An Independent Validation And Extended Examination Of The Complex Trial Protocol, A P300-Based Concealed Information Test

Michel Funicelli

A Thesis
In the Department
of
Psychology

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Psychology) at
Concordia University
Montreal, Quebec, Canada

January 2020

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Michel Funicelli

Entitled: An independent validation and extended examination of the Complex Trial Protocol, a P300-based Concealed lnformaton Test

and submitted in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy (Psychology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Xavier Ottenwaelder

_____ External Examiner
Dr. Ewout Meijer

_____ External to Program
Dr. Vivek Venkatesh

_____ Examiner
Dr. Andrew Chapman

_____ Examiner
Dr. Aaron Johnson

_____ Thesis Supervisor
Dr. Jean-Roch Laurence

Approved by _____
Dr. Andrew Chapman, Graduate Program Director

March 18, 2020 _____
Dr. André Roy, Dean
Faculty of Arts and Science

ABSTRACT

An independent validation and extended examination of the Complex Trial Protocol, a P300-based Concealed Information Test

M. Michel Funicelli
Concordia University, 2019

The P300-based Concealed Information Test (CIT) is a memory detection technique where an examiner presents an examinee a crime relevant piece of information called '*probe*' along with an assortment of neutral alternatives called '*irrelevants*' while measuring P300 amplitudes. The Complex Trial Protocol (CTP) is a four-stimuli method of presenting probes and irrelevants while maintaining the examinee's attention on the computer screen with the additional presentation of '*target*' and '*Non-target*' items. The CTP has only been validated once by a team of independent researchers. In addition, CIT examiners are often faced with examinees intent on using countermeasures (CM) meant to influence the test outcome in their favour. We presented a literature review in chapter 1 in relation to CIT & CTP related research. In chapter 2 our aim was twofold. First, to validate the CTP's performance with autobiographical data, and then to test it with a novel mental (counting backwards) CM. We met both objectives by obtaining excellent detection results in terms of sensitivity (98%) and specificity (100%). The goal in chapter 3 was to verify if deep and shallow levels of processing (LOP) in a dual modality (words versus pictures) presentation had any influence on CTP results. Our findings pointed to the absence of an LOP effect on either verbal or visual stimuli, but we identified a pictorial superiority effect over words. Notwithstanding, these conclusions of our true positive (sensitivity) detection rates were unacceptably low, ranging from 7% (words) to 60% (pictures), while true negative (specificity) levels were excellent at 100%. Finally, in chapter 4 we exposed our participants to a mock terrorism scenario as we sought to explore the usage of three successive pictorial blocks of stimuli (mock bomb, mock crime scene, and male accomplice face) and to replicate the work of Hu, Bergström, Bodenhausen, and Rosenfeld (2015) and Ward and Rosenfeld (2017) on a memory inhibition CM. Depending on the stimuli type, our results showed diagnostic rates ranging from 64% to 79% in the innocent condition (true negatives), to 54% to 77% in the

simply guilty group (true positives), and 71% to 93% in the guilty CM group (true positives). Experiment 4 highlights the importance of the careful selection of probe type for an effective CIT. Relative to participants in the two guilty groups, innocent persons unexpectedly reacted to the mock bomb, whereas the facial probe of a male accomplice produced better distinguishing results. Lastly, we successfully replicated findings on the memory inhibition CM.

ACKNOWLEGMENTS

The road to here has been very long and not without its many personal and professional challenges. I owe a tremendous amount of gratitude to my supervisor Dr. Jean-Roch Laurence. Beginning with the day I walked into his office in 2005 with plans to pursue graduate studies in forensic psychology, Dr. Laurence has been there along my side to guide me through a master's degree and doctoral studies. His words, upon my completion of the MA program, that the "real school start at the doctorate level" still resonate in my head. Now I know why. Thank you, 'JR', as he is affectionately known in the lab, for being patient with me, for allowing me to pursue my passionate projects at my own pace, for pushing the limits of my intellect, and for giving me the opportunity to taste the fantastic world of teaching, in forensic psychology no less. My thanks also go to Dr. Peter Rosenfeld for permitting me to work in his laboratory at Northwestern University and to discover the P300 and the Complex Trial Protocol.

