Two Approaches to Assessing Eyewitness Accuracy

by

Mario Joseph Baldassari
BA, Lake Forest College, 2011
MSc, University of Victoria, 2013

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Psychology

Supervisory Committee

Two Approaches to Assessing Eyewitness Accuracy

by

Mario Joseph Baldassari
BA, Lake Forest College, 2011
MSc, University of Victoria, 2013

Supervisory Committee

Dr. D. Stephen Lindsay, Psychology
Co-Supervisor

Dr. C. A. Elizabeth Brimacombe, Psychology
Co-Supervisor

Dr. Rebecca Johnson, Law
Outside Member

Abstract

Supervisory Committee
Dr. D. Stephen Lindsay, Psychology
Co-Supervisor
Dr. C. A. Elizabeth Brimacombe, Psychology
Co-Supervisor
Dr. Rebecca Johnson, Law
Outside Member

This dissertation presents two individual-difference measures that could be used to assess the validity of eyewitness identification decisions. We designed a non-forced two-alternative face recognition task (consisting of mini-lineup test pairs, half of which included a studied face and half of which did not). In three studies involving a total of 583 subjects, proclivity to choose on pairs with two unstudied faces weakly predicted mistaken identifications on culprit-absent lineups, with varying correlation coefficients that failed to reach the value $r = 0.4$ found in Baldassari, Kantner, and Lindsay (under review). The likelihood of choosing correctly on pairs that included a studied face was only weakly predictive of correct identifications in culprit-present lineups (mean r of .2). We discuss ways of improving standardized measures of both proclivity to choose and likelihood to be correct when choosing.

The second measure is based on the Guilty Knowledge Test (GKT), a lie detection method that utilizes an oddball paradigm to evoke the P300 component when a witness sees the culprit. This GKT-based lineup was intended to postdict identification accuracy regardless of witnesses' overt responses, thus faces are used as stimuli. Half of participants were instructed to respond as if they knew the culprit and wanted to falsely exonerate him. P300 amplitudes evoked by the culprit were indistinguishable from those evoked by a different learned face but were larger than P3s evoked by unfamiliar faces in

both the described lying condition and the group of participants who intentionally told the truth.

Table of Contents

List of Tables

List of Figures

Acknowledgments

My well-being throughout this process would have gone absolutely haywire without Dimitra's love and support. Special shout outs to Calum, Elliott, Julie, Kaitlyn, and Tanjeem, for being open ears when I needed it.

And of course the entire experience would have been impossible without the guidance, advice, and coaching I received from Steve since I arrived in Victoria in 2011. Thank you so very much for welcoming me then and for all your help in getting here, Steve.

Dedication

This dissertation is dedicated to my parents and grandparents. I am only here because of your lifetimes of hard work that gave me the access and freedom to pursue the high-minded ideals of psychological science. I will be lucky for the rest of my life, Dimitra, that you stuck around throughout all this. Love to you all.

Chapter 1 – Lineup Skills Test

Individual differences may predispose some people toward making more accurate eyewitness identification decisions than others across many types of witnessing conditions. Indeed, such differences have been in the hive mind of psychologists since Munsterberg first published *On the Witness Stand* at the beginning of the 20th century: "The courts will have to learn, sooner or later, that the individual differences of [people] can be tested to-day by the methods of experimental psychology far beyond anything which common sense and social experience suggest" (1908/2009, p. 47). Despite Munsterberg's early assertion, surprisingly little individual differences research has been done in the eyewitness memory domain. Witnessing conditions have since been systematically manipulated by researchers (see Granhag, Ask, & Giolla, 2014; Valentine, 2014, for reviews), but some data have shown varying levels of performance in identification tasks among participants when all have comparable encoding conditions (Darling, Martin, Hellman, & Memon, 2009; Valentine, Pickering, & Darling, 2003). These variations in performance are likely due to both skill in encoding a new face and individual differences in response bias (Kantner & Lindsay, 2012; Megreya & Burton, 2007). If such variation is stable within a participant, a measure of both face recognition ability and face memory response bias should be a reliable predictor of eyewitness identification skill. There is a separate literature on relationships among different face recognition tasks, some of which reveal correlations around $r = 0.6$ (McKone, Hall, Pidcock, Palermo et al., 2011; Megreya & Burton, 2006, 2007). As lineups are a face recognition task, scores on lineup tasks may also correlate with other measures in this domain at around the same strength.

Bindemann, Brown, Koyas, and Russ (2012) hypothesized that the apparent similarity between identification tasks and face recognition tasks should mean that one is predictive of the other. Consistent with that idea, some applied researchers already use face recognition tasks to approximate lineup presentation when testing new methods. Weber and colleagues have used mini-lineups with four members as methodological stand-ins for full lineups (e.g., Weber & Varga, 2012). Weber and Varga tested a new lineup procedure in which participants studied a list of labelled faces and then were asked to identify a specific studied face (based on the label) out of a lineup of four faces. Responses to these mini-lineups were compared to another set of mini-lineups presented slightly differently (as in Weber & Varga; Weber & Brewer, 2004), and were also used as a proving ground for a hypothesis before application of the idea to a traditional video-lineup paradigm with six-person lineups (Sauer, Brewer, & Weber, 2008). The act of testing new procedures with this method implies that a procedure yielding higher accuracy for mini-lineups will translate well to full sized lineups. This assumption seems reasonable and converges with the conclusion reached by Bindemann et al. (2012), but there is no direct exploration of the relationship between mini-lineups and 6-person photospread lines in the published literature. The current research provided such tests.

The literature on the Cambridge Face Memory Test (CFMT, an extensively-tested measure of face recognition ability) aided us in setting expectations for the size of the correlations between face recognition task and a lineup task. Scores on the CFMT have been thoroughly examined for correlation with related measures (Bobak, Hancock, & Bate, 2016; Bowles, McKone, Dawel, Duchaine, Palermo, Schmalzl, Rivolta, Wilson, & Yovel, 2009; McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012). Strengths of

these correlations range from $r = 0.26$ to $r = 0.61$, see Table 1 for predictors and specific findings. Some of the fluctuation in the strength of the relationships between these seemingly very similar tasks may call into question the test-retest reliability of such measures, as well as the possible upper bound of these correlations. The reliability of the CFMT is well established, both originally by Duchaine and Nakayama (2006) and in many studies since. Internal reliability scores within and correlations between two variations of the CFMT (traditional CFMT and new CFMT-Aus, McKone et al., 2011) indicated a hypothetical upper bound of $r = 0.86$, based on a measured $r(72) = 0.61$ (see Table 1 for details). The upper bound of the correlation between face memory tasks and lineup tasks is likely not so large, but if it approached $r = 0.6$ we could begin to construct a predictive task useful for real world police to assess the quality of their eyewitness IDs.

Individual differences in face recognition ability have been used as a predictor of lineup identification accuracy with some success, though few have found relationships stronger than $r = 0.4$. Hosch (1994) reported the first data of this kind in which participants' scores on the Benton Facial Recognition Test (BFRT) were significantly correlated with accuracy on a lineup in which participants identified the experimenter who gave their task instructions. Half of these lineups contained the experimenter (culprit-present, or CP) and half did not (culprit-absent, CA). See Table 2 for r values, sample sizes, and 95% confidence intervals around r. This correlation held fairly steady around $r = 0.45$, though noisily, across three small-N studies with slightly different procedures, but two other studies using the BFRT did not produce significant correlations larger than $r = 0.05$. Using two new samples, Hosch tested the relationship between accuracy on the same lineup task and measures of sensitivity and response bias on a

yes/no face recognition task. The number of trials in the face task was not reported, but the first study yielded no correlation between sensitivity and ID accuracy and a significant correlation between response bias and ID accuracy. Also, participants who produced a false alarm on a CA lineup had a more neutral bias on average (B" mean = -0.1) than those who produced a correct selection on a CP lineup, who tended toward conservative in their face recognition decisions (B" mean = 0.59). A second study weakly replicated these findings, and these also appear in Table 2. The samples in Hosch's studies were not large enough to produce a stable estimate of the true correlation strength (Schönbrodt & Perugini, 2013). Nonetheless, these data established the "common knowledge" that face recognition scores can predict eyewitness accuracy.

Data from Kantner and Lindsay (2012, 2014) indicated that individual differences in willingness to endorse items in a face recognition task may be sufficiently large and reliable to be useful in evaluating eyewitness identification decisions. Several studies showed evidence of stable, trait-like differences in old/new recognition memory response bias across face, word, and painting stimuli and across testing contexts. Kantner and Lindsay (2014) also observed a statistically significant correlation between response bias in a yes/no recognition test with face stimuli and number of identifications made on a set of culprit-absent lineups, but replication with a larger sample size would strengthen their findings considerably (see Table 2).

The relationship between the BFRT and lineup task accuracy reappeared in a replication of Hosch's original findings (Geiselman, Tubridy, Bkynjun, Schroppel, Turner, Yoakum, & Young, 2001). Participants who chose the culprit from either of two CP lineups tended to have higher scores on the short form of the BFRT, but the scores

were not predictive on easier lineups in which most participants chose the culprit. What Geiselman et al. refer to as a difficult lineup is likely the most plausible type to be deployed in the real world, especially since the large scale adoption of lineup administration practices suggested by psychologists in the 1990's (Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998). Therefore it remains likely that a face recognition test such as the BFRT could be useful in predicting lineup accuracy when the culprit is present, but Geiselman et al.'s data do not measure the predictive utility of response bias. Additionally, these and all studies using the BFRT should be considered with appropriate skepticism, as there is evidence that one can ignore face identities and still score highly on the BFRT by focusing on eyebrows (Duchaine & Nakayama, 2004).

Bindemann et al. (2012) used an altered version of the face matching task designed by Bruce et al. (1999) to predict lineup performance. To turn the matching test into a memory task, Bindemann et al. had participants study the target face on a separate slide before presenting the 10-person test array. The data showed that participants who made a correct ID from a CP lineup tended to have higher hit rates in the Bruce test than did participants who had not made a correct ID, reported Cohen's $d = 0.71$, our calculated 95% CI [0.05, 1.59] (see Table 2 for correlations). Participants who correctly rejected a CA lineup tended to have higher correct rejection rates in the adapted Bruce test than those who chose from a CA lineup, $d = 0.93$, 95% CI [0.26, 1.63]. In a second experiment, participants who made a correct lineup response (either choosing or rejecting) tended to have higher correct rejection rates in the modified Bruce task, choosers $d = 0.42$ [0.003, 1.07], nonchoosers $d = 0.54$ [0.12, 0.98]. That an individual witness's proclivity to choose (PTC, to be thought of like response bias) on a lineup was

predicted by their proclivity to choose in the modified version of the adapted Bruce task makes intuitive sense because the task is much like a 10-person lineup. However, that a witness's tendency to choose correctly from a CP lineup was also predicted by their proclivity to choose in the adapted Bruce task (replicating some of Hosch's findings) suggested that an individual's proclivity to choose on a face memory task may be a robust predictor of accuracy above and beyond system or situational factors that influence the likelihood the witness will answer a lineup correctly.

There is also evidence of a relationship between face recognition test performance and eyewitness identification in a stressful, realistic setting. Morgan, Hazlett, Baranoski, Doran, Southwick, and Loftus (2007) observed a positive relationship between face recognition ability and eyewitness accuracy in a group of 46 Army trainees. The trainees underwent a stressful interrogation, and their ability to later identify the interrogator from a 10-person sequential lineup (CP for 58% of participants) was predicted by scores on the face subtest of the Weschler Intelligence Test. Out of 48 possible correct responses, the trainees who were correct on their lineup judgment had an average score of 33.8, while those who made an incorrect judgment on the lineup had an average score of 27.3. This difference was driven by the finding that trainees who made a correct decision on the lineup tended to have produced fewer false negatives and more true positives in the Weschler test (MANOVA p's $< .01$). Tukey post hoc tests split these findings into types of eyewitness decisions and found that participants who produced false positive ID's were in fact the drivers of the effect, as this group tended to have made fewer true positive responses and more false negatives in the Weschler test (p's between 0.1 and

.05). That false positives drove Morgan et al.'s effects is further evidence that proclivity to choose on a lineup is a predictable individual difference.

Andersen, Carlson, Carlson, and Gronlund (2014) aimed to measure both face recognition skill (FRS, akin to sensitivity) and PTC from a lineup by inserting multiple predictors into four separate logistic regressions for CP and CA simultaneous and sequential lineups. Of their 238 participants, each watched two videos and saw one CP and one CA lineup. Half of the participants were shown sequential lineups, the other half saw simultaneous. One predictor was participants' score on the Cambridge Face Memory Test (CFMT, developed to replace the BFRT, Duchaine & Nakayama, 2006). Odds ratios indicated that for every unit increase in CFMT score (ranging from 0 to 100), there was a 1% higher likelihood of a correct simultaneous lineup ID, and a 1% lower likelihood of a simultaneous or sequential false positive ID, see Table 2 for correlations derived from a logistic regression. Thus Anderson et al. (2014) supported the hypothesis that the predictive utility of face recognition for identification tasks can be two-sided, in that witnesses showed individual differences in FRS and PTC.

As stated by Megreya and Burton (2007), any test measuring whether a witness is "good at faces" should incorporate a test of both the witness's ability to choose correctly from a CP array and to correctly reject a CA array, or the witness's FRS and PTC. Across these five studies, 4 found support for face recognition tasks predicting a witness's likelihood to select correctly from a CP lineup (Andersen et al., 2014; Bindemann et al., 2012; Geiselman et al., 2001; Hosch, 1994), and 4 found support for the same tasks predicting witness's likelihood to correctly reject a CA lineup (Andersen et al., 2014; Bindemann et al., 2012; Hosch, 1994; Kantner & Lindsay, 2014). Essentially, all five

supported the predictive utility of the side of being "good at faces" that the authors set out to test, but the strength of the relationships varied considerably. Other unpublished studies seemed to show effects of a similar size (See Deffenbacher, Brown, & Sturgill, 1978, in Table 2) that did not reach significance because the samples were underpowered. Deffenbacher et al. presented otherwise unpublished efforts to predict eyewitness accuracy at the Practical Aspects of Memory conference in Cardiff (1978), in which an overall score on a yes/no face recognition test was not significantly correlated with accuracy in a very difficult lineup. Lastly, Hosch (1994) wrote that unpublished findings from Shepherd, Davies, and Ellis (1980) showed that recognition bias was predictive of eyewitness accuracy but sensitivity was not.

