross onset and the face onset of 650-850ms. Faces appeared on the screen for 1.5 s with the response options "Known Face" and "New Face"; a gray screen appeared in place of the face and was, in turn, replaced by a new fixation cross after 1s. Reaction times were measured from the time a face stimulus appeared on the screen to the time the response key was pressed.

Half of participants were randomly selected to be placed in the lie group and were instructed to lie during the lineup task just before the lineup phase began. They were told "Some witnesses may not want to identify the criminal for various reasons. Sometimes they are threatened by associates of the criminal, sometimes they do not trust police enough to want to help, and other times they may be secret co-conspirators. We want you to pretend to be one of these witnesses. Thus, when you see him, you will not press the button to identify him." This group was instructed to only identify the known-innocent

and to pretend to not recognize the culprit by categorizing him as a “New face.” For the truth group, the other half of participants were given no special instructions, providing a control group.

Upon completion of 30 cycles through the 12 faces, participants were presented with a traditional simultaneous lineup for the known-innocent face and a separate lineup for the culprit (both lineups contained their intended target). Participants in the lie group were asked to not lie any longer. Thus, these identifications served as confirmation that they did, in fact, recognize both faces during the previous task. Confidence ratings were also collected at this stage.

ERP data acquisition and analysis. The EEG was recorded using 41 electrode sites organized according to the extended international 10-20 system (Jasper, 1958). Signals were acquired using Ag/AgCl ring electrodes mounted in a nylon electrode cap with an abrasive conductive gel on scalp electrodes and a non-abrasive conductive gel on eye, face, forehead, and mastoid electrodes (EASYCAP GmbH, Herrsching-Breitbrunn, Germany). Signals were amplified by a low-noise differential amplifier with a frequency response of DC 0.017-67.5 Hz (90dB-octave roll-off) and digitized at a rate of 250 samples per second. Digitized signals were saved using Brain Vision Recorder software (from Brain Products GmbH, Munich, Germany). Electrode impedances were maintained below 20k Ω . One electrode was placed on each of the left and right mastoids, and the EEG was recorded using the average reference. The electrooculogram (EOG) was recorded for later artifact correction: horizontal EOG was recorded from the external canthus of each eye, and vertical EOG was recorded from the suborbit of the right eye and electrode channel Fp2.

Postprocessing and data analysis were conducted using Brain Analyzer software (also from Brain Products). The digitized signals were filtered using a fourth-order digital Butterworth filter with a passband of 0.10-20Hz. Trials were segmented by condition (culprit, known-innocent face, filler faces). A 1600 ms epoch of data extending from 100 ms prior to and 1500 ms following the onset of each face stimulus was extracted from the continuous data file for analysis. Ocular artifacts were corrected using the eye movement correction algorithm describe by Gratton et al. (1983). The EEG data were re-referenced to averaged mastoid electrodes and were baseline corrected by subtracting the mean voltage associated with each electrode during the 100 ms interval preceding stimulus onset from each sample. Muscular and other artifacts were removed using a $\pm 50 \mu\text{V}$ step threshold as a rejection criterion.⁶ ERP's were then created for each electrode and participant by averaging single-trial EEG according to condition (culprit, known-innocent, and fillers). Culprit and known-innocent ERP conditions contained up to 30 trials each, which varied slightly when trials were removed due to artifacts (to reiterate, the data of participants with fewer than 24 usable trials with either face were discarded). The filler ERP condition contained up to 300 trials per participant, the number of which also varied due to artifacts in the data.

The P300 was calculated at channel Pz, where it typically reaches maximum amplitude (Fabiani, Gratton, Karis, & Donchin, 1987). In these data, the P300 reached maximum amplitude in different locations across conditions, groups and conditions. However, we confirmed that in every condition and group that the P300 was not maximal

⁶ EOG channels were not included in artifact rejection.

at channel Pz, its value at the maximal channel (which varied between P4 and PO8) was nevertheless not statistically different from the value at Pz (see P300 Maximal Channel Report on OSF). When comparing waveforms across participants, the time course of the P300 may be such that it reaches maximum sooner for a particular person and later for another. Thus, the P300 was defined as the average ERP voltage between 450ms and 550ms, as the maximum voltage occurred between these two time points for all participants.