The pathway to the end of the PhD program would have been impossible without the insights, cooperation, humour, time, effort, and friendship of my research assistants Simon, Yasmine, Sarah, Lauren, and Sabina. You helped me develop into the junior researcher that I am, and I will be forever grateful.

Being a part-time graduate student and a full-time member of the Royal Canadian Mounted Police was not easy. Just as some supervisors and managers were putting up roadblocks across my academic pursuit, other work colleagues stepped up to the plate and made the journey possible. Pierre and Patrick, I thank you for seeing in my academic advancement what others could not or would not see and allowing me to work with a flexible schedule. Julie, I thank you for being there for me during some difficult episodes at work while I encountered harassment and discrimination from supervisors.

My dear Priscilla. The love of my life. I cannot ever thank you enough for being my support system, for the sacrifices you made while I spent time away studying, and most importantly for being my guiding light. The seas were rough at times, but your shining light was always there for me to find my way home safely. You are my rock. I love you.

To my mother and father, I am eternally indebted for raising me with the set of values where you finish what you start, setting me off on an academic path I never thought I would end up taking. Your endless and unconditional love were nothing but inspirational.

IN MEMORIAM

Luc Funicelli

1990-2019

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CONTRIBUTION OF AUTHORS

This dissertation consists of an introduction, four experiments and a concluding discussion. The dissertation was composed by myself and was edited by my supervisor Dr. Laurence and committee members Dr. Chapman and Dr. Johnson.

All four experiments were designed and conceptualized by myself with guidance from Dr. Laurence. I was responsible for the recruiting, all testing sessions, and data analysis.

The following individuals assisted me with the testing of participants and managing the schedule: Lauren White and Sabina Ungureanu for experiment 1 (Chapter 2), Sarah Salphati and Sabina Ungureanu for experiments 2 and 3 (Chapter 3) and 4 (Chapter 4).

CHAPTER 1: GENERAL INTRODUCTION

A certain fascination exists about detecting deceitful behaviors in others (Lee et al, 2002), and this obsession harkens back for millennia (Vrij & Verschuere, 2013). "For as long as human beings have deceived one another, people have tried to develop techniques for detecting deception and finding truth" (National Research Council, 2003, p. 1). Unsurprisingly, the scientific community is equally intrigued by both behaviors, the deception and its detection. In fact, as of a little more than a decade ago, more than 150 articles on deception and lie detection were published annually in peer-reviewed journals (Vrij, 2008).

Deceiving others or being the subject of deception is a common occurrence (DePaulo et al, 2003; Vrij, 2008; Porter & ten Brinke, 2010). How often people lie in day-to-day activities is a difficult question to answer. Researchers employ a few methods such as self-report, diary studies, and get-acquainted conversations in laboratory settings, but it is nearly impossible for investigators to ascertain the truth about the answers that participants provide (Vrij, 2008). Nevertheless, findings generally arrive at the conclusion that lying happens frequently in human transactions (see Vrij, 2008, for a complete review). Not only is lying a constant event, but DePaulo, Kashy, Kirkendol, Wyer, and Epstein (1996) demonstrated that it occurs daily. According to their diary studies, community members told one lie a day, while college students told two. During one-on-one interaction over the course of one week, community participants lied to 30% of their interlocutors, and in one out of every five social interactions. College students were untruthful to 38% of the people in their lives, and in one out of three interactions. Around 50% of the lies in that study were for self-serving reasons (i.e. protecting one's own interest), and participants admitted that the overwhelming majority of their lies were not serious. In a follow up study, DePaulo, Ansfield, Kirkendol, and Boden (2004) found that when the stakes were higher (e.g. lying about an illegal or immoral act), 90% of the lies were for self-serving reasons. Consequently, it is a logical inference that actors of the judicial system (i.e. police, probation, parole, lawyers, judges, jurors, etc.) are routinely confronted with acts of deception during the course of their respective functions and professional duties.

Some lies are considered benign while others are a lot more serious. There is a world of a difference between complimenting a host for their excellent cooking when it was not, or telling someone you liked the gift just handed to you when you disliked it, and falsely claiming your innocence about a crime the police alleged you have committed, or lying about your involvement

in an act of violent extremism or membership in a terrorist organization. Somewhere in between are also inaccurate truths where individuals erroneously, but firmly believe, to have experienced an event that either did not take place at all, or some parts of it were mistakenly committed to memory.