Sample size issues aside, the lack of consistency in these findings has also been due to the variety and the nature of the face tests used, as no study has yet produced correlations near the upper bounds suggested by the CFMT data in Table 1 (apart from the low-*N* findings by Hosch, 1994). The CFMT and BFRT may not be optimal indices of eyewitness skill. After all, these measures were not initially developed for this use and were intended to diagnose prosopagnosia by assessing sensitivity in face recognition, not response bias. In Baldassari, Kantner, and Lindsay (under review), we aimed to develop and test superior measures of both sides of being "good at faces" in the context of eyewitness identification lineups. To that end we crafted a new procedure that we have dubbed the Lineup Skills Test (LST). The long-term ambition of this line of research is to develop a standardized test of eyewitnesses that assesses both (a) ability to recognize a culprit's face when it is present in a lineup and (b) proclivity to choose an innocent suspect when the culprit is absent from a lineup.

Baldassari et al. designed a new face recognition test to predict eyewitness accuracy based on a previous finding of a correlation between response bias on a yes/no face recognition task and number of rejections of a series of lineups (Kantner & Lindsay, 2014). We tested face memory with a two-alternative non-forced choice recognition task in which 50% of the trials contain a studied face and an unstudied face and the other 50% contain two unstudied faces. Scores on this Lineup Skills Test (LST) were compared to performance on five lineups. The LST paired a measure of Facial Recognition Skill (FRS) similar to sensitivity (accuracy when choosing on pairs containing one studied face and one non-studied face) with a measure of PTC (rejection rates of pairs containing two non-studied faces). False positive selection rates on these pairs of unstudied faces reliably predicted false positive selection rates on five CA lineups completed before the face recognition study list begins across 4 samples, $r \approx 0.4$. The relationship held steady through two local samples of university students, two samples of workers recruited from Amazon's Mechanical Turk, and procedures that included a two day or a five minute delay between video viewing and lineup completion. If tweaks to the procedure or materials of the LST produce stronger relationships with lineup accuracy, then such a test could provide police with a measure of an eyewitness's likelihood to make an accurate lineup decision. Such a measure could strengthen the evidentiary value of an identification or lineup rejection from a high-scoring witnesses in court, thereby ensuring that the truly guilty are found so. An LST that accounts for much of the variance in eyewitnesses would also enable police to treat the lineup decision of a low-scoring witness with appropriate skepticism to avoid unnecessary and wrongful arrests.

Study 1

The following studies reflect attempts to strengthen both the PTC and FRS relationships. We have thus far begun testing a larger, more diverse (in age and ethnicity) set of faces in the LST with a change based on a suggestion from Jacoby, who hypothesized (Personal Communication, 2016) that 2-alternative recognition tests which paired similar words at test would result in increased confusion and error compared to randomly selected test pairs. This is akin to description-matching in face recognition, which is the traditional way lineups are filled out with foil members. Thus, an LST with description-matched pairs would present the opportunity for the type of error likely to be made on full sized lineups. We also sought to take advantage of the possibility of diversity in our university sample by diversifying the faces in the test. A test with all Caucasian faces would be easy for Caucasian participants and hypothetically a bit more difficult for other-race participants with limited exposure to Caucasian faces (Malpass & Kravitz, 1969; Tanaka & Pierce, 2009). An additional advantage of diversifying the face set was that it made it more likely to be more widely useful in real world practice. The forthcoming studies, therefore, provided tests of these changes compared to the original LST in Baldassari et al.

Method

Participants. Participants were recruited through the psychology research participation system at the University of Victoria (N = 182) and were compensated with course credit.

Materials and procedure. Participants met in groups of 2 to 25 in a computer laboratory on campus. After participants signed in to their individual computers, an

experimenter directed attention to the presentation board on which five videos were shown in succession. The five videos were clipped from British television crime dramas, all of which depicted middle-aged Caucasian male culprits committing crimes. A clip of a man breaking into a home (about 18s of exposure to culprit) was obtained from *Vincent*, a clip of a man and woman arguing and a clip of a woman's car exploding as she leaves her home (about 13s and 15s, respectively, of exposure to culprit) were obtained from *MI-5*, a clip of a man destroying cabinets of fine china with a shotgun (about 16s of exposure to culprit) was obtained from *Dalziel and Pascoe*, and a clip of a man shooting another man (about 35s of exposure to culprit, most from distance) was obtained from "Murder City." Clips ranged from 47 to 83 seconds in length and were presented with the original sound tracks; any gory shots of violence were removed. After a lengthy distractor task in which participants judged 96 high quality digital scans of paintings, participants responded to a lineup for each video. Lineups consisted of six photos of men who fit a description of the culprit. The photos were gathered from various internet sources then edited so that all were wearing similar clothing. The filler face we thought most resembled the culprit was predesignated the "innocent suspect" in the CA lineup for each crime. Groups were split as evenly as possible, 95 participants saw five CA lineups and 87 saw five CP lineups. Crime and lineup order were reversed for half of the sample.

Next, participants studied one of four fixed random sets of 50 digital photos of faces from a larger set of 80 men and 120 women. Of these 200 faces, 36 people were of African descent, 144 were of Caucasian descent, and 20 were of South Asian descent. The youngest face was 18 years, and the oldest was 89. In the study portion, we used

faces making a neutral expression with a 1s gray mask between faces.[1] The photos were shown in a head-and-shoulders view in color (Minear & Park, 2004). Photos were selected from the much larger set uploaded by the Park Aging Mind Laboratory if the set included a neutral and a smiling photograph of the person. Two independent lab members organized the available photographs into description-matched pairs; in cases where more than two photos matched one description, the lab members agreed on the best match based on factors beyond the basic descriptions. Photos were 640x480 pixels on screen, and were presented in one of four fixed random orders. The instructions referred to the face recognition test as a Lineup Skills Test and informed participants before the test phase that it was meant to measure their ability on the preceding identification task (See Appendix A for a full set of instructions). After a 5 minute distractor task, participants began the LST. In each of 100 trials a pair of digital photos of faces appeared to the right and left of the mid-point of the screen; half of the trials consisted of one studied or "old" face and one unstudied or "new" face (the Face Recognition Skill portion, consisting of Old/New pairs). The other 50 trials each consisted of two unstudied faces (the Proclivity to Choose portion, containing New/New pairs). The two types of trials were randomly interleaved, and the faces in the test phase were photos taken in the same session as those in the study phase but with the subject smiling to encourage face recognition rather than photo recognition (Bruce & Young, 1986). The first two and last two faces in the study list were not used in the test list to avoid primacy and recency effects. Test trials displayed selection options of Left, Neither, and Right with corresponding keyboard

[1] See [https://osf.io/nptmy/] for the entire set of faces from which our sets were drawn, as well as downloadable programs of our entire procedure. The faces were downloaded from the Park Aging Mind Laboratory at the University of Texas – Dallas.

buttons. Participants then rated confidence in each response on the test list on a 11-point scale (by tens, 0-100).

Results

Figure 1 displays proportion correct on N/N pairs and proportion correct on lineups from Study 1, $r(93)= 0.29$, $p = .002$,[2] 95% CI [0.09, 0.46]. Figure 2 displays proportion correct on O/N pairs and proportion correct on CP lineups for Study 1, $r(85)= 0.26$, $p = .008$, 95% CI [0.05, 0.45]. There was also a significant correlation between accuracy rates on O/N pairs and CP lineups when choosing, $r(85)= 0.19$, $p = .04$, 95% CI [-0.02, 0.39].

Discussion

As predicted, participants who falsely chose more often on N/N pairs also tended to falsely choose more often on later CA lineups than participants who correctly rejected more N/N pairs. A PTC correlation of almost 0.3 would be potentially useful in the real world, but it does not reach the maximum possible correlation strength suggested by the studies already discussed. Study 1 was a replication of both critical measures from Baldassari et al. in that the correlations were significant, but they were somewhat weaker in the current study. It is reasonable, though, that an effect's size would fluctuate from sample to sample, and an *r* value of 0.29 is well within the potential range for the expected value of 0.4. We replicated the study to determine whether this was merely an occurrence of random noise or was a sign of some weakness in the new materials set adopted since Baldassari et al.

[2] *P* values report one-tailed tests for correlations, as it was hypothesized that these tasks would correlate positively.

Study 2

Method

Participants. Participants were recruited through the psychology research participation system at the University of Victoria ($N = 202$) and were compensated with course credit.

Materials and procedure. The materials and procedure were identical to Study 1, except that the distractor task was a series of personality inventories. Participants completed self-report versions of the Autism Spectrum Quotient (AQ; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) and the Liebowitz Social Anxiety Scale (LSAS; Fresco, Coles, Heimberg, Leibowitz, Hami, Stein, & Goetz, 2001) in a Qualtrics survey designed in the lab, and they completed the Multidimensional Social Competence Scale (MSCS; Yager & Iarocci, 2013) in its native web-based survey. These inventories were intended to address different questions than those being investigated here, thus they will not be discussed further. Additionally, after going once through the lineups participants were told "Our research shows that people sometimes reject a lineup even though they have a hunch that one of the lineup members might be the culprit. We would like you to go through the lineups again and pick someone from the lineup on each one. If it helps, you may think of this as an academic exercise rather than a police lineup with any consequences," and were sent through the lineups again but were forced to choose. Groups were split as evenly as possible, 113 participants saw five CA lineups and 89 saw five CP lineups.

Results

Figure 1 displays proportion correct on N/N pairs and proportion correct on lineups from Study 1, $r(111)= 0.10$, $p = .15$, 95% CI [-0.09, 0.28]. Figure 2 displays proportion correct on O/N pairs and proportion correct on CP lineups for Study 1, $r(87)= 0.22$, $p = .02$, 95% CI [0.01, 0.41]. There was not a significant correlation between accuracy rates on O/N pairs and CP lineups when choosing, $r(87)= 0.15$, $p = .08$, 95% CI [-0.06, 0.35]. There was a slightly stronger correlation between accuracy rates on O/N pairs and the second, forced-choice round of CP lineups, $r(87)= 0.27$, $p = .005$, 95% CI [0.07, 0.45]. The majority of participants ($N = 37$) made one more correct selection, and the next highest group ($N = 36$) had no change in their CP lineup score when forced to choose.

Discussion

That this replication produced an even smaller correlation than Study 1 suggested that the critical factor was not random noise but something to do with the change in materials between Baldassari et al. and the current study. This hypothesis was supported by the fact that participants in Study 1 and Study 2 exhibited a smaller range of N/N pair scores than those in the paper (N/N standard deviations in Baldassari et al.: 0.26, 0.24, 0.21, 0.18; standard deviations here: 0.2, 0.19), though it should be noted that the earlier samples were smaller. Standard deviation of O/N scores has been consistent throughout. It is possible that the process of completing the various personality scales was too long and participants simply did not remember the study list as well as in previous studies, or that participants focused on a perceived task demand suggested by the scales. Another possible explanation for the reduction in test accuracy is sampling bias: undergraduates at

UVic do not typically have high rates of exposure to people of African descent, therefore they may have found the presence of so many African faces surprising and overly focused on them in order to not fall into the perceived trap of the Other Race Effect. We will investigate that possibility in the forthcoming item analyses. It is also possible that so many other-race faces made the task more difficult, but this idea is a tempered somewhat by the surprising finding that the participants performed better on the Black face pairs than they did on the Caucasian face pairs (See Figure 3). We thought this was likely due to the distinctiveness of some of the older women in the set, and we set upon an item analysis to investigate this hypothesis.

A potential pitfall of the basic design of the LST is the unnatural nature of the acquisition of new faces in the study phase. Eyewitnesses in the real world experience the criminal naturalistically, and they likely exhibit individual differences in what they attend to. An orienting judgment was added to the study phase of the LST for Study 3 to try to encourage deeper processing and natural observation and discourage feature-based memorization.

Interim Item Analysis

As data collection progressed on Study 2, the data from Study 1 were further explored. Item analyses revealed several face pairs for which discrimination was very high and others for which it was low. The item analysis also revealed face pairs for which response bias was highly conservative or liberal. It was decided that the 20 most extreme values would dictate the face pairs to be removed in order to keep the list as long as possible while still removing all items that were obviously not contributing to any

individual differences in LST scores. Study and test lists were shortened to 40 faces and 80 pairs. See Table 3 for a full list of face pairs removed.

Study 3

Method

Participants. Participants were recruited through the psychology research participation system at the University of Victoria (N = 199) and were compensated with course credit.

Materials and procedure. The materials and procedure were identical to Study 2, except for a few critical aspects. First, there was no extra round of forced-choice decisions through the lineups. The distractor task was reduced to only include the MSCS. Faces deemed too easy to be diagnostic of skill in the above item analysis were removed from the test. We hypothesized that the LST would produce slightly lower accuracy levels, because more easy than difficult test pairs were removed. To counteract this difference, participants completed an orienting task to each study phase face. If the face was older than 30 years, participants were to press the 'o' key, and if the face was younger than 30 they pressed the 'y' key. This manipulation was meant to encourage holistic processing of the face. Groups were split as evenly as possible, 99 participants saw five CA lineups and 100 saw five CP lineups.

Results

Figure 1 displays proportion correct on N/N pairs and proportion correct on lineups from Study 1, $r(97)= 0.33$, $p = .0004$, 95% CI [0.14, 0.50]. Figure 2 displays proportion correct on O/N pairs and proportion correct on CP lineups for Study 1, $r(98)= 0.11$, $p = .14$, 95% CI [-0.09, 0.30]. There was not a significant correlation between

accuracy rates on O/N pairs and CP lineups when choosing, $r(98) = -0.002$, $p = .49$, 95% CI [-0.21, 0.20].

Average accuracy on N/N pairs was significantly higher in Study 3 than in Study 2, $t(399) = 4.44$, $p < .0001$, Cohen's $d = 0.44$, 95% CI [0.24, 0.64], see Table 4 for group means. The reverse was true of ON pairs, $t(399) = 9.33$, $p < .0001$, Cohen's $d = 0.93$, 95% CI [0.73, 1.14].

Discussion

Removal of the 20 test pairs from the item analysis reduced average accuracy on the FRS half of the LST despite the addition of the younger/older judgment, as the average score on O/N pairs was significantly lower than in Study 2. The range of O/N accuracy scores is still quite restricted, as evidenced by the weaker FRS correlation (see Figure 2) and the slightly smaller standard deviation. The younger/older judgment may have aided N/N pair performance beyond the reduction that was expected from the removal of the easy test pairs, but the increase in the strength of the correlation appears to be more because of a lack of outliers than an increased range of responses overall.

The main PTC correlation showed renewed significance, suggesting that the low *r* value in Study 2 was either a product of random noise or was a result of a more restricted range of N/N face pair scores due to the overly easy face pairs and difficulty of remembering the faces across all the personality tests. Though it recovered to its more regular value for our studies, it still does not reach toward the 0.6 potential value shown in other face recognition work. The return of the PTC correlation coincided with the loss of a significant FRS correlation between Studies 2 and 3, thus leaving the LST not quite

accounting for enough variance in eyewitness skill to be presented to police or triers of fact.