Bayesian Hypothesis Tests. For groupwise and individual comparisons, we performed Bayesian hypothesis tests comparing participants' average identification accuracy, median reaction time, and average P300 amplitudes across conditions and groups (using the BayesFactor package in r, which performs tests with prior odds of an effect set at 1). Bayesian hypothesis tests produce a statistic called the Bayes Factor, which can be interpreted as a probability ratio: values greater than 1 reflect the probability that the alternative hypothesis is true (given the data), while values smaller than 1 reflect the reciprocal probability that the null hypothesis is true (Rouder et al., 2009). We first ran a Bayesian ANOVA, which tests models of one main effect, the other main effect, two main effects, and two main effects with an interaction each against a basic model. We then conducted a one-on-one model comparison, in which the model with the most evidence in favor was compared to the next-best model. Bayesian t-tests were then conducted as post hoc tests for specific differences between groups and conditions where appropriate.

Individually, if a participant's mean P300 amplitudes to the known-innocent and to the culprit were not different (i.e. the test resulted in a Bayes Factor smaller than 1)

and that participant's mean P300 amplitude to the culprit was greater than their mean P300 amplitude to the fillers (i.e., a Bayes Factor larger than 1), then that participant was classified as having wave forms that revealed memory for the culprit. By contrast, if a participant's mean P300 amplitude to the known-innocent was greater than their mean P300 amplitude to the culprit and that participant's mean P300 amplitude to the culprit and to the fillers were not different, the data would represent a participant or group who either did not remember the culprit's face or successfully lied and fooled the test. If a participant's data matched this second signature, we conducted a third test to determine whether their mean P300 amplitude to the known-innocent differed from their mean P300 amplitude to the fillers, as a difference between these two values could enable successful detection of the culprit under less-strict guidelines. No difference between any conditions would indicate that the test was not robust for that particular participant (due to any of a series of possibilities, including simply noisy data). Cohen's d is reported alongside Bayes Factors for the Bayesian t-tests of the ERP data.

All data and R scripts are available on the Open Science Framework (OSF) at <https://osf.io/b8tk9/>. Hypotheses were pre-registered at <https://osf.io/hrktf>.

Results

Behavioral data. Data from participants who did not correctly identify either the known-innocent or the culprit from the simultaneous lineups were removed from analyses ($N = 6$ in lie group, 1 in truth group). We also dropped data from participants who failed to identify the culprit or the known-innocent in more than 20% of the lineup phase ($N = 2$). Thus, the set consists of data from 42 participants for analysis.

During the training phase for the known-innocent lineup member, performance was near perfect (99.6% accuracy). Average reaction time for all correct responses was 676 ms. Average RT for correct identifications of both culprits and known-innocents was 700ms, while average RT for correct rejections of fillers was 670 ms.

Across the 360 trials in the lineup phase, 21 participants made fewer than 3 errors, 14 participants made between 4 and 8 errors, and 4 participants made more than 15 errors. Overall accuracy was 98.3%, and the mean false alarm rate on fillers was 0.10%. In the truth group, the mean rate of identifications of the culprit was 98.39% and the mean rate of identifications of the known-innocent was 98.25%. In the lie group, the mean rate of identifications of the culprit (contrary to the instruction to lie) was 7.55% and the mean rate of identifications of the known-innocent was 94.21%. A 2x3 within-participants Bayesian ANOVA (group x condition) model comparison preferred a model with two main effects and an interaction above the next best model (only a main effect of condition) with a Bayes Factor (BF) of 346.44.

For participants in the truth condition, a Bayesian t-test comparing accuracy in identifying the culprit and the known-innocent face indicated over 9 times as much evidence in favor of a difference as that in favor of no difference, $BF = 9.59$. A Bayesian t-test comparing accuracy in identifying the known-innocent and the fillers also showed evidence of a difference between the conditions, $BF = 30.44$. There was evidence against a difference between accuracy on trials showing the culprit and trials showing fillers, $BF = .80$. In the lie group, participants' average accuracy in identifying the culprit and the known-innocent face showed evidence against a difference, $BF = .31$, very little evidence in favor of a difference between average accuracy identifying the known-innocent and the

fillers, $BF = 2.73$, and some evidence in favor of a difference between average accuracy identifying the culprit and the fillers, $BF = 18.11$.