Police detectives, national security agents, and other investigative agencies often have the difficult task of determining truths from falsehoods in many of their investigations. Suspects of criminal or terrorist activity, witnesses and victims may have good natural reasons to lie (e.g. avoiding prosecution, incarceration or deportation, protecting a loved one). Others might be deceitful as a result of a pathological or a psychological condition. Birch, Kelln, and Aquino (2006) reported the Canadian case of a 22-year-old female assessed with pseudologia fantastica. She convincingly made false accusations against a number of people who ended up criminally charged, and also resulted in the criminal conviction and yearlong imprisonment of an innocent friend. Dike, Baranoski, and Griffith (2005) described the case of Patrick Couwenberg, also diagnosed as a pseudologue, who misrepresented himself to become a superior court judge in Los Angeles, pretended he was a CIA operative, lied to lawyers, other judges and state officials before the California State Commission on Judicial Performance ousted him from office in 2001. In the early 1980s, while serving a sentence for murder, Henry Lee Lucas managed to persuade police authorities in Texas over an 18-month period that he had committed over 600 murders across the USA before they realized they were being duped (Gudjonsson, 1999). When interviewed in prison by Gudjonsson (1999), Lucas estimated that he had made over 3,000 false confessions to murder. The consequences of undetected deception in legal, forensic or security settings can be astronomical. Aside from the obvious repercussions for the wrongly convicted innocent, the integrity of an entire justice system is seriously jeopardized, the reliability of law enforcement agencies is called into question, valuable tax dollars, time, and human resources are wasted, and worse, in some cases a criminal suspect runs free.

Discovering the truth in criminal investigations, and alternatively detecting deception, is without question paramount. Facts dictate the course of a police investigation, and discretionary powers are devolved onto the police to take the case further or not. In Canada, when the police determine that laying a charge is warranted, the fact pattern as investigated by the police is then submitted to a public prosecutor for an assessment of the evidence and to approve or not the laying of a charge. These facts, first gathered by investigators, then tendered as evidence through

sworn testimonies before a judge, ultimately will form the basis of future judicial decisions. Throughout this process, every actor of the justice system will have to sift through the lies, mistaken truths, and real truths from the ground truth.[1]

An in-depth look at the motivation behind dishonesty and the typology of lies (e.g. exaggerations, minimizations, lies of omissions, etc.) is beyond the scope of this paper. Nevertheless, the way we define deception is important in discussions about this behavior. Researchers have come up with several definitions over the years (see Masip, Garrido, & Herrero, 2004a for a full review). But the common denominator that defines deception, and on which most psychologists agree on, is that the sender's piece of information contains at least one of the three "components of the prototypical lie: The objective falsity of the proposition, the sender's belief in this falsity and the intention of the sender to deceive the receiver." (Masip et al., 2004a, p. 165). We will retain one definition for the purpose of this essay, "a successful or unsuccessful deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue" (Vrij, 2008, p. 15). This definition is broad and versatile enough to encompass liars who fail in a deliberate attempt at lying to their intended interlocutor, as well as to capture individuals, such as magicians, who tip-off in advance, by the nature of their theatrical performance, their intentions to deceive their audience, and properly excludes individuals who do not categorize an honest mistake as a lie, as shown by the results of a survey reported in Vrij (2008), where 88% of children and 95% of adults felt that lying necessitates an intentional act. Yet, it is also important to discern false beliefs and misremembering from lying. If two eyewitnesses to the same event recall it differently, and firmly believe in their memory, it does not necessarily mean that one of them is deceitful and the other truthful. In fact, both could be wrong, or both could equally be right.

Lie catchers have made great strides since early societies. Trials by ordeal, where some kind of higher power was thought to intervene in favor of truthful persons, or other creative torturous procedures, were some attempts at lie-detection in Greece, pre-Christian Scandinavia, Iceland, Polynesia, Japan, and Africa (Palmer, 1989; Grubin & Madsen, 2005). "In the Middle Ages, the honest man in some parts of Europe was expected to be able to hold his arm in boiling water longer than a liar, while in Scandinavia, a woman accused of adultery was required to

[1] Ground truth is defined here as if the events had been recorded by audio-visual means from every possible angle and made available to decision makers.