General Discussion: LST Studies

Through the three LST studies detailed here, we encountered measurement issues that had not presented in earlier iterations of the test (Baldassari et al., under review). Changing from a set of university-aged, Caucasian faces taken within the lab to a more ethnically and age-diverse set from the Park Lab was the main difference between the earlier tests and those in the current study. The intent to make the test more difficult and more externally valid may have backfired in that the diversification of the face set made it more heterogeneous. The heterogeneity then made the faces in the set easier to distinguish from one another based on distinctive, lower level information in each face. See Figure 4 for item analyses from Study 1 by race of the face pair. As most participants were Caucasian, the expected Other Race Effect (ORE; Meissner & Brigham, 2001) does not appear here, likely due to the uniqueness of the African and Indian faces within the set. This effect should have been counteracted by the fact that test pairs only contain two members of the same race, but as the list becomes one mental object perhaps the crossover between and among the faces in the mind leads to the advantage gained by the heterogeneity of the list. The reversal of the ORE seen here might be counteracted by use of a longer list, but that would require a study/test cycle about twice as large as that we have been using. It is already likely that the LST as currently designed requires high memory load capabilities and a long period of focus and fatigue rather than the quick recognition skill demanded by the CFMT and the Bruce task. Thus, a delayed match to sample (DMTS) task may more readily predict lineup skill. A DMTS LST would present

a face for less than a second, mask it, then replace the mask with a pair of faces. The participant's task would be the same: determine whether either of the replacement faces was the face just studied. Easing the high memory load participants have been under in previous LST studies should ensure that performance differs on the basis of face recognition ability rather than ability to hold large loads in memory. A DMTS procedure would also serve to make the task based more in face perception than in longer-term face memory, as it would hinge more closely to tests like the CFMT. Such a test would also require only comparisons within a single race on a single trial, rather than memory judgments based on ability to recognize anyone from the study set.

Another area that the LST could continue to be improved is in its materials. The faces from the Park Lab are fantastic, but they may not be the best for the more basic skill we intended to measure. It remains possible that participants study distinctive features like eyebrows, piercings, clothing, or hair to make their LST judgments. Cropping the Park Lab's faces to ovals of just the face information would remove this possibility, and gray-scaling the faces would further do so. On the other hand, Duchaine and Weidenfeld (2003) reported that cropping faces in a test to ovals resulted in similar results, so while prosopagnosics may use outer features when available, typical participants mostly attend to internal facial features. Apart from modifying the Park Lab's faces, a future version of the LST could contain a different face set. We attempted to measure person recognition rather than photo recognition by presenting faces with different expressions at study and test, but perhaps the photos were still too similar to prevent participants from focusing on individual features rather than holistic faces. A set of face photographs offering different

angles and expressions might enable a more clear measure of identity recognition rather than recognition of familiar features.

Chapter 2 – ERP Lineup

When law enforcement officials and witnesses perform their duties to the best of their abilities, eyewitness identification (ID) is still an unreliable form of evidence in a criminal trial. An ID becomes much stronger evidence, though, when the witness knows the perpetrator well. However, this increase in accuracy is eliminated when the witness has reason to lie about knowing the perpetrator. Witnesses may not wish to divulge recognition of a perpetrator in cases in which such recognition would implicate the witness, in which the perpetrator is a friend of family member of the witness's, in which the witness might be under threat from associates of the perpetrator, or others. The second section of my dissertation will describe a study in which we attempt to solve the problem of an uncooperative but knowledgeable eyewitness.

In cases where a subject's overt responses cannot be trusted, researchers sometimes turn to responses elicited from other, uncontrollable behaviors of the mind or body to find truthful answers to critical questions. One such method of studying responses in the body is electroencephalographic Event-Related Potentials (ERP). The development of ERP technology and methodology in the 1980's offered an insight into neural activity that was before inaccessible (Luck, 2005). The tight time-course of electrical responses in the brain to stimuli offered a clear connection between neural activity and actions in the world. Researchers have, for example, discovered separate components after a familiar face is presented to a participant that represent awareness that it is a face (N170), the identity of the face (N250), and connections to context and experiences involved with that person (P300; Eimer, Gosling, & Duchaine, 2012).

Electroencephalography Overview

Electroencephalography (EEG) was first used as a medical tool, but the first studies to use EEG as a measure of cognition were published in the 1950's and 1960's. Cognitive psychologists had previously only been able to infer the workings of the mind and brain from cleverly designed tests, but EEG offered the first opportunity to bridge the gap between the performance of the mind and the physicality of the brain. It also offered the possibility of linking behaviors that were previously thought to be disparate but turned out to have similar cortical activity. In the cortex, a neuron fires by opening channels through which positively charged sodium and potassium ions flow, then allowing them to slowly leave, which essentially cause the overall electric potential of the cell to increase from below its action potential to far above it and back down again. These individual neuron firings result in changes in electricity that are far too small to be detected without inserting a probe directly into brain tissue, but when enough neurons fire in the same direction at the same time, the charge grows so that it is just strong enough to be detected on the skin of the head. A modern EEG system utilizes a ground electrode near the front-center of the head to measure the hum electricity flowing through and around the body as well one or more reference electrodes to establish a baseline of the connection strength between an electrode and the skin. The automation of mathematical techniques that enable the elimination of artifacts such as eye blinks or sneezes on a by-trial basis have made EEG research many times more reliable in the decades since. In the context of cognitive psychological research, EEG data are most useful when the researcher focuses on the few seconds immediately after the appearance of a stimulus. In fact, if many stimuli are presented similarly within an experiment, the researcher can

average together the EEG voltages from all those trials in the second or two after the stimulus appears and create what is known as an event-related potential (ERP). Combining EEG signals from many stimulus presentations into an ERP enables researchers to improve the signal-to-noise ratio for tests comparing the conditions of an experiment to one another.

Once researchers began creating ERPs, patterns emerged depending on the context and type of stimulus being presented. Some examples included a larger negative voltage 170ms after the onset of a face compared to other stimuli, with an accompanying negative voltage 250ms after the onset if the face was known to the participant. These distinguishable, specific changes in voltage are known as ERP components. The aforementioned components, the N170 and the N250 (so named for their negative direction and the time of their typical peak), were present in the current study but were not of interest in the analysis. The component of interest was the P300, which Sutton, Braren, Zubin, and John (1965) first introduced as the ERP correlate of stimulus uncertainty. Sutton et al.'s participants heard pairs of sounds or saw pairs of lights in either a predictable or an unpredictable scenario. The trials that were unpredictable resulted in a larger positive voltage about 300ms after the onset of the stimulus, and so researchers in the years since have taken to calling this component the P300 (sometimes shortened to P3). The less expected the event, the larger the P300 (Johnson & Donchin, 1980). The more different the event is from its surrounding events, the larger the P300 (Gill & Polich, 2002). Contemporary understanding of this component has expanded to include its sometimes-longer duration, and now most large positive voltages happening between 250ms and 500ms after stimulus onset are considered part of the P300 family of

effects (Donchin, 1980). The P300 appears in parietal regions of the scalp when the so-called 'oddball' stimulus appears, and it can be produced in a variety of contexts as long as the participant and experimenter agree upon the context of the current stimulus list and upon how to classify the stimuli.

The P300 and the Guilty Knowledge Test

Some psychologists have endeavored to use the P300 to aid law enforcement and other truth-seekers in lie detection by use a method called the Concealed Information Test (CIT), also known as the Guilty Knowledge Test (GKT). Researchers hypothesized that a criminal should have intimate knowledge about his crime that would betray his guilt if he could be coerced into revealing that knowledge. If, for example, the gun used in a bank robbery was only known to those who were at the scene (its type had not reached news outlets, and all bystanders had been exculpated) a suspect would not want to reveal knowledge of what type of gun was used. If the police showed the suspect a slideshow of guns that included the gun in question, one of two situations should arise: (1) the suspect is not the criminal, therefore to him the slideshow is just a series of guns or (2) the suspect is the criminal and the slideshow contains the gun he used to rob the bank. If these suspects were connected to an EEG monitor, their ERPs should be easily distinguishable from one another because the gun used to rob the bank would stick out of the list as an oddball to the culprit and elicit a P300, but it would not be an oddball to an innocent suspect.

Farwell and Donchin (1991) published the first attempt at such a test adapting Lykken's original GKT (1959), and the follow-up by Allen, Iacono, and Danielson (1992) shortly after established the general method that many would adopt. They

presented three types of stimuli to participants: known-familiar (infrequent items the experimenters knew participants would recognize), known-unfamiliar (frequent items the experimenters knew participants would not recognize), and unknown-familiar (infrequent items that only 'guilty' participants would recognize). Thus, the important test is whether P300 amplitude to unknown-familiar items is more like that of the known-familiar or the known-unfamiliar items (See Figure 4).

In the application of this test, a suspect for an attack by knife would be shown a slideshow of photographs of knives and other implements (the known-unfamiliar filler items) with the clear instruction that if he sees the knife from the crime or one particular other implement (perhaps a pair of garden shears) he should say so. Though most criminals would not self-incriminate by choosing the knife used in the crime, they would still watch for and identify the shears. The shears would thus elicit a P300 as the known-familiar item. The amplitude of the suspect's P300 to the shears could then be compared to the amplitude of the P300 to the knife used in the attack, the unknown-familiar item. An innocent suspect, on the other hand, would not know the knife used in the crime from any of the other knives in the list and so would not view it as an oddball. The differences in the P300 amplitudes between and among conditions are likely to be small and noisy, thus statistical tests are usually performed through Bayesian estimation or comparison of bootstrapped mean amplitudes (e.g. Meixner & Rosenfeld, 2014). Such methods enable researchers to estimate whether the unknown stimulus elicits a waveform more like that of one or the other type of known stimulus.

There are, of course, several ways the GKT can go wrong. The simplest way to attempt to fool the test might be to ignore the known-familiar stimulus, but then the test is

not failing so much as the suspect is simply refusing to take it. A criminal wise to the method might assign false importance to another item in the list and watch for it to reappear, thus eliciting a P300 to it as well and inflating the P300 amplitude to filler items. A bootstrapping procedure has been shown to be somewhat resistant to this countermeasure (Winograd & Rosenfeld, 2011), but it nonetheless remains possible and certainly reduces the accuracy of the GKT's classification of suspects as guilty or innocent. Complications could also arise if there is, by accident, a known-unfamiliar stimulus that happens to be familiar or significant to the suspect for reasons unknown to the investigators. If, for example, a guilty knife-wielder from the above example was a knife collector and saw one or several from his collection in the list, the amplitude of his P300 to the knife used in the attack would be reduced and the amplitude of his P300 to all the other familiar knives might be increased. The current study proposes to apply the GKT to witnesses by showing sets of faces, so both the desire to fool the test and the chance of accidentally familiar items appearing is low.

However, there is still a possibility that witnesses viewing a GKT composed of faces might, in the course of seeing a set of faces several times in a row, begin to notice small differences in the faces that serve to weaken the unfamiliarity of the known-unfamiliar faces. It is critical that the faces match the description well enough that they cease to stand out as individuals for most of the procedure. It is equally critical, though, that the suspect and the known-familiar face remain discoverable. On the other side, if the suspect is too unique within the set, he may elicit a strong P300 just by virtue of his perceived physical difference. Thus there was a delicate balance to be achieved in preparing the materials for the current study.

Some researchers have already looked to apply the ERP technique to detect face recognition (Sun, Chan, & Lee, 2012; Treese, Johansson, & Lindgren, 2010) and lineup performance (Friesen, 2010; Lefebvre, Marchand, Smith, & Connolly, 2007, 2009). The work of Lefebvre et al. (2007, 2009) is most similar to the current study. Lefebvre et al. (2007) used four crime videos (each depicting the same 60s crime but with a different male perpetrator and female bystander/victim, approximately 15s of exposure to the criminal) and experimented with varying time delays between video and lineup presentation. Participants saw sequential lineups of 6 faces that repeated 40 times, then rated confidence that each face was the culprit at the end. CP lineups contained the culprit, five photos matched approximately to the appearance of the culprit, and the victim to encourage attentive responding. For CA lineups, the criminal was replaced by another face found via the same search method. Each crime had a wholly unique set of faces. Lineups were pilot tested and found to be unbiased. Each participant completed one CP and one CA lineup in immediate test conditions in addition to a CP lineup for the 1-hour and 1-week delay conditions.

Grand average P300 amplitude across participants was larger to the culprit than to fillers when collapsing across all central parietal electrodes and time delays. Also, P300 amplitude for correct ID's was larger than P300 amplitude for a falsely identified foil, which was in turn higher than P300 to unselected filler faces. However, 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 effect comes from the nature of the GKT task, namely that the participant must have a strong memory of the critical item for it to register as an oddball and elicit a P300. It is

somewhat surprising, then, that the authors found group differences when many of the participants presumably did not have strong memories of the culprit. For CA lineups, when averaged over all participants, no one face had a significantly higher P300 than any others. P300 amplitude for culprit selections was larger than that for any false positives from CA lineups.[3]

Lebevbre et al. (2009) followed up by replicating their previous study with new but similar videos and adding an instruction to deceive the experimenters. The authors' main stated goal with this follow up was to individually classify identifications of culprits, regardless of each participant's instruction to lie or tell the truth. The manipulation of delay was removed. New lineups were designed to accompany the new videos, and the authors assumed the lineups were nonbiased because they were created using the same description-matching method as those in the 2007 paper. They describe the filler faces as having "some overlapping attributes with the culprit." The authors also tested differences between two methods of data analysis: (A) comparing bootstrapped averages of P300 amplitude elicited by the culprit versus average amplitudes elicited by all the other faces and (B) comparing the bootstrapped averages elicited by culprit to those elicited by the foil with the next highest amplitude. The authors were able to identify culprit photographs from the individual ERP's of every participant in the truth condition using both methods and 18 of 20 participants in the lie condition using method A. However, method A produced more false positive ID's to CA lineups (reanalyzed data from the 2007 dataset).

[3] As is typical for CA lineups, interpretation of confidence ratings was less clear than for CP lineups.

Levebvre and colleagues did important work in finding that eyewitness ID accuracy could be assessed by 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 could be made more ready for real world application. First, the current study more closely mirrored real-world witnessing situations by showing each participant only one video. Second, witnesses were exposed to the culprit for much longer and told before the crime video that the later goal will be to identify the culprit from a lineup. Though there is some debate about asking witnesses to actively process criminal faces, this method mimicked a scenario in which a participant would be motivated to lie about recognizing the culprit (e.g., because the witness has been threatened against identifying anyone). Third, the Lefebvre et al. procedure used the victim as the known-familiar member of the lineup, which would lead to confusion if the victim matched the description of the culprit. Changing the second target face into a known-familiar based on the method used in typical GKT research addressed this issue. It may also be possible that splitting the identification decisions into three types enabled lying participants to ignore the culprit, as participants never used the button to which the culprit was assigned. In the current study, the culprit and the learned face were grouped together as "known" faces. Last, the bootstrapping method used by Lefebvre et al. would be unnecessary if the procedure evoked a larger P300 to the culprit. The current procedure included more filler faces to make the culprit a more infrequent oddball in order to evoke a larger P300. These changes along with a unique set of materials enabled the current study to expand on the findings of the Lefebvre team.