Error trials in the lineup phase were removed for RT analysis (see Figure 1, figure captions follow manuscript text). We report a group average of individual medians in order to reduce the influence of outlier trials. In the truth group, participants responded about 55 ms slower to both the culprit and known-innocent faces than to the filler faces [average medians of 605.2, 605.0, and 549.3 ms, respectively]. A 2x3 within-participants Bayesian ANOVA (group x condition) model comparison preferred a model with two main effects and an interaction above the next best model (only a main effect of condition), $BF = 160425130$.

Paired samples Bayesian t-tests were conducted to discover specifics of the effects. For participants in the truth condition, a Bayesian t-test comparing RT to the

culprit and the known-innocent face resulted in a Bayes Factor less than one, indicating no difference between the conditions, $BF = 0.41$, and a test comparing RT to the culprit and the fillers showed strong evidence of a difference between the conditions, $BF = 270.24$. Participants in the lie group produced a similar pattern, but the data were more variable [average medians of 612.2, 639.1, 523.4 ms, respectively]. A Bayesian t-test comparing RT to the culprit and the known-innocent face resulted in a Bayes Factor less than one, indicating no difference between the conditions, $BF = 0.30$, and a test comparing RT to the culprit and the fillers showed some evidence of a difference between the conditions, $BF = 30.26$.

P300. Application of the exclusionary rules resulted in the loss of more than 20% of lineup trials for 3 additional participants; their data were not further analyzed. Of the 14,040 trials seen by the 39 remaining participants, 215 were removed (1.5%) because participants responded incorrectly to the stimulus (1.6%). Of the remaining 13,825 trials, 79 were removed based on the artifact rejection algorithms (0.6%). We calculated the average amplitude of the P300 at Pz out of the remaining trials for all participants in our predesignated window.

A 2x3 within-participants Bayesian ANOVA (group x condition) model comparison preferred a model with main effects of condition and group above the next best model (two main effects and an interaction), $BF = 5.37$. Paired samples Bayesian t-tests were conducted to discover specifics of the effects. For participants in the truth group, a Bayesian t-test of the difference between P300 amplitudes evoked by the known-innocent and the culprit resulted in weak support for the null hypothesis [$BF = 0.305$, Known-innocent mean = 14.9 μV , culprit mean = 14.7 μV , Cohen's $d = 0.31$]. There was

strong evidence in the data that amplitudes evoked by the culprit were different from those evoked by filler faces [$BF = 1,048,000$, filler mean = $4.8 \mu V$, Cohen's $d = 5.15$], see Figure 2.

For participants in the lie group, a Bayesian t-test of the difference between P300 amplitudes evoked by the known-innocent and the culprit again resulted in weak evidence of no difference between the conditions [$BF = 0.32$, known-innocent mean = $20.6 \mu V$, culprit mean = $20.9 \mu V$, Cohen's $d = 0.22$]. Amplitudes evoked by the culprit were different from those evoked by filler faces [$BF = 90.86$, filler mean = $8.5 \mu V$, Cohen's $d = 2.41$], see Figure 3.

All these effects maximized in the parietal lobe along the midline as expected for the culprit and the known-innocent, but data were somewhat less clean for the fillers with maximization showing slight lateralization for participants in both the truth and lie groups (which is to be expected given the low amplitude of the signal, see Figure 4). Subtracting foil averages from culprit and known-innocent averages showed a clear maximization of the difference in the parietal lobe along the midline for truth tellers, but there was a bit of lateralized occipital difference for liars, which was also likely driven by the low amplitude of the foil signal and the fact that P300 amplitudes were larger for truth tellers than for liars (see Figure 5).

Individual Bayesian t-Tests. Dependent-measures Bayesian t-tests for differences between mean amplitude evoked by the culprit and mean amplitude evoked by the known-innocent, combined with tests for differences between the culprit and fillers enabled a test to determine whether a participant's information about the culprit was concealed.⁷ If a participant's mean amplitude to the known-innocent and the culprit was not statistically different but that participant's mean amplitude to the culprit was larger than their mean amplitude to the fillers, that person's identification was considered to have been successfully detected by the ERP-CIT lineup task data. Table 1 shows that for 9 of 18 participants asked to lie and 16 of 21 participants asked to tell the truth, mean

⁷ Graphs of mean amplitude in the window 450-550 ms after face onset by participant can be found on the OSF, osf.io/b8tk9, in the Individual Figures folder of the Files section.

amplitude 450-550ms after face onset to the culprit was different from fillers and not distinguishable from the known-innocent. Widening the time window to 450-650ms and 450-1000ms did not improve detection rates. Of the participants whose results did not enable detection of the culprit (the remaining 14), 6 participants produced three waves that were all statistically different from each other.