'clear herself with the iron,' that is, hold a red-hot iron for a short time: if her hands burnt she was guilty of adultery" (Grubin & Madsen, 2005, p. 358). In China, suspects were asked to chew rice powder and then spit it out. If the powder was dry the person was guilty (Grubin & Madsen, 2005). Another variation, during the Inquisition, implicated swallowing a slice of bread and cheese. If the slice stuck to the suspect's palate the person was declared a liar (Grubin & Madsen, 2005). Despite the crudity of these techniques, an association had been made between decreased saliva flow and deception. Centuries later, Hugo Munsterberg, at Harvard university, pioneered the idea of physiological indicators to deception, from which evolved the invention of the polygraph (Grubin & Madsen, 2005).[2]

Currently, psychological science uses three approaches to make credibility assessments (Masip, Garrido, & Herrero, 2004b). First, behavioral indicator practices (i.e. Philippe Turchet's synergology, Behavior Analysis Interview, Neuro-Linguistic Programming, thermal facial imaging, and voice stress analysis) attempt to decipher non-verbal manifestations through the observation of a person's behavior for signs of deceit (Turchet, 2004; Vrij, Mann, & Fisher, 2006; Vrij, 2008; Mann et al., 2012). Second, verbal methods involve the examination of the written content of a person's declaration for cues of deception, through a variety of statement analysis techniques (i.e. Scientific Validity Analysis – SVA, Criteria Based Content Analysis – CBCA, Reality Monitoring – RM, Scientific Content Analysis - SCAN) (Vrij, 2015). Third, psychophysiological techniques record changes in electrodermal activity, blood pressure and respiration (e.g. polygraph) (National Research Council, 2003) while others measure hemodynamic and electrical activity in a variety of deception-based paradigms. Respectively, this last category includes fMRI, which localizes the blood flow and oxygen consumption in brain structures during deception (Ganis, 2018), and EEG which displays electroencephalographic signals in relation to a particular Event Related Potential (ERP) component called P300 associated with memory recognition (Rosenfeld, 2011).

If progress over centuries, in identifying liars from truth tellers, is calculated in terms of humaneness, then it is nothing short of remarkable. The methods designed thus far, are much more civilized than the torturous ones of centuries ago. However, the precision of human

---

[2] In 1915, William Marston, a student of Hugo Munsterberg, inspired by his wife's observation that her blood pressure increased when she got angry or excited, developed a systolic blood pressure deception test, and found high correlates between lying and changes in systolic rates (Grubin & Madsen, 2005).

judgments at identifying deception without technical aid is far from perfect. In a review of 107 studies published in English where lay people and trained investigators were asked to detect truth tellers and liars based solely on verbal and non-verbal behavior of adults, the success rate hovered marginally above chance level at 54.27% for laypersons and 55.91% for professional lie catchers (Vrij, 2008). When chance level is established at 50%, lie detection without the assistance of some device is not much better than the flip of coin.

Detection rates improve when some kind of instrumentation is used to detect mendacious statements, but methodological and foundational issues then become critical considerations. The polygraph is an instrument that measures physiological signals, generated by the autonomic nervous system (ANS), such as respiration, heart rate, blood pressure, and electrodermal activity. Several polygraphic tests can be conducted while an examiner questions an individual as those signals are collected. Among the more popular techniques are the Control Question Test (CQT) and the Concealed Information Test (CIT) (National Research Council, 2003).[3] The CQT is by far the most favored of the two amongst law enforcement practitioners in the USA (Maschke & Scalabrini, 2005) and in Canada.[4] The general assumption underlying the CQT is that relevant questions will elicit stronger arousal, and thus larger responses, only in guilty persons (Meijer & Verschuere, 2018). Despite considered dubitable by many scientists, polygraph proponents claim that the CQT has a 90% or higher accuracy rate (Iacono & Ben-Shakhar, 2019). Albeit the favorite technique, the CQT is very controversial because of its unknown error rate and questionable methodological premises (Vrij, 2008). Iacono and Ben-Shakhar (2018) reviewed the literature since the landmark report by the National Research Council of the National Academies of Sciences in 2003. According to them, the main conclusions reached by the panel more than a decade ago, that the CQT is based on weak scientific foundation with indeterminate accuracy, still stand. Despite these identified deficiencies, the CQT's appeal over the CIT is its polyvalence, in that the CQT can be applied to a wide array of criminal infractions and in circumstances where little or no tangible evidence has been uncovered by police (Podlesny, 1993; Podlesny, 2003). As described in more details further below, the CIT requires from

---

[3] Other less popular techniques are the Reid Comparison Question Test, Zone Comparison Test, Utah Probable-Lie Test, Utah Directed-Lie Test, Test of Espionage and Sabotage, and Stimulation Test (National Research Council, 2003).