Current Study

The basic goal of an ERP lineup is to uncover any evidence of oldness of the criminal's face in the participant/witness's memory. Uncovering this evidence would only be useful to the criminal justice community if it were applicable in situations in which the criminal would not otherwise be identified by the witness's overt decision or confidence level. Since the GKT 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 criminal easily but are compelled to falsely claim that they do not because of a threat from associates of the criminal or because the witness is secretly a co-conspirator. As Canadian law considers refusing to identify a known culprit to be perjury, ethical and legal guidelines around extracting an identification from an unwilling witness this study does not propose to use the ERP lineup as such a tool (Farah, Hutchinson, Phelps, & Wagner, 2014, deal with these issues in reference to lie detection with fMRI). A witness whose ERP lineup provided evidence that put away a connected criminal might still be considered for retribution from their associates, as we assume such folks would be uninterested in splitting hairs between a verbal and a neural identification. The real world niche of this test is more likely as an aid in the variety of scenarios in which witnesses wish to be cooperative but are unable to do so. 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 could prove useful. Witnesses with normal brain activity who are unable to verbally or physically identify the culprit could also make an identification through this method. The test would be best deployed in scenarios in which witnesses may not identify the criminal but are not purposefully lying.

Method

Participants and procedure. Participants ($N = 48$) were recruited through the UVic psychology participation pool of undergraduate students who were compensated with course credit. Participants reported an average age of 23.34, with a range from 18 to 43. 12 were male, 36 female, 4 left-handed, and 44 right-handed. None reported regular seizures or recent brain injury.

Participants entered the lab and were asked to find a comfortable, seated position in front of a 19-inch computer monitor in an electromagnetically shielded booth. After providing consent, they were informed of the procedure. Participants watched one of two crime videos, and then watched it a second time while narrating it aloud. Each video depicted a young male of the same physical description commit a car theft in almost-identical scenes. Participants were instructed to pay attention to the culprit "Because we will ask you to identify him later." Participants then completed a yes/no recognition test in which they studied the face of a man called “Chris” for as long as they liked then responded to a series of test faces (inspired by the Joe/No Joe paradigm in Tanaka, Curran, Porterfield, & Collins, 2006). For each participant, Chris was the criminal from the video they did not see, thus he matched the same description. Then, in a series of 6 yes/no recognition memory practice trials (4 cycles) they were instructed to identify Chris. If a participant did not achieve 80% correct decisions, the study/test cycle was repeated.

In the next phase, the ERP lineup, a set of 10 new faces matching the description of the culprit (none of which were used during Chris training) were cycled 30 times with appearances by Chris and the culprit in each cycle while EEG data were recorded. The

procedure thus displayed 12 faces 30 times in 6 different pseudo-random orders. 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). Thus, familiar faces were not repeated without at least 2 intervening unfamiliar faces. Participants were told that there were two faces to recognize, Chris and the culprit, and to press a button every time they see either known face. Buttons were the 'm' and 'z' keys on the keyboard, and the category of each button was counterbalanced between subjects. 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 seconds 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 button was pressed.

Just before the lineup cycle began, half of participants 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 Chris and to pretend to not recognize the culprit by categorizing him as a "New face" according to the categories established by the computer program. Upon completion of the 30 cycles, participants were presented with a classic culprit-present simultaneous lineup for the learned face "Chris" and a separate

culprit-present lineup for the criminal. Participants who were originally asked to lie were asked to not do so any longer, thus these identifications served as proof that the witness did, in fact, recognize both faces during the previous task. Confidence ratings were also collected at this stage. Participants who did not correctly identify either Chris or the culprit from the simultaneous lineups were removed from analyses ($N = 7$ did not ID culprit, 6 in lie condition) as were those who missed identifications of one or the other in more than 20% of the ERP lineup phase ($N = 2$). These exclusions left 39 participants' data to be analyzed. Two more participants made errors on exactly 20% of trials (6 of 30). Of the 12950 trials seen by participants left in the sample, 206 were removed (1.6%).

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 20kΩ. One electrode was placed on each of the left and right mastoids, and the EEG was recorded using the average reference. 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 face type (suspect, learned face, unfamiliar faces). A 1600ms epoch of data extending from 100ms prior to and 1500ms 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 100ms interval preceding stimulus onset from each sample. Muscular and other artifacts were removed using a ±50μV step threshold as a rejection criterion.[4] These rules lead to 3 participants losing more than 20% of ERP lineup trials; their data were not further analyzed. Of the remaining 12960 trials, 206 were removed because participants responded incorrectly to the stimulus (1.6%). Of the remaining 12754 trials, 73 were removed based on the artifact rejection algorithms (0.6%). ERP's were then created for each electrode and participant by averaging single-trial EEG according to face type (criminal, Chris, and foils).

Statistical analysis. The P300 was calculated at Pz, where it typically reaches maximum amplitude in the literature (Fabiani, Gratton, Karis, & Donchin, 1987). For each participant, the P300 was defined as the maximum ERP voltage between 450ms and 550ms, as the maximum voltage occurred between these two time points for all participants.

[4] EOG channels were not included in artifact rejection.

Results

Behavioral data. Of the participants included in the ERP analysis, 20 made fewer than 3 errors on the ERP lineup, 12 participants made between 4 and 8 errors, and 4 participants made more than 15 errors. Overall, each participant made errors on fewer than 7% of trials, and most made errors on fewer than 4%. Errors were removed on a by-trial basis for the ERP analysis.

P300. The grand average waveforms for participants not excluded by any of the behavioural or EEG criteria described above are shown in Figures 5 and 6. Scalp distributions are shown in Figures 7-12. Paired samples *t*-tests were conducted on the amplitude of the P300 at its maximal location between 450 and 550ms. Bonferroni corrections were applied to the series of 4 tests, leading to an alpha level of 0.0125. When participants truthfully identified the culprit, P300 amplitudes evoked by Chris and the culprit were not different from each other, $t(20) = 0.69$, $p = .50$, Cohen's $dz = 0.10$, 95% CI [-0.17, 0.36], and amplitudes evoked by the culprit were different from those evoked by foil faces, $t(20) = 11.51$, $p < .0001$, Cohen's $dz = 1.68$, 95% CI [1.08, 2.28], see Figure 5. When participants lied by not identifying the culprit, P300 amplitudes evoked by Chris and the culprit were not different from each other, $t(14) = 0.42$, $p = .68$, Cohen's $dz = -0.15$, 95% CI [-0.79, 0.49], and amplitudes evoked by the culprit were different from those evoked by foil faces, $t(14) = 4.5$, $p = .0004$, Cohen's $dz = 1.12$, 95% CI [0.48, 1.76], see Figure 6.

Discussion

Groupwise effects confirmed the hypothesis that P300 amplitudes evoked by Chris would be indistinguishable from P300 amplitudes evoked by the culprit while P300 amplitudes to the foil faces would be different from those evoked by the culprit across both conditions. Truth tellers produced larger amplitude P300s to the critical faces than did liars, likely because P300 amplitudes are typically larger for tasks that require active responding (Bennington & Polich, 1999).

The intention with the current design was to present the culprit as a very infrequent oddball to elicit a large P300. That a P300 was evoked by both the learned face and the culprit can be informative for future researchers in that memory of a face need not be to the strength of personal familiarity to be detected using this ERP lineup if the situation allows for groupwise analysis. Because more participants failed to identify the culprit from the simultaneous lineup in the lying condition, it may be possible that lying through the ERP lineup lead them to be less certain when they saw the final, simultaneous lineup.

These findings build upon those by Lefebvre and colleagues (2007; 2009), as they aimed to distinguish correct identifications from false positives as well as lies from true lineup rejections. An interesting 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 buttons for the culprit, the learned target face, and the fillers whereas we grouped the culprit and the learned face together into a "known faces" group that required the same button response. The current study also used a longer identification task with 10 filler faces compared to Lefebvre et al.'s 5 fillers. 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. Thus, we provide more evidence of the utility of the ERP lineup as a method of identification for witnesses who are physically or otherwise unable to identify the culprit despite having high levels of familiarity with the person.

Future analyses. A number of analyses of other components might be informative in these data in the future. It is possible that the feeling of a correct selection elicited a reward positivity component. We may also see a general increase in positivity across the entire waveform for familiar compared to unfamiliar faces. This difference seems likely to be found given the differences further from presentation of the face in the truth-telling condition, but this effect has been inconsistent in the GKT literature.

The literature on face recognition and its typical ERP components as indices of memory indicate that finding any differences between familiar and new faces will be unlikely except for the N250 component (Caharel, Poiroux, Bernard, Thibaut, Lalonde, & Rebai, 2002; Eimer, Gosling, & Duchaine, 2012; Trenner, Schweinberger, Jentzsch, & Sommer, 2004), as N170 differences and all post-300ms effects are mixed. We would expect to see a larger N250 for faces the participants have seen before as opposed to new faces. Analysis of N250 amplitudes would enable additional tests of whether culprits were truly recognized as known faces, and this test could be informative in cases where the P300 effects are unclear. It seems plausible, though, that N250 effects would take stronger familiarity to appear in these data. For lineups in which the culprit is present, this difference may increase over the course of the procedure. There is reason to expect

that familiarity with all the faces (including fillers) presented in the ERP lineup grew through repetition within the procedure of the current study, but this effect should be compounded for the task-relevant face (as shown by Tanaka et al., 2006 and Friesen, 2010) and thereby should not reduce the P300 or N250 amplitude to the learned and culprit faces relative to components evoked by filler faces. A split-half analysis will be critical to this point.

With P300 data, future work should also aim to categorize ERP waveforms elicited by suspects as more similar to those elicited by celebrities or those by unfamiliar faces. A bootstrapped method of comparison would enable differentiation between truth tellers and liars on an individual basis (Meixner & Rosenfeld, 2014; Lefebvre et al., 2007 & 2009). If suspects are categorized as familiar faces, future directions will include combination with other measures to develop a test of likelihood of correct identification for eventual use by police.

Limitations and future directions. In any P300-based lie detection study, the possibility arises that an element of the test was a trigger for some other emotion. Perhaps some participants recognized another face in the set, or perhaps one face looked like someone else in a participant's life. Either situation would have caused that face to evoke a P300. While every effort was made to make the face set as close to the ideal between homogeneity and heterogeneity, a small number of participants and several research assistants pondered aloud after completing the study about whether one of the two culprits had blonder hair than the rest of the photographs in the set. If he truly stuck out in this way, then any oddball effects to the presentation of his face could have been due to his different hair color. The experimenters did not find the difference obvious, thus he

was used throughout the study and no participants were removed for mentioning that they noticed a difference. Nonetheless, future studies with these materials should contain a new video with a more closely matched culprit, should have a higher occurrence of blonde haired men in the ERP lineup slideshow, or should digitally alter the video and/or photo of this culprit to darken his hair. Other participants found the other culprit to be quite similar to one of the fillers in the lineup, and it is indeed true that these two faces are used as suspects for one another in a different study exactly because they look quite similar. Thus, that filler might have also elicited a P300, especially for participants who falsely identified him. He should be removed from the set in a future study.

To add to this phenomenon, of the four participants who made more than 6 errors in the 30 critical trials during the ERP lineup all made those errors by not identifying this second culprit. Perhaps the presence of his doppelganger confused participants, as most errors were early in the sequence. Some of these missed identifications may have also been a result of the instructions. In pilots of the procedure, participants reported having watched for the culprit to reappear later in the list from a different viewpoint (perhaps another result of the doppelganger's presence). Instructions were modified to counteract this type of searching, but perhaps the note that the faces would repeat many times and not change viewpoints lead some participants to take each individual identification less seriously because they knew many more would take place.

Beyond those participants who made errors during the lengthy and repetitive ERP task, seven others failed to identify the culprit from the simultaneous lineup at the end of the procedure. Most participants made errors on the final lineup by rejecting it, thus less data might be excluded from a future sample if participants were forced to choose

someone on the final lineup. An experimenter might maintain the current procedure and add a forced choice lineup for any participant who rejects the final lineup, as such a witness might have different ERP responses to the culprit during the ERP lineup. One could also position the video after the Chris study/test cycle in the procedure. It is additionally interesting to note that six of these seven excluded participants rejected the final lineup after lying throughout the ERP lineup. Putting aside the possibility that participants misunderstood that the instruction to lie had finished, it could be the case that memory for the culprit at the end of the procedure was hindered by efforts to not remember him during the ERP lineup, as is often found in the unwanted memory suppression literature (Anderson et al, 2004). Investigating the EEG of these participants may also be interesting, as there is some evidence that actively trying not to remember an item can conceal guilty knowledge in the ERP GKT (Bergström, Anderson, Buda, Simons, & Richardson-Klavehn, 2013).

General Discussion

In Baldassari et al. (under review), we aimed to determine the exact nature of the relationship between face recognition skill and lineup skill. This paper presents the next studies in that line of work, through which the end goal is to develop a tool that can be combined with other known predictors of eyewitness identification accuracy in a larger model that can account for as much of the variance among witnesses as possible. We can certainly conclude that there are measureable individual differences in both face recognition skill and proclivity to choose on a face memory task. The studies presented in Chapter 1 continue development of a test that reliably measures both sides of accuracy on a face memory task. The potential impact of a LST that is strongly correlated with lineup

accuracy is of import in the world of policing and investigation, as investigators may not be correctly estimating the likelihood that their witnesses have made accurate lineup decisions. Data from a variety of labs showed that judges (Benton, Ross, Bradshaw, Thomas, & Bradshaw, 2006; Brigham & Wolfskeil, 1983; Wise & Safer, 2004) and students acting as mock jurors (Cutler, Penrod, & Stuve, 1988; Desmarais & Read, 2011; Yarmey & Jones, 1983) did not accurately report the factors that make a good eyewitness. Students playing the role of investigator for a crime have been similarly unable to differentiate good from bad witnesses (Boyce, Lindsay, & Brimacombe, 2008; Dahl, Lindsay, & Brimacombe, 2006; Lindsay, Nilsen, & Read, 2000; MacLean, Brimacombe, Allison, Dahl, & Kadlec, 2011). Rather than attempt to train police investigators and jurors to assess witness accuracy, the field would be better served by developing an objective test of the likelihood a given witness will respond accurately to a lineup and teaching investigators and jurors how to interpret the results of such a test. The LST would be the first step toward a much larger measure that would include other known predictors of eyewitness accuracy such as confidence, latency, decision-making style, and perceptual basis (Kaminski & Sporer, 2017). The variance in witness accuracy accounted for by these individual differences could be buoyed by measures of the distance between the criminal and the witness, the lighting at the scene, the length of time the witness could see the criminal, and the presence of anything that might make identifying more difficult (such as a hat, a disguise, or smoke).