Discussion

Group-wise effects confirmed the hypothesis that P300 amplitudes evoked by the culprit were equal to those evoked by the known-innocent, and larger than those evoked by the filler faces, in both the truth and lie groups (amplitudes and differences were smaller in the lie group). Amplitudes evoked by the culprit also peaked later than those evoked by either other condition, which was not reflected in the behavioral data (in which average median RT's to the culprit did not differ from those to the known-innocent in both groups). The procedure was further validated in that the P300 amplitudes evoked by the known-innocent were statistically indistinguishable from P300 amplitudes evoked by the culprit in both groups.

Using Bayesian t-tests to classify participants' mean ERP amplitude in a 100 ms window enabled 50% successful differentiation between culprit trials and filler trials for participants in the lie group (about the average rate of successful lie detection by laypeople) and 76% successful differentiation for participants in the truth group. These individual participant datasets showed a difference between P300 amplitude evoked by the culprit and P300 amplitude evoked by fillers as well as no difference in P300 amplitude evoked by the culprit and the known-innocent. Two participants in the lie group (#'s 5 and 22, see Table 1) showed a larger P300 to the known-innocent and

statistically identical P300 amplitudes to the culprit and the fillers. These were the only participants who exhibited the data signature of a liar who fooled the test, as such a data pattern would be identical to a person who had never seen the culprit. Thus, most of the unsuccessful differentiations were likely due to noisy data rather than to participants lying successfully.

The current findings extend the results of Lefebvre and colleagues (2007; 2009), in which they distinguished correct identifications from false positives as well as lies from true lineup rejections. An important difference between the current study and those by Lefebvre is the organization of the identification task. Lefebvre et al. asked participants to use different response keys for the culprit, the victim, and the fillers, enabling participants to completely ignore the entire category of the culprit when instructed to lie. We grouped the culprit and the known-innocent together into a “known faces” group that required the same key response, preventing participants in the lie group from ignoring a response category that could have dulled the salience of the culprit’s photograph during the procedure. The current study also used a longer identification task with 10 filler faces compared to Lefebvre et al.’s 4. Both of these differences likely contribute to the larger amplitude of the P300 found in our study (Lefebvre et al.’s peaks were around 10mV), as the culprit was a less-frequent oddball and participants could not ignore the “known face” response as they could ignore the “culprit” response in Lefebvre’s studies. Their work was also missing a condition in which participants would recognize the culprit and were asked to do so, as they were in our truth group. Thus, our redesigned ERP-CIT lineup is a useful method of identification for witnesses who are unwilling to identify the culprit despite having high levels of familiarity with the person.

Limitations and future directions. As currently designed, this ERP-CIT lineup enabled classification of the culprit for half of participants in the lie group; conversely, only 2 of 18 participants produced waveforms that might have successfully fooled the test (with equivalent amplitudes evoked by the culprit and the fillers, and larger amplitudes evoked by the known-innocent: participants 5 and 22 in Table 1). This detection rate is as good as that of undergraduates attempting to detect deception (Bond & DePaulo, 2006), and with small methodological tweaks the ERP-CIT lineup could likely be improved to exceed human lie detection abilities. As the test failed to identify the culprit in three participants in the truth group, it seems to be conservative in its current form. An important next step is to investigate whether this procedure is robust against false positives. A condition in which participants are not exposed to the culprit at all would enable conclusive testing with regard to the test's decision criterion, as would a condition with a culprit-absent lineup. Continuing to improve on these procedures with the suggestions herein and investigating boundary conditions will undoubtedly result in reliable tests that could inform both police practice and basic memory research.⁸

Legal issues. These situations are fraught with legal and ethical issues when one considers their real-world implementation (Farah, Hutchinson, Phelps, & Wagner, 2014, deal with these issues in reference to lie detection with fMRI). As refusing to identify a known culprit is typically treated as obstruction during an investigation and perjury during a trial, we do not wish that the ERP-CIT lineup designed here be used as a tool to detect perjurers. Likewise, a witness whose ERP-CIT lineup results provided evidence

⁸ For discussion of additional robustness through combination of participants or measures, see Appendix B.

that put away a connected culprit might still be in danger of retribution from the culprit's associates. If police had reason to believe the witness was exposed to the culprit for a lengthy amount of time but the witness was still nervous about getting the ID wrong, this test would perhaps be at its most useful in a philosophically unambiguous situation. The test would be best deployed in scenarios in which witnesses might not identify the culprit but are not purposefully lying.