[4] There is only one polygraph school in Canada. It is located at the Canadian Police College in Ottawa, Ontario, and the course is managed and dispensed by the Royal Canadian Mounted Police.

investigators to present criminally relevant pieces of information (also called undisclosed case details) known only to them and the examinee. But police are often challenged by the absence of physical traces of evidence when investigating a criminal offence. Indeed, Baldwin and McConville (1980) found that forensic evidence was either unavailable or not important in 95% of criminal cases in England, and Horvath and Meesig (1996) concluded that forensic clues were gathered in only 10% of offences investigated by police in the United States. The difficulty for police in seeking tangible evidence where none exists or is impossible to find, transfers the investigatory focus onto possible suspects where the CQT becomes the go-to technique to extract incriminating evidence in the form of admissions of facts or culpable confessions.

With respect to brain imagery, two types of paradigms dominate deception research. The differentiation of deception (DoD) paradigm requires participants to respond truthfully to a question and deceptively to another question of a matched pair (Furedy, Davis, & Gurevich, 1988). Thus, comparing counterbalanced honest and deceptive conditions among groups of participants provide investigators with indications of which neural processes are engaged in deceptive responses versus honest responses (Ganis, 2018). The CIT is based on the idea that a bodily function will manifestly respond to an infrequently presented item of interest (i.e. a crime relevant piece of information) only known by the author of the crime, called 'probe', compared to frequently shown neutral items of equal plausibility, called 'irrelevants' (Lykken, 1998; Rosenfeld, 2011). For example, in a homicide investigation where the murder weapon, say a revolver, is recovered at the crime scene, the latter becomes the probe and irrelevants could be items such as a baseball bat, a knife, a rifle, etc. However, neither of these two paradigms are well suited for neuroimaging deception research. First, matching questions over significance and familiarity in lying and truth telling conditions are usually problematic in DoD (Ganis, 2018). Moreover, the DoD paradigm was not designed to study detection but rather the psychological and physiological underpinnings of deception as a behavior (Suchotzki, Berlijn, Donath, & Gamer, 2018)[5]. In the CIT, probes and irrelevants are inherently different from each other, the former being more salient than the latter by their infrequent presentation. Second, participants are often instructed to lie in these research paradigms, whether in mock crime scenarios or not, which makes it difficult to evaluate the processes involved in deception (Ganis, 2018).

[5] The Sheffield Lie Test (SLT) is a variant of DoD paradigm. In the SLT the participant is asked to answer the same question twice, once for each truthful and deceptive condition (Suchotzki, Berlijn, Donath, & Gamer, 2018).

Nevertheless, depending on the paradigm used, accuracy rates with neuroimaging methods show good promise. Ganis (2015) reviewed 10 neuroimaging studies and found the average accuracy rate to be about 82%. Kozel et al. (2009) used a variant of the DoD paradigm in a mock theft scenario and achieved an impressive 100% (9/9) sensitivity rate, but 33% (5/15) specificity rate. Peth et al. (2015) employed a CIT paradigm in a mock theft scenario as well, but they also manipulated encoding levels by comparing the planning of the theft (guilty intention group) versus enacting the theft (guilty action group). For example, a CIT question was: How much money did you steal from the locker? followed with five alternative answers (e.g. 100€?, 10€?, 5€?, 50€?, 20€?). Relative to innocent controls, their signal detection theory derived data showed an Area Under the Curve (AUC) of .91 for the guilty intention group and .98 for the guilty action group. AUC varies from 0.0 to 1.0, with .5 indicating chance level and 1.0 signifying perfect detection. Dealing specifically with CIT findings, Meijer, Verschuere, Gamer, Merckelbach, and Ben-Shakhar (2016) found that the average of AUC detection scores from four brain imaging studies was 0.94.