Until the LST is combined with these other measures to produce a strong predictor of eyewitness accuracy, the best thing police investigators can do to minimize errors is to try to only use eyewitness identification as a strong indicator of guilt when

there is already good evidence of that guilt (Wells, Yang, & Smalarz, 2015). If the base rate of suspects placed in lineups who are actually guilty is high, then much of the eyewitness identification research will merely scuttle on the edges. Unfortunately, it seems that this is not the case (Greene & Evelo, 2015), and that some precincts insert suspects into lineups with little other evidence of guilt as regular practice. I recommend that they avoid doing so until more objective measures of eyewitness accuracy are in place.

In situations where the witness does not identify the culprit but should have a strong memory of him, the ERP lineup examined here could be a welcome addition. One immediately plausible use of this test would be for a witness in a very important or high profile case who could not speak or move. The ERP lineup could enable this witness to identify the culprit. Then, this lineup administration method could function more like a P300 speller in extracting a statement from an otherwise incommunicado witness and would be worth the cost (Krusienski, Sellers, McFarland, Vaughan, & Wolpaw, 2008). Further, this study was the first to demonstrate the large P300 evoked by the culprit and a learned face when participants intend to tell the truth. If individual bootstrapped means enable successful classification of truth tellers and liars, then the lack of perfect identification accuracy in the ERP and the simultaneous lineup across both culprits could be viewed as additional support for the utility of the procedure for a nervous truth teller, if participants who made false identifications of filler faces or missed the culprit on occasion during the ERP lineup nonetheless produced larger P300 components to the face of the actual culprit.

These results also have a potential impact on the eyewitness identification literature. Some labs, primarily those of and surrounding Brewer, already use mini-lineups based on study photos (albeit with a slightly different structure) as stand-ins for full sized lineups based on videos (Sauer, Brewer, & Weber, 2008; Weber & Varga, 2012). That the LST only accounts for a small portion of the variance in lineup performance suggests mini-lineups in a photo based face recognition task are not perfect one-to-one stand-ins for full sized lineups based on videos. Psychologists should be careful about publicizing conclusions based only on mini-lineup data before testing the same questions on full sized lineups.

From a basic science perspective, the LST and the ERP lineup both contribute to their respective literatures. The so-far moderate success in predicting witness skill shown by the LST speaks to the specificity, and perhaps a lack of transfer of processes, of the task. The disconnect between lineup research and real-world lineups has been discussed at length in the eyewitness literature, but this study is one of few that has measured the connection between lineup research and other face memory research. Though the literature within face recognition has shown strong predictability across face recognition tasks, their overlap in necessary skills may be due to the specific nature of laboratory exposure to face stimuli. Most studies expose participants to new faces with a single photograph, some use several photos, but very few begin with video or lifelike exposure. This type of exposure works well when both tasks use it, but it appears that the transfer of static face acquisition to dynamic, real-world face acquisition may not be viable. I posit that this disconnect is a result of the difference between automatic and intentional memory as detailed by Jacoby (1991). For all but the very-experienced (perhaps highly

trained) witness, any memory related to the appearance of a perpetrator would be automatic. Someone who has testified as a witness before or a witness with police or military training might be able to focus their attention to intentionally memorize the culprit's face, but such a witness would also be distracted by all the other aspects of the crime at hand. In fact, some famous cases of false imprisonment include witness statements that detail intentional processing, so the benefit to memory of actively attempting to memorize the culprit's face is negligible. This line of reasoning supports the holistic theory of face processing, as intentional memorization often leads participants to investigate individual features rather than to take in the face as a whole. Participants are not warned of the coming lineups before they watch the videos, thus they watch them somewhat naturally. Some may focus on faces, but others likely focus on the details of the criminals' actions or the social moments within the videos.[5] In contrast to the mostly automatic processing of a criminal's face at the scene, the LST demands only intentional processing. The same distinction can be made between the culprit and the known-target face in the ERP lineup. Participants are told they are taking a memory test, and there is no distraction from the faces as they appear. As the eponymous goal of the LST is to measure lineup skill, it would be closer to the real experience if participants learned LST faces using automatic processing. A different study phase, such as watching a video of people in a mall followed by a LST containing folks from the mall scene, might be more

[5] Some data show that preparing a witness for a forthcoming lineup may not affect their accuracy (Shapiro & Penrod, 1986; Yarmey, 2004). It may be the case that intentionally processing the culprit's face leads to unhelpful focusing on features or is negatively affected by other distracting aspects of the crime, thus leading witnesses who intentionally learn the face to fail to outperform other witnesses who learned the face automatically. Whether this change in processing is reflective of a completely different mental skill seems unlikely given the correlations between tasks like the CFMT and the Cambridge Face Perception Test (Fysh & Bindemann, 2017).

predictive of lineup accuracy. Another potential idea would be to take a sample of undergraduate students in a 100-level course in a fairly large university such that their exposure to one another is mostly limited to time spent in class. A LST containing only classmates might be an interesting predictor of eyewitness accuracy.

The ERP lineup is the second implementation of a new use of the P300, namely as a lie detector with face stimuli. This test importantly shows the utility of the P300 in detecting lies with a second set of materials and using a longer procedure to ease the statistical burden. It remains to be seen whether these findings would generalize to other materials sets, but the most important future direction of this line of work should be to find its boundary conditions. At what strength of memory will the criminal in an ERP lineup still evoke the P300? How similar or dissimilar can the faces in the set be, if a particular police precinct is limited in the photos they can access?

Both approaches detailed here represent early efforts to apply so far underutilized realms of cognitive psychology to the obvious problems in eyewitness identification in the real world. The LST measures some of the variance in individual differences in face memory to enable prediction of lineup accuracy that would not be possible with a traditional groupwise analytic approach. The ERP lineup also enables individual analysis through an oddball paradigm with many exposures to extract identifications from unable or nervous witnesses. Continuing to improve on these procedures with the suggestions herein will undoubtedly result in reliable, useful tests that will inform both police practice and basic memory research.

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.

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality And Individual Differences*, *60,* 36-40.

Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., Glover, G. H., & Gabrieli, J. E. (2004). Neural systems underlying the suppression of unwanted memories. *Science*, *303*(5655), 232-235.

Baldassari, M. J., Kantner, J. D., & Lindsay, D. S. (Under review). Presenting the Lineup Skills Test: A 2-alternative unforced-choice face recognition task. *Cognitive Research: Principles and Implications*.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal Of Autism And Developmental Disorders*, *31*(1), 5-17.

Begleiter, H., Porjesz, B., & Wang, W. (1995). Event-related brain potentials differentiate priming and recognition to familiar and unfamiliar faces. *Electroencephalography & Clinical Neurophysiology*, *94*(1), 41-49.

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.

Bergström, Z. M., Anderson, M. C., Buda, M., Simons, J. S., & Richardson-Klavehn, A. (2013). Intentional retrieval suppression can conceal guilty knowledge in ERP memory detection tests. *Biological Psychology*, *94*(1), 1-11.

Bindemann, M., Brown, C., Koyas, T., & Russ, A. (2012). Individual differences in face identification postdict eyewitness accuracy. *Journal of Applied Research in Memory and Cognition, 1*, 96-103.

Bobak, A. K, Hancock, P. J. B., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology, 30*, 81-91.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., Rivolta,

D., Wilson, E., & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology, 26*(5), 423-455.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*, 338-360.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.

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.

Darling, S., Martin, D., Hellmann, J. H., & Memon, A. (2009). Some witnesses are better than others. *Personality and Individual Differences, 47*, 369-373.

Donchin, E. (1981). Surprise! … Surprise?. *Psychophysiology*, *18*(5), 493-513.

Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. *Neurology*, *62*(7), 1219-1220.

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.

Duchaine, B. & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. *Neuropsychologia, 41*, 713-720.

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.

Fabiani, M., Gratton, G., Karis, D., & Donchin, E. (1987). The 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.

Fresco, D. M., Coles, M. E., Heimberg, R. G., Leibowitz, M. R., Hami, S., Stein, M. B., & Goetz, D. (2001). The Liebowitz Social Anxiety Scale: A comparison of the

psychometric properties of self-report and clinician-administered formats. *Psychological Medicine*, *31*(6), 1025-1035.

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.

Fysh, M. C., & Bindemann, M. (2017). The Kent face matching test. *British Journal Of Psychology*, doi:10.1111/bjop.12260

Geiselman, R. E., Tubridy, A., Bkynjun, R., Schroppel, T., Turner, L., Yoakum, K., & Young, N. (2001). Benton Facial Recognition Test scores: Index of eyewitness accuracy. *American Journal of Forensic Psychology, 19*, 77-88.

Gill, O., & Polich, J. (2002). P300 stimulus sequence effects in children and adults. *Perceptual and Motor Skills, 94*, 509-520.

Granhag, P. A., Ask, K., & Giolla, E. M. (2014). Eyewitness recall: An overview of estimator-based research. In D. S. Lindasy & T. J. Perfect (Eds.), *The SAGE handbook of applied memory* (pp. 541-558). New York, NY: SAGE Publications.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55*, 468-484.

Greene, E., & Evelo, A. J. (2015). Cops and robbers (and eyewitnesses): a comparison of lineup administration by robbery detectives in the USA and Canada. *Psychology, Crime, & Law, 21*, 297-313.

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.

Guillaume, F., & Tiberghian, G. (2013). Impact of intention on ERP correlates of face recognition. *Brain and Cognition, 81*, 73-81.

Hosch, H. (1994). Individual differences in personality and eyewitness identification. In D. F. Ross, J. D. Read & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 328-347). New York, NY: Cambridge University Press.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541.

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. *Psychophymology 17*(2), 167-178.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40*, 1163-1177.

Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review*, *21*(5), 1272-1280.

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.

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.

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.

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.

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.

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

Lykken, D. T. 1959. The GSR in the detection of guilt. *Journal of Applied Psychology, 43*, 385-388.

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

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. Journal of personality and social psychology, 13(4), 330.

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.

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.

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.

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research, 69*, 10-22.

McKone, E., Hall, A., Pidcock, M., Palermo, R., Wilkinson, R. B., Rivolta, D., Yovel, G., Davis, J. M., & O'Connor, K. B. (2011). Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test—Australian. *Cognitive Neuropsychology, 28*(2), 109-146.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*(4), 865-876.

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69*(7), 1175-1184.

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.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, And Law*, *7*(1), 3-35.

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.

Minear, M. & Park, D.C.(2004). A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments, & Computers. 36, 630-633.

Morgan, C. A., Hazlett, G., Baranoski, M., Doran, A., Southwick, S., & Loftus, E. (2007). Accuracy of eyewitness identification is significantly associated with performance on a standardized test of face recognition. *International Journal of Law and Psychiatry, 30*, 213-223.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of

eyewitness memory. *Journal Of Experimental Psychology: General*, *137*(3), 528-
547.

Sauer, J. D., Brewer, N., & Weber, N. (2012). Using confidence ratings to identify a target among foils. *Journal Of Applied Research In Memory And Cognition*, *1*(2), 80-88.

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.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size to correlation stabilize? *Journal of Research in Personality, 47*, 609-612.

Schweinberger, S. R., Pickering, E. C., Jentzsch, 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.

Shapiro, P. N., & Penrod, S. D. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin*, *100*, 139–156.

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.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science, 150*(37), 1187-1188.

Tanaka, J. W., Curran, T., Porterfield, A. L., & Collins, D. (2006). Activation of Preexisting and Acquired Face Representations: The N250 Event-related Potential as an Index of Face Familiarity. *Journal Of Cognitive Neuroscience*, *18*(9), 1488-1497.

Tanaka, J.W. & Pierce, L. J.(2009). The Neural Plasticity of Other-Race Face Recognition. Cognitive, Affective, & Behavioral Neuroscience, 9(1), 122-131.

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.

Trenner, M. U., Schweinberger, S. R., Jentzsch, I., & Sommer, W. (2004). Face repetition effects in direct and indirect tasks: an event-related brain potentials study. *Cognitive Brain Research,* 21, 388-400.

Valentine, T. (2014). Estimating the reliability of eyewitness identification. In T. J.

Perfect & D. S. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 579-
594). New York, NY: SAGE Publications.

Valentine, T., Pickering, A., & Darling, S. (2003). Characteristics of eyewitness identification that predict the outcome of real lineups. *Applied Cognitive Psychology, 17*, 969-993.

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.

Weber, N., & Brewer, N. (2004). Confidence-Accuracy Calibration in Absolute and Relative Face Recognition Judgments. *Journal Of Experimental Psychology: Applied*, *10*(3), 156-172. doi:10.1037/1076-898X.10.3.156

Weber, N., & Varga, M. (2012). Can a modified lineup procedure improve the usefulness of confidence? *Journal of Applied Research in Memory and Cognition, 1*, 152-157.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *22*(6), 603-647.

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law And Human Behavior*, *39*(2), 99-122.

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

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.

Yager, J., & Iarocci, G. (2013). The development of the multidimensional social competence scale: A standardized measure of social competence in autism spectrum disorders. *Autism Research*, *6*(6), 631-641.

Yarmey, A. D. (2004). Eyewitness recall and photo identification: a field experiment. *Psychology, Crime & Law*, *10*(1), 53-68.

Appendix A – Figures and Tables

Figure 1. Proclivity to Choose scatterplot, y-axis jittered.