Some more rare situations would also benefit from implementation of this idea. An ERP-CIT lineup could, in theory, enable a witness suffering from locked-in syndrome to identify the culprit, as there is no need to speak or move. Further, some studies suggest that a P300 can be evoked by subliminal visual oddball stimuli (e.g. Bernat, Shevrin, & Snodgrass, 2001), thus this procedure might also be useful when a witness or victim who might be negatively emotionally triggered by presentation of the culprit's face could be spared that additional trauma.

Author's contributions

MJB conceived, developed, and conducted the procedures with input from DSL, JWT, and CBH. CBH maintained EEG facilities and provided trained RA's, as well as training for MJB. DSL provided funding to access EEG facilities and infrastructure to recruit and train RA's. MJB wrote the initial draft of the manuscript, and all authors discussed the manuscript in depth over many edits.

Acknowledgements

The authors wish to acknowledge the contributions of Ryann Tansey and the rest of the large team of research assistants during data collection. Thanks also to Sepideh Heydari, whose help creating scalp maps and testing P300 peaks at different electrode sites was essential to the final product.

Math Formulae

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are available on the Open Science Framework, [<https://osf.io/b8tk9/>]. Hypotheses were pre-registered at osf.io/hrktf.

Ethics approval and consent to participate

This study was approved by the University of Victoria Human Research Ethics Board, which is accredited and audited by the Canadian Tri-Council of research ethics and funding.

Competing interests

The authors declare no competing interests.

Funding

These studies were funded by a grant from the Natural Sciences and Engineering Research Council of Canada to D. Stephen Lindsay.

PREPRINT

Figure Captions

Figure 1. Average of individual participants' median RT. Error bars represent 95% confidence intervals.

Figure 2. Groupwise ERP average waveforms for participants in the truth group with individual participant averages. 95% confidence ribbons around ERPs are basic nonparametric bootstraps without assuming normality (See osf.io/b8tk9/ for r code and data).

Figure 3. Groupwise ERP average waveforms for participants in the lie group with individual participant averages. 95% confidence ribbons around ERPs are basic nonparametric bootstraps without assuming normality.

Figure 4. Scalp distributions of the P300 averaged across the 450-550 ms window by group.

Figure 5. Difference between culprit to foil and known-innocent to foil contrasts in scalp distributions of the P300 averaged across the 450-550ms window by group.

References

- Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The identification of concealed memories using event-related potential and implicit behavioural measures: A methodology for prediction in the face of individual differences. *Psychophysiology*, 29(5), 504-522. doi:10.1111/j.1469-8986.1992.tb02024.x
- Anderson, J. (2007). Gang-related witness intimidation. *National Gang Center Bulletin*, 1. Retrieved from <https://www.nationalgangcenter.gov/content/documents/gang-related-witness-intimidation.pdf>
- Bennington, J. Y., & Polich, J. (1999). Comparison of P300 from passive and active tasks for auditory and visual stimuli. *International Journal of Psychophysiology*, 34(2), 171-177. doi:10.1016/S0167-8760(99)00070-7
- Bernat, E., Shevrin, H., & Snodgrass, M. (2001). Subliminal visual oddball stimuli evoke a P300 component. *Clinical Neuropsychology*, 112, 159-171. doi:10.1016/S1388-2457(00)00445-4
- Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review*, 10(3), 214-234. doi:10.1207/s15327957pspr1003_2
- Caharel, S., Poiroux, S., Bernard, C., Thibaut, F., Lalonde, R., & Rebai, M. (2002). ERPs associated with familiarity and degree of familiarity during face recognition. *International Journal Of Neuroscience*, 112(12), 1499-1512. doi:10.1080/00207450290158368
- Donchin, E. (1981). Surprise! ... Surprise?. *Psychophysiology*, 18(5), 493-513. doi:10.1111/j.1469-8986.1981.tb01815.x

- Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology*, 62(7), 1219-1220.
doi:10.1212/01.WNL.0000118297.03161.B3
- Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, 44, 576–85.
doi:10.1016/j.neuropsychologia.2005.07.001
- Eimer, M., Gosling, A., & Duchaine, B. (2012). Electrophysiological markers of covert face recognition in developmental prosopagnosia. *Brain: A Journal Of Neurology*, 135(2), 542-554. doi:10.1093/brain/awr347
- 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. In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 2, pp. 1–78). Greenwich, CT: JAI Press.
- Farwell, L. A. & Donchin, E. (1991). The truth will out: Interrogative polygraphy (“Lie detection”) with event-related brain potentials. *Psychophysiology* 28(5), 531-547. doi:10.1111/j.1469-8986.1991.tb01990.x
- Friesen, K. B. (2010). Electrophysiological correlates of correct and incorrect eyewitness identification: The role of the N250 and P300 in real-world face recognition. Masters’ Thesis completed at the University of Victoria: Victoria, BC, Canada.
- Georgiadou, K., Chronos, A., Verschuere, B., & Sauerland, M. (2019). Reaction time-based concealed information test in eyewitness identification is moderated by picture similarity but not eyewitness cooperation. *Psychological Research*, online

publication. [doi:10.1007/s00426-018-1139-8](https://doi.org/10.1007/s00426-018-1139-8)

Gill, O., & Polich, J. (2002). P300 stimulus sequence effects in children and adults.

Perceptual and Motor Skills, 94, 509-520. doi:10.2466/PMS.94.2.509-520

Gordon, I., & Tanaka, J. W. (2011). The role of name labels in the formation of face representations in event-related potentials, 5, 3280–3293.

<https://doi.org/10.1111/j.2044-8295.2011.02064.x>

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography & Clinical Neurophysiology*, 55(4), 468-484. doi:10.1016/0013-4694(83)90135-9

Guillem, F., Bicu, M., & Debruille, J. B. (2001). Dissociating memory processes involved in direct and indirect tests with ERPs to unfamiliar faces. *Cognitive Brain Research*, 11(1), 113-125. doi:10.1016/S0926-6410(00)00070-7

Guillaume, F., & Tiberghian, G. (2013). Impact of intention on ERP correlates of face recognition. *Brain and Cognition*, 81, 73-81. doi:10.1016/j.bandc.2012.10.007

Jasper, H. H. (1958). The ten twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10, 371-375.

Johnson, R., & Donchin, E. (1980). P300 and stimulus categorization: Two plus one is not so different from one plus one. *Psychophysiology*, 17(2), 167-178. doi:10.1111/j.1469-8986.1980.tb00131.x

Kaltwasser, L., Hildebrandt, A., Recio, G., Wilhelm, O., & Sommer, W. (2014).

Neurocognitive mechanisms of individual differences in face cognition: A replication and extension. *Cognitive, Affective & Behavioral Neuroscience*,

14(2), 861-878. doi:10.3758/s13415-013-0234-y

Krigolson, O. E., Pierce, L. J., Holroyd, C. B., & Tanaka, J. W. (2009). Learning to become an expert: Reinforcement learning and the acquisition of perceptual expertise. *Journal of Cognitive Neuroscience*, 21, 1833–1840.

doi:10.1162/jocn.2009.21128

Krusienski, D. J., Sellers, E. W., McFarland, D. J., Vaughan, T. M., & Wolpaw, J. R. (2008). Toward enhanced P300 speller performance. *Journal of Neuroscience Methods*, 167, 15-21. doi:10.1016/j.jneumeth.2007.07.017

Lefebvre, C. D., Marchand, Y., Smith, S. M., & Connolly, J. F. (2007). Determining eyewitness identification accuracy using event-related brain potentials (ERPs). *Psychophysiology*, 44, 894-904. doi:10.1111/j.1469-8986.2007.00566.x

Lefebvre, C. B., Marchand, Y., Smith, S. M., & Connolly, J. F. (2009). Use of event-related brain potentials (ERPs) to assess eyewitness accuracy and deception. *International Journal of Psychophysiology*, 73, 218-225.

doi:10.1016/j.ijpsycho.2009.03.003

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

Lukács, G., & Ansorge, U. (2019). Information leakage in the response time-based concealed information test. *Applied Cognitive Psychology*. doi: 10.1002/acp.3565

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, 43, 385-388. doi:10.1037/h0046060

Lykken, D. T. (1960). The validity of the guilty knowledge technique: the effects of faking. *Journal of Applied Psychology*, 44, 258-262. doi:10.1037/h0044413