Reaching perfect accuracy in brain imagery is, however, constrained by five limitations (Ganis, 2018). First, neuroimaging has not been able to distinguish neural processes involved between individuals actually committing a crime (i.e. guilty knowledge) and those having acquired crime related knowledge through other means (i.e. incidental acquisition of crime knowledge) such as being a crime witness or due to media leakage. Second, fMRI-based detection is vulnerable to countermeasures (CM).[6] For instance, in Ganis, Rosenfeld, Meixner, Kievit, and Schendan, (2011) 26 participants were asked to respond to six dates, truthfully to five non-significant dates, and deceptively to their respective date of births. Investigators achieved a 100% detection rate in individual events when no CM were used, but the accuracy rate fell to 33% when CMs (i.e. imperceptibly moving the index finger, middle finger or left toe) were applied. Third, brain imaging methods cannot differentiate between the truth from what one believes to be the truth, "the pattern of neural activity elicited by old items correctly believed to be old (hits) seems to be indistinguishable from that of new items incorrectly believed to be old (false alarms)" (Ganis, 2018, p. 161). Fourth, specificity is poor in neuroimaging studies as

[6] Countermeasures are either physical (e.g. biting tongue, moving limbs or digits) or mental (e.g. doing silent arithmetic operations, thinking about a salient stimuli) actions designed to influence the outcome of a test, without the examiner's knowledge, so as to enhance the saliency of neutral items and produce a false negative reading, or inhibit the reaction to a probe item and generate a false positive result.

neural patterns believed to be associated to deception (e.g. prefrontal regions) are also engaged in many other general-purpose cognitive control processes that do not involve deception. Finally, when the cost of testing equipment is considered, the accuracy of brain imagery in deception is not much different from that of much less expensive psychophysiological methods, such as EEG, which would likely make the latter more attractive and available to front line investigative agencies.

P300-based CIT

Another technique is attracting the attention of psychologists in their pursuit of deception detection. An ERP component called the P300, is a special series of electrical brain waves peaking positively generally about 300ms (Luck, 2014), 250-900ms (Andreassi, 2007), or somewhere between 500-800ms (Farwell, 2012), "whenever a *meaningful* piece of information is *rarely* presented among a random series of more frequently presented, non-meaningful stimuli of the same category as the meaningful stimulus." (Rosenfeld, 2011, p. 64). This waveform does not represent a lie per se but is rather considered a memory recognition phenomenon (Rosenfeld, 2011).

The leading theoretical framework that underpins the P300 waveform is the orienting reflex (OR) (Gamer, 2011). The OR was first uttered by Pavlov in 1910 at a lecture in St-Petersburg (Zernicki, 1987), as a behavioral response to unexpected novel stimuli in the environment drawing the organism's attention to a particular stimulus in order to extract more information from it. Neurologically speaking, the OR is a component of a complex functional system that involves the integration of different parts of the brain, having its genesis "in a mismatch of extrapolatory impulses and afferent signals reaching common efferent neurons." (Sokolov, 1963, p. 576). From an evolutionary perspective, the OR was seen as a means to survival (Bradley, 2009). An organism's reflexive signal requiring it first, to direct its attention towards an unusual and significant incoming stimulus from its surroundings, then to assess its hazardous or safety value, followed by the mobilization of the necessary mental or physical resources, and finally to adapt its behavior to the changing environment. Pavlov called this response the "what is it? reflex" or "orienting reflex" (Zernicki, 1987, p. 240).

> It is this reflex which brings about the immediate response in man and animals to the slightest changes in the world around them, so that they immediately orientate their

> appropriate receptor organ in accordance with the perceptible quality in the agent bringing about the change, making full investigation of it (Pavlov, 1927, cited in Lynn, 1966, p. 1).

The Complex Trial Protocol

In 2008, Rosenfeld et al. developed a four-stimuli protocol and named it the Complex Trial Protocol (CTP). The CTP is a P300-based CIT technique of memory detection. Since then Rosenfeld and colleagues have conducted a wide range of laboratory studies to assess the CTP's performance. What follows is a review of the pertinent experiments.