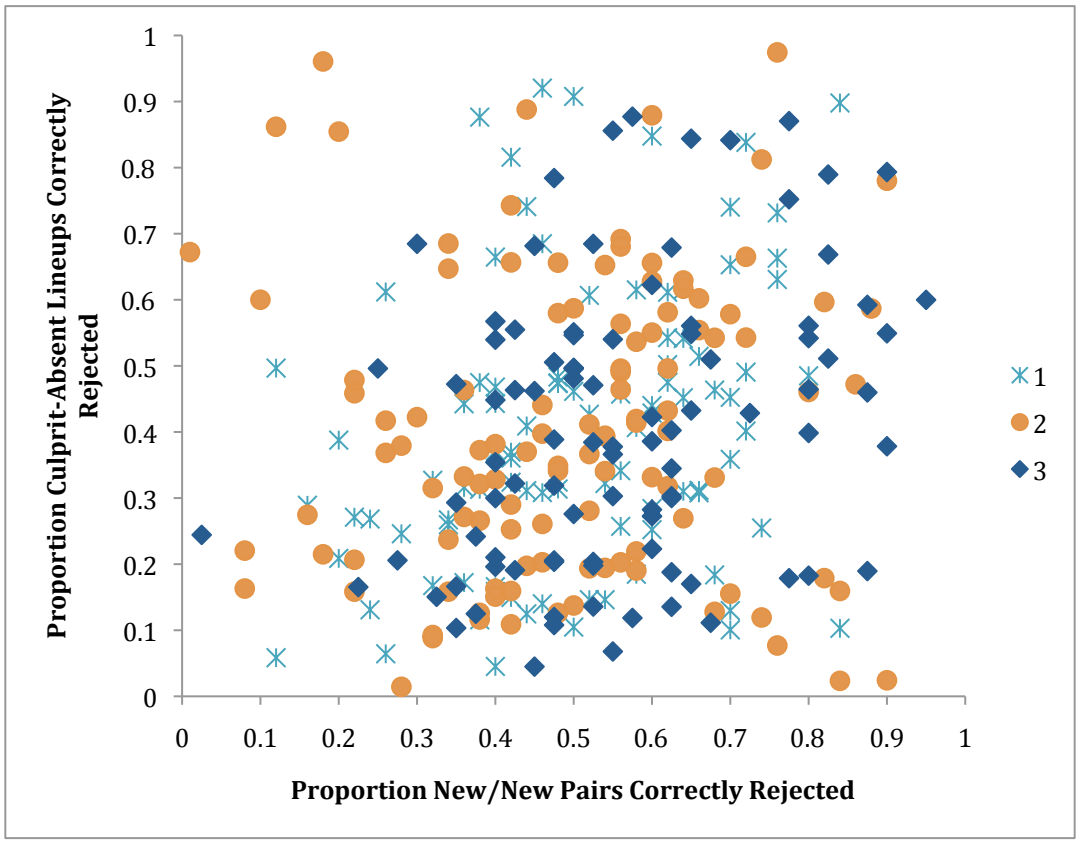

Figure 2. Face Recognition Skill scatterplot, y-axis jittered.

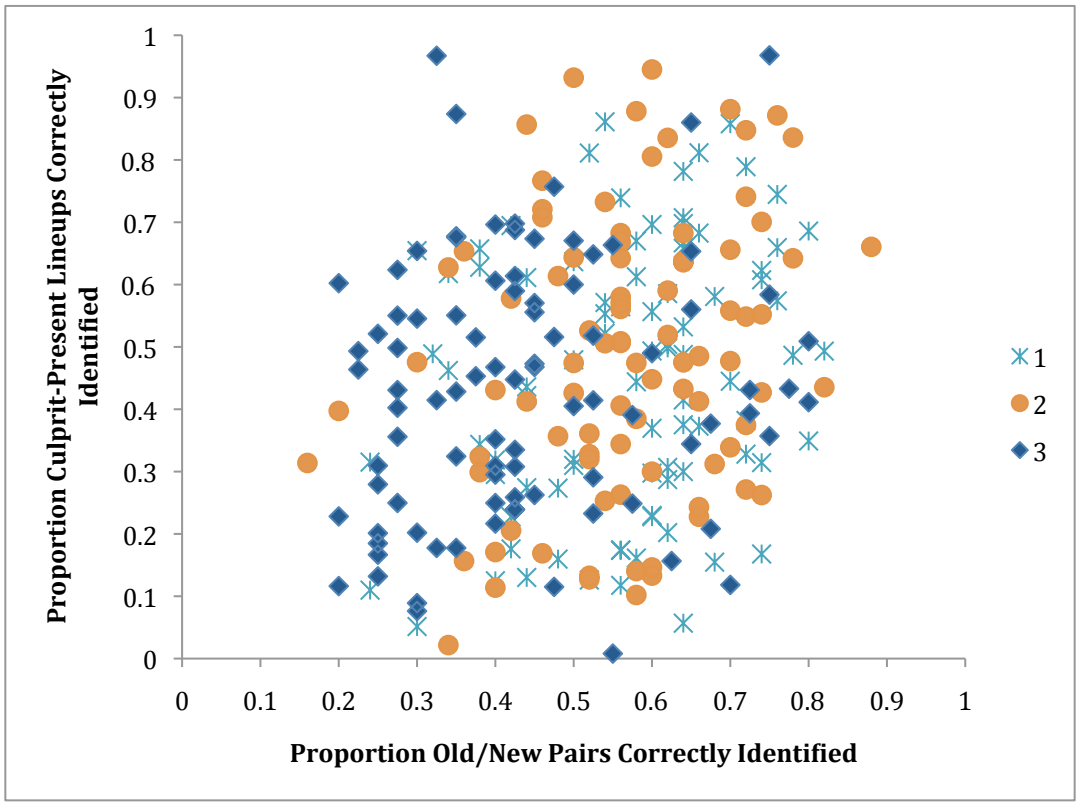

Figure 3. Item analyses by descent of photographed person, Study 1.

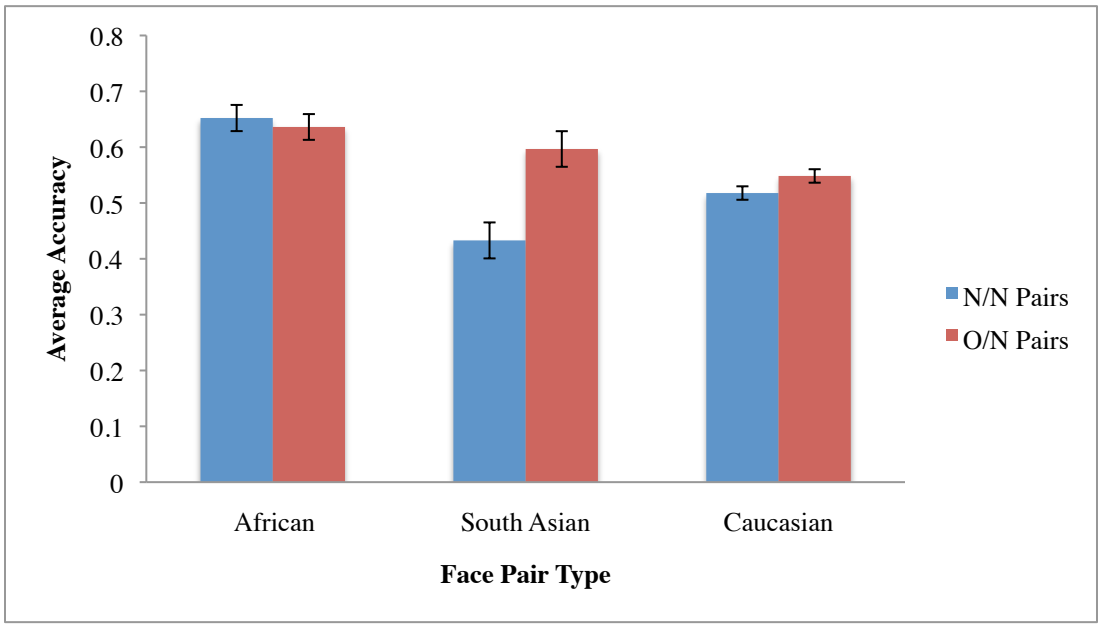

Figure 4. Theoretically ideal data for the traditional GKT/CIT P300 lie detectors.

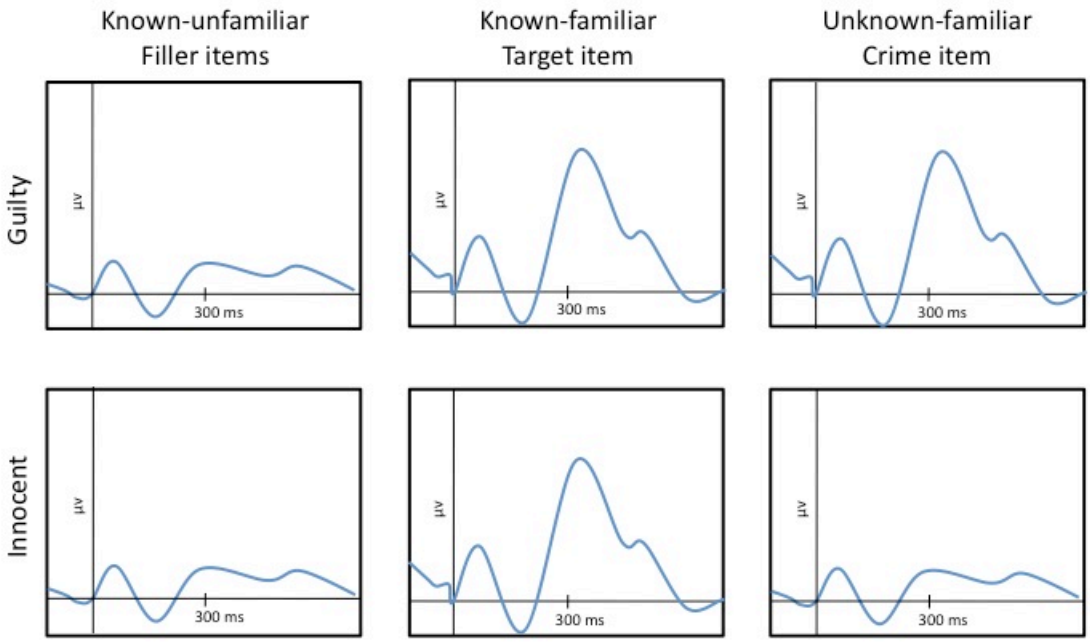

Figure 5. Groupwise ERP average waveforms for truth tellers with individual participant averages. 95% confidence ribbons around ERPs are basic nonparametric bootstraps without assuming normality (See osf.io/dzkez for r code and data).

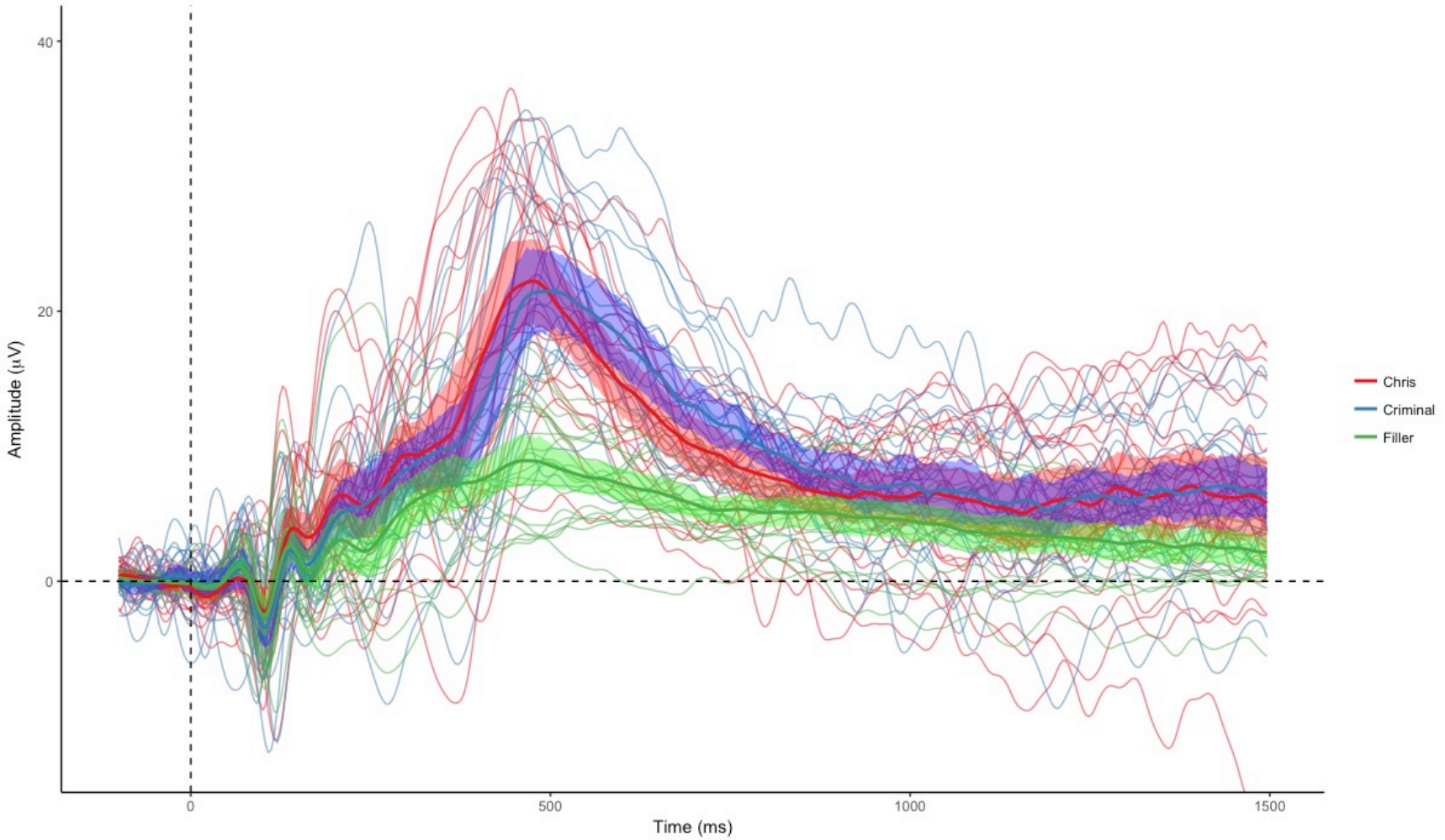

Figure 6. Groupwise ERP average waveforms for liars with individual participant averages. 95% confidence ribbons around ERPs are basic nonparametric bootstraps without assuming normality.