- Mann, S., Vrij, A., & Bull, R. (2004). Detecting True Lies: Police Officers' Ability to Detect Suspects' Lies. *Journal Of Applied Psychology*, 89(1), 137-149.
doi:10.1037/0021-9010.89.1.137
- Marchand, Y., Inglis-Assaff, P. C., & Lefebvre, C. D. (2013). Impact of stimulus similarity between the probe and the irrelevant items during a card-playing deception detection task: The 'irrelevants' are not irrelevant. *Journal Of Clinical And Experimental Neuropsychology*, 35(7), 686-701.
doi:10.1080/13803395.2013.819837
- Martens, U., Schweinberger, S. R., Kiefer, M., & Burton, A. M. (2006). Masked and unmasked electrophysiological repetition effects of famous faces. *Brain Research*, 1109(1), 146-157. doi:10.1016/j.brainres.2006.06.066
- Meijer, E. H., Smulders, F. T. Y., Merckelbach, H. L. G. J., & Wolf, A. G. (2007). The P300 is sensitive to concealed face recognition. *International Journal of Psychophysiology*, 66, 231-237. doi:10.1016/j.ijpsycho.2007.08.001
- Meixner, J. B., & Rosenfeld, J. P. (2014). Detecting knowledge of incidentally acquired, real-world memories using a P300-based concealed-information test. *Psychological Science*, 25(11), 1994-2005. doi:10.1177/0956797614547278
- Palmer, M. A., Brewer, N., & Weber, N. (2012). The information gained from witnesses' responses to an initial 'blank' lineup. *Law and Human Behavior*, 36(5), 439-447.
doi:10.1037/h0093939
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225-237. doi: 10.3758/PBR.16.2.225

- Schacter, D. L., Reiman, E., Curran, T., Yun, L. S., Bandy, D., McDermott, K. B., & Roediger, H. L. (1996). Neuroanatomical correlates of veridical and illusory recognition memory: Evidence from Positron Emission Tomography. *Neuron*, 17, 267-274. doi:10.1016/S0896-6273(00)80158-0
- Schweinberger, S. R., Pickering, E. C., Jentsch, I., Burton, M., & Kaufmann, J. M. (2002). Event-related potential evidence for a response of inferior temporal cortex to familiar face repetitions. *Cognitive Brain Research*, 14, 398-409. doi:10.1016/S0926-6410(02)00142-8
- Sun, D., Chan, C. H., & Lee, T. C. (2012). Identification and classification of facial familiarity in directed lying: An ERP study. *PloS one*, 7(2), e31250. doi:10.1371/journal.pone.0031250
- Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, 150(37), 1187-1188. doi:10.1126/science.150.3700.1187
- Tanaka, J. W., & Pierce, L. J. (2009). The neural plasticity of other-race face recognition. *Cognitive, Affective & Behavioral Neuroscience*, 9(1), 122-131. doi:10.3758/CABN.9.1.122
- Treese, A., Johansson, M., & Lindgren, M. (2010). ERP correlates of target–distracter differentiation in repeated runs of a continuous recognition task with emotional and neutral faces. *Brain And Cognition*, 72(3), 430-441. doi:10.1016/j.bandc.2009.12.006
- Trenner, M. U., Schweinberger, S. R., Jentsch, I., & Sommer, W. (2004). Face repetition effects in direct and indirect tasks: an event-related brain potentials study.

Cognitive Brain Research, 21, 388-400. doi:10.1016/j.cogbrainres.2004.06.017

Walla, P., Endl, W., Lindinger, G., Deecke, L., & Lang, W. (2000). False recognition in a verbal memory task: an event-related potential study. *Cognitive Brain Research*, 9, 41-44. doi:10.1016/S0926-6410(99)00041-5

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, 7, 45-75. doi:10.1111/j.1529-1006.2006.00027.x

Westmarland, L. (2013). 'Snitches get stitches': US homicide detectives' ethics and morals in action. *Policing & Society*, 23(3), 311-327. doi:10.1080/10439463.2013.784313

Werner, N., Kuhnel, S., & Markowitsch, H. J. (2013). The neuroscience of face processing and identification in eyewitnesses and offenders. *Frontiers in Behavioral Neuroscience*, 7, 189. DOI 0.3389/fnbeh.2013.00189.

Winograd, M. R., & Rosenfeld, J. P. (2011). Mock crime application of the Complex Trial Protocol (CTP) P300-based concealed information test. *Psychophysiology*, 48(2), 155-161. doi:10.1111/j.1469-8986.2010.01054.x