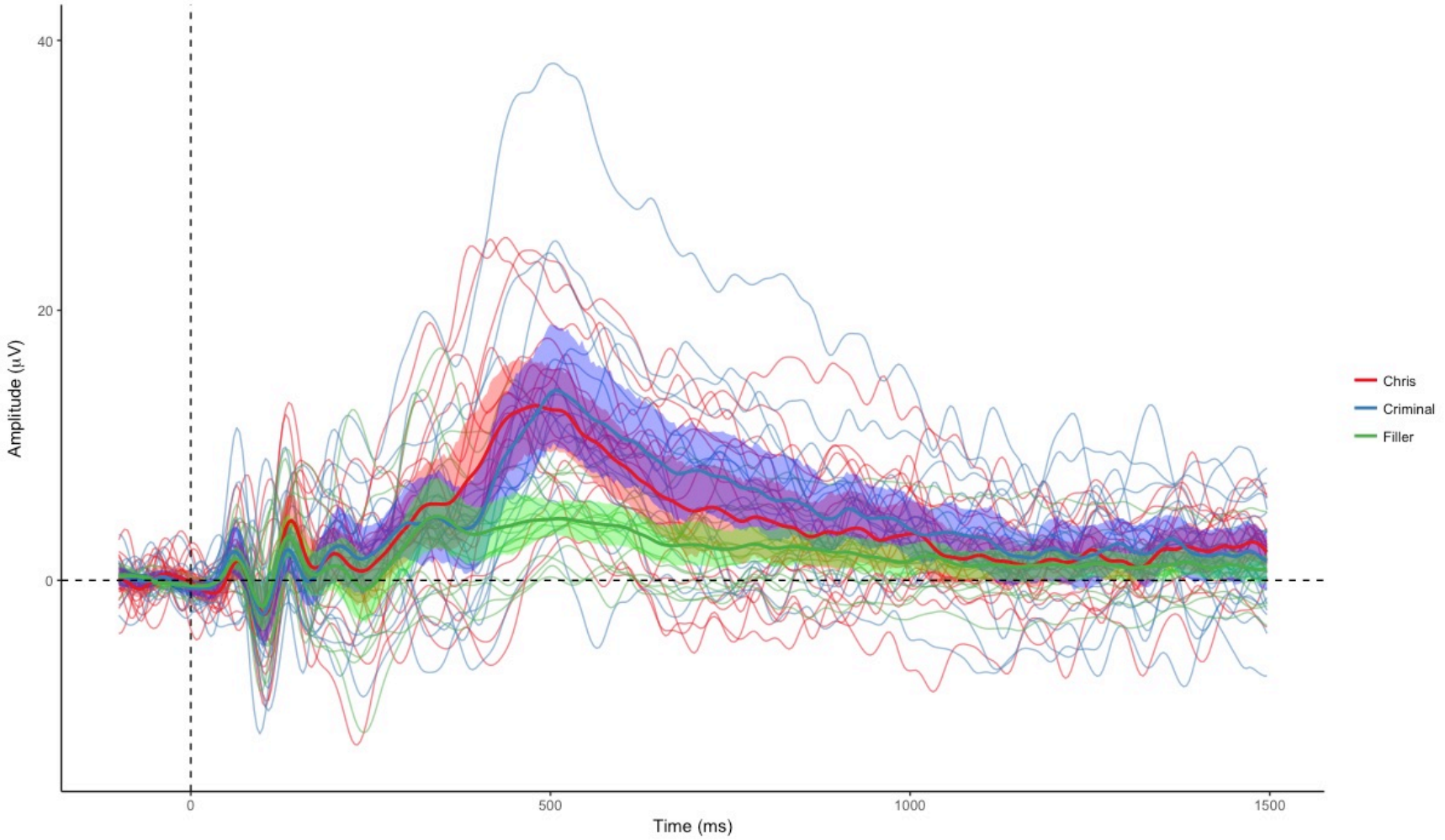

Figure 7. Scalp map of average response of truth-tellers to the face of the criminal across the ERP epoch.

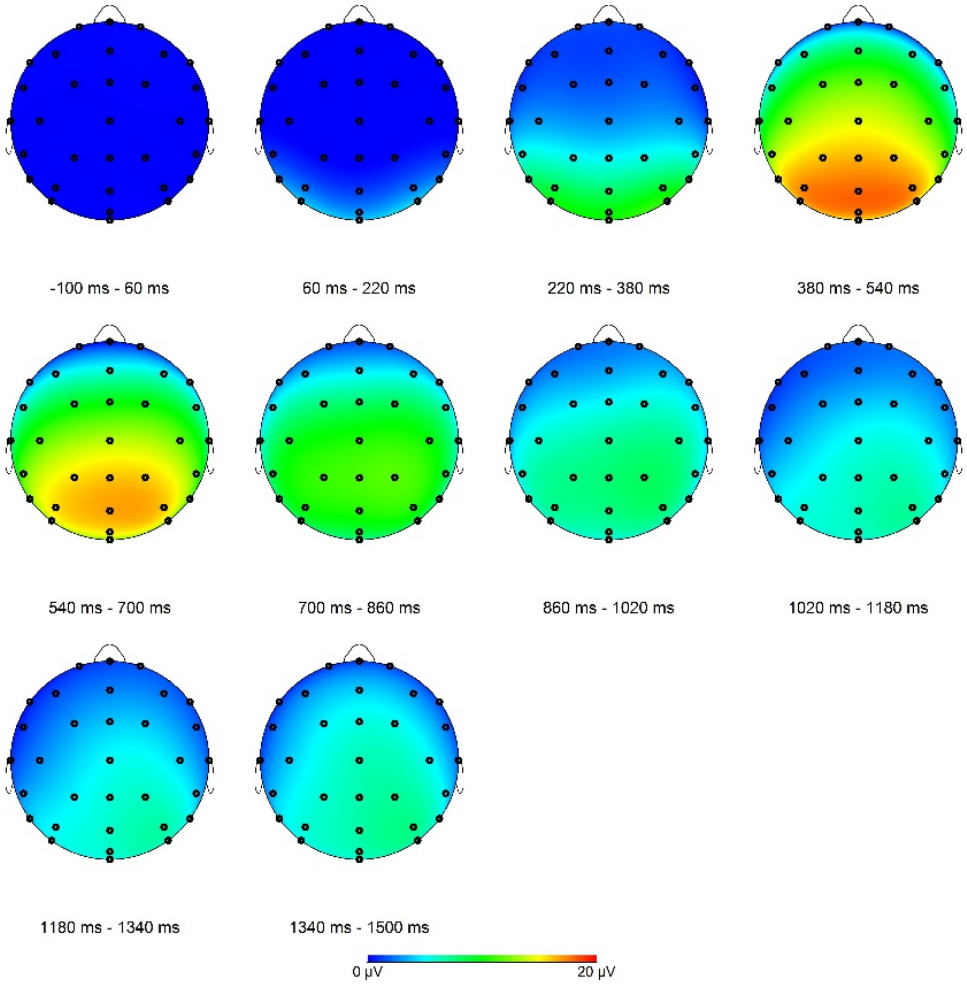

Figure 8. Scalp map of average response of truth-tellers to the face of the known lineup member (Chris) across the ERP epoch.

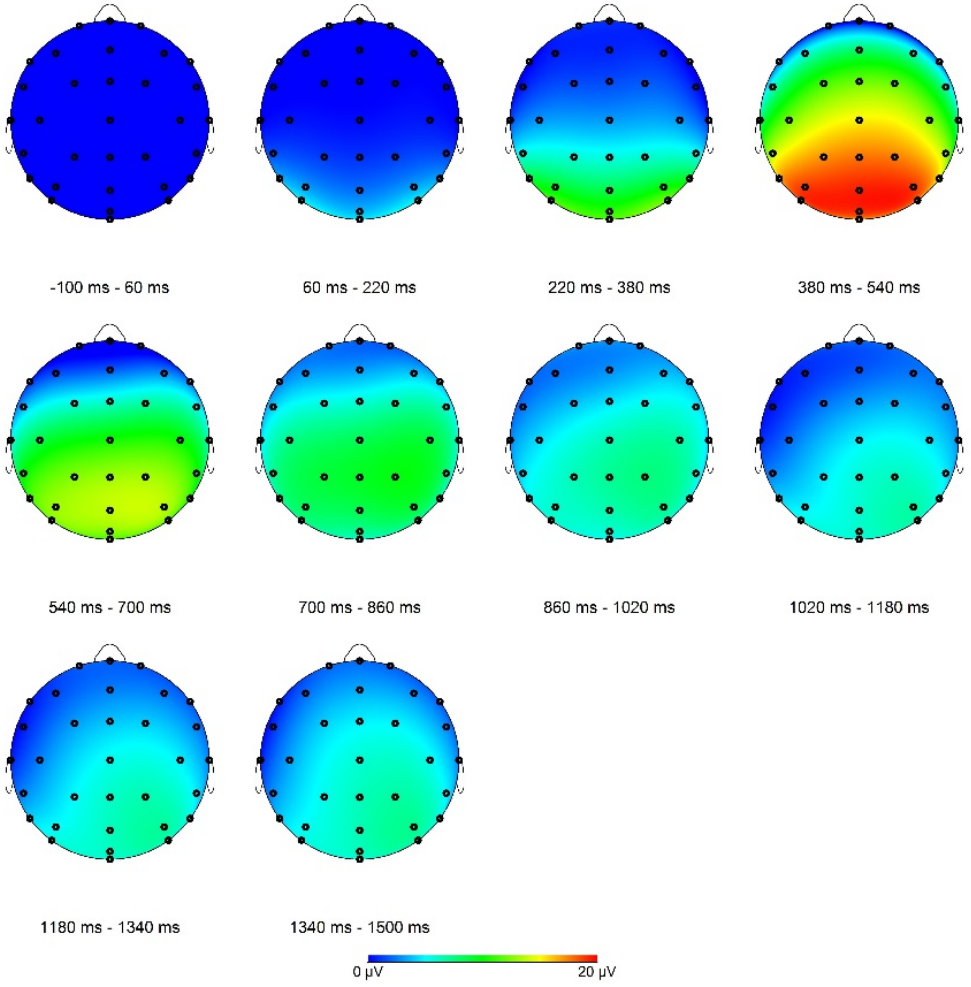

Figure 9. Scalp map of average response of truth-tellers to the filler faces across the ERP epoch.

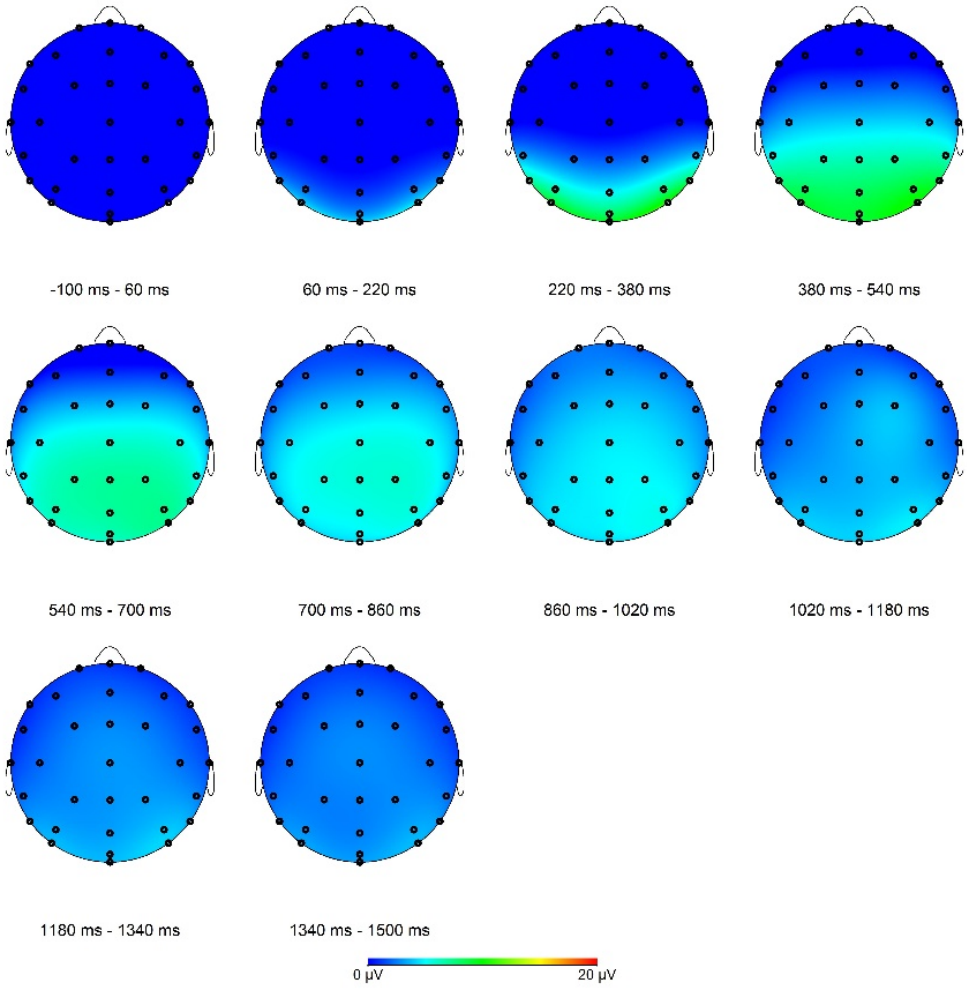

Figure 10. Scalp map of average response of liars to the face of the criminal across the ERP epoch.

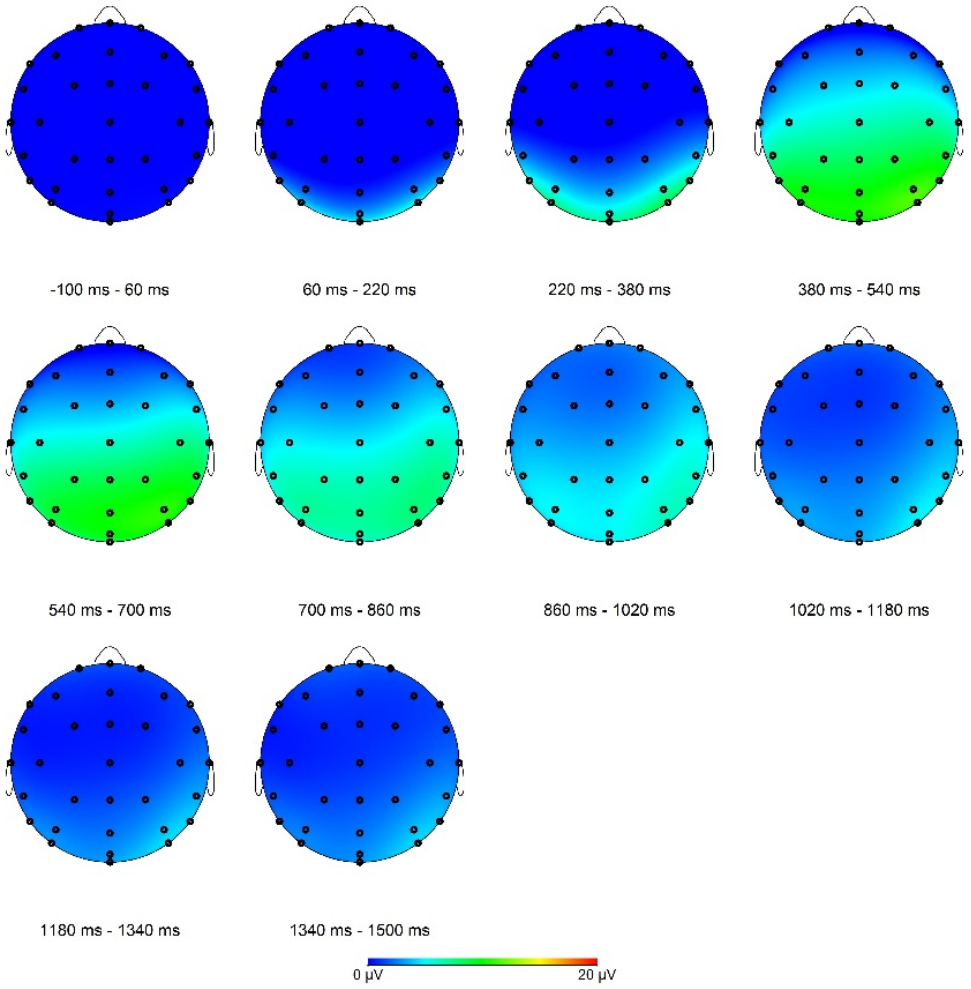

Figure 11. Scalp map of average response of liars to the face of the known lineup member (Chris) across the ERP epoch.

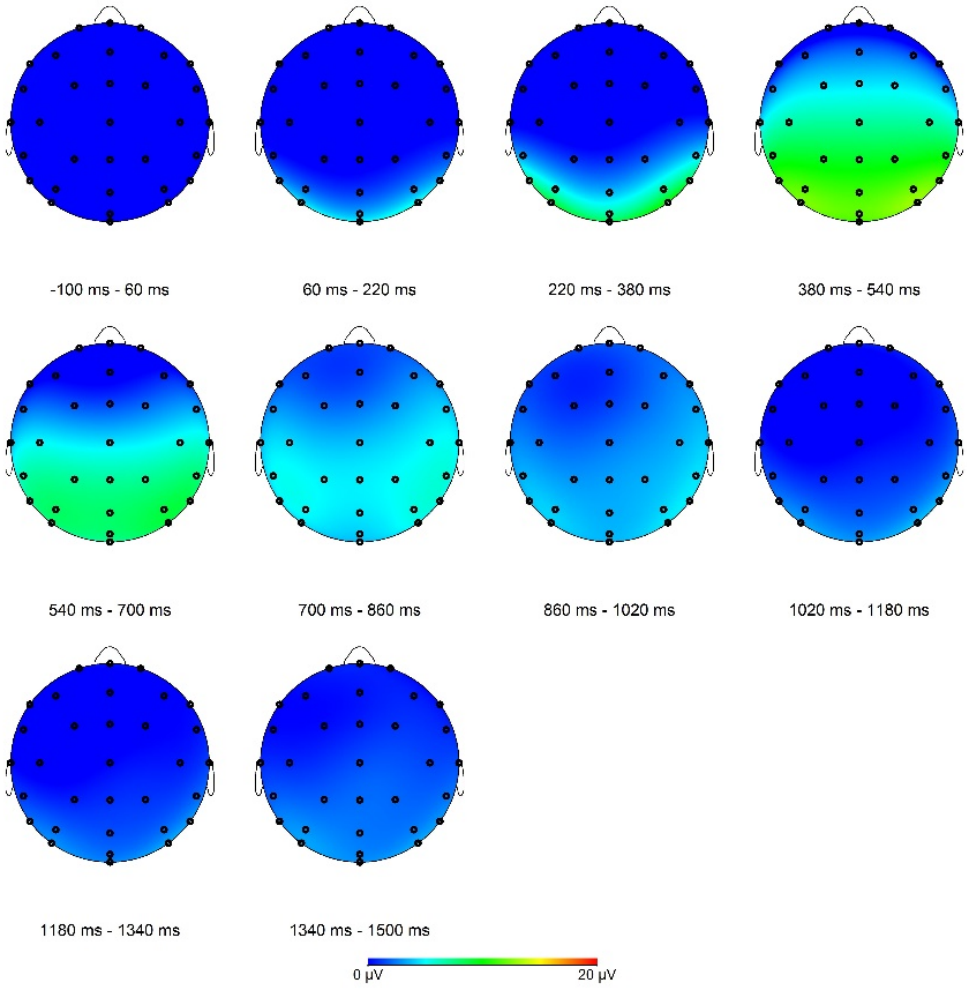

Figure 12. Scalp map of average response of liars to the filler faces across the ERP epoch.

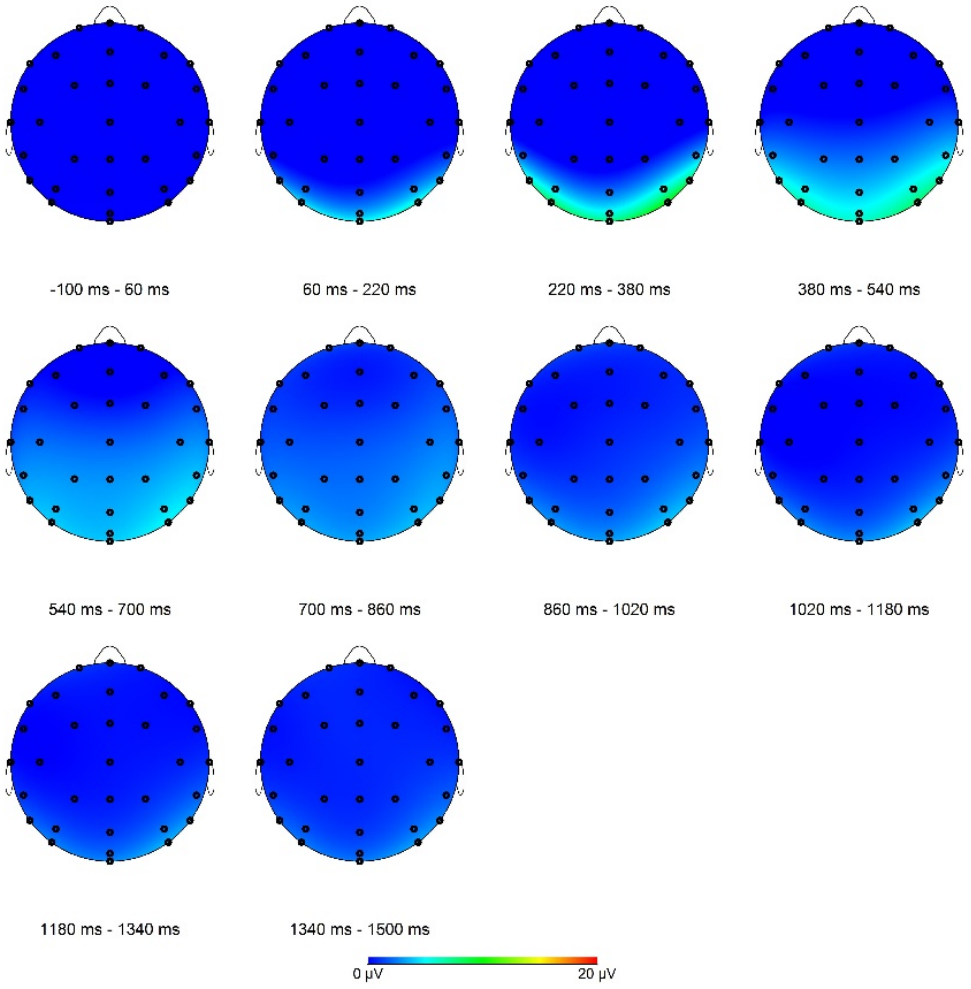

Table 1. Literature measuring correlations with Cambridge Face Memory Test

Paper	**Predictor**	**Outcome**	***r***	***N***	**CI lower**	**CI upper**
Bobak, Hancock, & Bate, 2016	Face matching HR	Cambridge Face Memory Test (CFMT)	0.61	27	0.29	0.8
	Face matching FAR	"	0.57	27	0.24	0.78
	Face memory TP trials	"	0.38	27	0	0.67
	Face memory TA trials	"	0.46	27	0.1	0.72
Bowels et al., 2009	CFPT	"	-0.61	124	0.24	0.8
McGugin et al., 2012	Holistic Processing Test	"	0.26	109	0.09	0.44
McKone et al., 2011	CFMT-Aus	"	0.61	74	0.44	0.74
[a]Spearman's rho calculated by authors, used here as well						
Note: Where not reported, 95% CI's calculated using vassarstats.net/rho.html						

Table 2. Literature measuring correlations with lineup accuracy.

Paper	Predictor	Outcome	*r*	*N*	CI lower	CI uppe
Andersen et al., 2014	CFMT	CP simultaneous lineup	0.26[a]	119	0.09	0.42
	"	CA simultaneous lineup	0.28[a]	119	0.1	0.44
	"	CP sequential lineup	*ns*	119		
	"	CA sequential lineup	0.27[a]	119	0.09	0.43
Bindemann et al., 2012	Hit rate, Bruce 1-in-10 as memory task	Groupwise probability of being a good witness (choosers)	0.7	37	0.49	0.83
	"	"	0.83	86	0.75	0.89
	FA rate, Bruce 1-in-10 as memory task	Groupwise probability of being a good witness (nonchoosers)	0.49	43	0.22	0.69
	"	"	0.38	99	0.2	0.54
Deffenbacher et al., 1978	Y/N Face Recognition score	4-person simultaneous lineup of class exam admins	-0.28	45	-0.53	0.01
Hosch, 1994	Benton Facial Recognition Task	Single lineup of experimenter (half CP)	0.54	32	0.24	0.75
	"	"	0.39	38	0.08	0.63
	"	"	0.41	27	0.04	0.68
	Y/N Face Recognition Sensitivity	"	-0.07	33	-0.4	0.28
	"	"	-0.21	36[b]	-0.5	0.13
	Y/N Face Recognition Response Bias	"	0.5	33	0.19	0.72
	"	"	0.28	36[b]	-0.05	0.56
Kantner & Lindsay, 2014	"	1 CP, 4 CA lineups	0.29	65	0.06	0.5
[a]Chi-squared values converted to correlation coefficients at campbellcollaboration.org/escalc/html/EffectSizeCalculator-R5.php						
[b]Sample sizes not reported, but are inferred based on reported *p*-values						
Note: Where not reported, 95% CI's calculated using vassarstats.net/rho.html						

Table 3. Face pairs removed for Study 3 with reason based on item analysis.

Left	Right	Reason
TSFWmale61-2happy.jpg	TSFWmale67happy.jpg	high c (conservative)
EMWfemale21happy.jpg	WWfemale29happy.jpg	high d'
EMBfemale21-2happy.jpg	EMBfemale22-2happy.jpg	high d'
TSFBfemale70-2happy.jpg	TSFBfemale87happy.jpg	high d'
TSFBfemale65happy.jpg	TSFBfemale84-2happy.jpg	high d'
EMWmale23-3happy.jpg	JWmale26happy.jpg	high d'
TSFWfemale67happy.jpg	JWfemale78(2)happy.jpg	high d'
EMBfemale20happy.jpg	EMBfemale22-4happy.jpg	high d'
EMBfemale21happy.jpg	EMBfemale22-3happy.jpg	high d'
EMImale28happy.jpg	TSFBmale50happy.jpg	high HR, dissimilar pair
WImale23happy.jpg	WImale24-2happy.jpg	low c (liberal)
WBfemale28happy.jpg	WBfemale21happy.jpg	low c
TSFBfemale78happy.jpg	TSFBfemale84-3happy.jpg	low c
TSFWfemale82-2happy.jpg	TSFWfemale64happy.jpg	low c
WWfemale22-3happy.jpg	WWfemale22-4happy.jpg	low c
TSFWfemale66-2happy.jpg	TSFWfemale69-2happy.jpg	low c, low d'
JWfemale63happy.jpg	JWfemale65-2happy.jpg	low d'
WWmale21-3happy.jpg	WWmale22-4happy.jpg	low d'
TSFWmale63-2happy.jpg	TSFWmale69happy.jpg	low d'
JWfemale71-2happy.jpg	JWfemale77happy.jpg	low d'

Table 4. Raw accuracy scores for Lineup Skills Test.

	LST Mean Accuracy (SD)			**Lineup Mean Accuracy (SD)**			
Study	**N/N**	**O/N**	***N***	**CA**	***N***	**CP**	***N***
1	0.53 (0.20)	0.57 (0.15)	*182*	0.41 (0.21)	*95*	0.44 (0.21)	*87*
2	0.51 (0.19)	0.57 (0.14)	*202*	0.40 (0.21)	*113*	0.49 (0.22)	*89*
3	0.59 (0.17)	0.43 (0.16)	*199*	0.39 (0.22)	*99*	0.42 (0.24)	*100*

Appendix B – LST Instructions

Study Phase:

We are attempting to create a Lineup Skill Test. Our test has two steps. First, we will present a long series of faces of people, one face at a time. Then you will take a test in which many pairs of faces will be presented (one pair at a time) and you will be asked to say which, if either, of the faces was on the study list. Please note that the faces used in this Lineup Skill Test are faces from a local community in the southern United States -- they have nothing to do with the faces you saw in the crime videos and are nobody you might have seen around Victoria.

Our hypothesis is that people who do well on our Lineup Skill Test (i.e., who pick the right face if one of the faces in a test pair had been studied, and who reject the pair if neither of the faces had been studied) also did well in the previous part of this study, in which the lineups for the crime videos were presented.

You will now see a series of faces one-at-a-time on the screen. Your task is to study them as they are presented and to try to memorize each. Your memory for the faces will be tested later.

Each face will be presented only once, and will only be on the screen for a short time. Therefore, please do your best to stay focused on the faces as they appear. As soon as you hit the space bar, the slide-show will begin.

IN STUDY 3:
You will now see a series of faces one-at-a-time on the screen. As they appear, you will designate each as older or younger than 30 YEARS OLD using the 'o' and 'y' keys.

Each face will only be presented once, and will only be on the screen for two seconds. After the face disappears, you will decide whether the person was older or younger than 30 before moving on to the next face.

Please do your best to categorize the faces accurately.

Test Phase:

You are now ready for the test phase of our Lineup Skills Test.

You will see a series of pairs of faces presented side-by-side. Some of the pairs will contain one face you studied earlier along with a new face, and other pairs will contain two new faces. As each pair of faces is presented, your task is to indicate which member of the pair you studied earlier (if either) -- the one on the left or the one on the right. You will make this decision and indicate your response by pressing the "a" key for the face on

the left and the "d" key for the face on the right. If you believe you are seeing a trial with two new faces, you should select the "s" key.

((Remember, like with real lineups, it is important for these face pairs that you pick the right face if one of the faces in a test pair had been studied and reject the pair if neither of the faces had been studied. Think of these face pairs like mini-lineups.))

After you enter your response, you will be asked to rank your confidence in the accuracy of that response. You will rank your confidence on a scale of 5 to 0, with 5 reflecting random guessing, or 50%. Then 6 will correspond to 60% and so on, with 0 standing for 100%.

Don't worry if you find yourself having trouble with the task; it is designed to be difficult.

Note that the pictures of the faces that you see will not be identical to those you saw earlier. Whereas the faces on the study list had neutral expressions, those on the test list are smiling. Your task is to determine which person you recognize in each pair even though old faces will have a different expression than before.

No face will appear twice within the test.

Ask now if you have any questions.

Note: ((Added for Study 3))