Too begin with, the CTP is made up of two parts. In part one, a participant is shown, following a baseline of 100ms of recorded pre-stimulus brain activity, either a rare ($p = 0.2$) probe or a frequent ($p = 0.8$) irrelevant stimulus (S1) and is instructed to respond (R1) as rapidly as possible with a single button press from one mouse, regardless if it is a probe or an irrelevant item (Rosenfeld et al., 2008; Rosenfeld, 2011). This is characterized as the "I saw it" implicit response (Rosenfeld et al., 2008). This response is unconditional and must be executed no matter what the stimuli is, a probe or an irrelevant item. In this part the probe stimulus is a crime relevant item known only to the author of a crime and the authorities, and irrelevant items are an assortment of similar stimuli acting as neutral alternatives to the probe item (Rosenfeld, 2011). After a pause, where a fixation cross appears in the middle of the computer monitor, part two of the CTP comes into play, and a second stimulus (S2) is displayed. The participant is asked to respond conditionally (R2) and rapidly on either one of two buttons from a separate mouse. If the stimulus is a target item (a string of numbers, typically 11111), the participant is instructed to press the right-hand button. If the stimulus is a non-target item (a series of strings of numbers, typically from 22222, … to 55555), the participant is instructed to press the left-hand button (Rosenfeld et al., 2008). This response is typified as explicit because of the decision and task the participant must execute after making the target/non-target differentiation. The inter-stimulus interval (ISI) can be randomized and vary from 1,100 to 1,800ms or presented at a fixed rate. According to the inventors, the S1/R1 task permits investigators to compare probe to irrelevants P300 amplitudes, while the S2/R2 task ensures the participant's attention is kept on the monitor. The S1/R1 and S2/R2 tasks run consecutively for as long as investigators require to meet their experimental purposes (e.g. generally 350-375 stimuli presentations overall). Finally, the CTP investigator usually interrupts the experiment periodically to quiz the participant about the

identity of the last stimulus seen. This step further ensures that the participant is fully attentive, cooperative and not employing any CMs (Rosenfeld et al., 2008).

Three methods of comparison are employed in CTP research to measure probe-irrelevants differences, Iall, Imax, and Blind Imax. Iall represents the average P300 amplitude of all irrelevant items from which the same data for the probe is compared against. This method is the more common. Imax is a more rigorous test. The average P300 amplitude for the probe item is compared with the irrelevant item having registered the largest mean amplitude level (Meixner & Rosenfeld, 2014). Finally, Blind Imax is conducted when the probe is unknown. For example, if a child has been kidnapped and sequestered in an unknown location, a suspect could be shown several possible locations suspected by the police to be hiding spots. The stimulus with the largest P300 amplitude average is assumed to be the probe, against which the second largest P300 is compared against (Meixner & Rosenfeld, 2011).

Another issue worth mentioning about CTP-based research are search windows. These windows can be set by the experimenter to allow an algorithm to search for the most positive and negative peaks. However, no CTP-based paper has ever explained thus far how those look windows have been determined. It is unknown what impact, if any, the manipulation of those search windows could have on CTP results. Albeit not a research aim of ours, we have taken an initiative in this regard and developed an objective method which is described in the method section of each subsequent chapters. But this is certainly an area of CTP research that warrants further attention in the future.

The most positive peak, following stimulus presentation, is typically where one finds the P300 waveform. This is well established. What follows this peak though is less clear. The most negative peak that follows can be described summarily as a rebound, or post-peak recovery, located towards the end of an ERP epoch (Rosenfeld, 2005; Hu, Bergstrom, Bodenhausen, & Rosenfeld, 2015). Depending whether positivity is plotted upwards or downwards, this negative rebound type of peak has been called NEG (Soskins, Rosenfeld, & Niendam, 2001), Late Positive Potential (LPP) (Luck, 2014), or Late Posterior Negativity (LPN), but it is generally found somewhere between the most-negative (if positivity is plotted downwards) 100ms segment from the P300 latency to 1500ms, the end of the ERP epoch (Hu, Bergstrom, Bodenhausen, & Rosenfeld, 2015). However, several mechanisms have been ascribed to this late component,

namely sustained attention, motivational responses, attentional resources, and stimulus valence (Benning et al., 2016). Research into the LPN is beyond the scope of this paper.

Two methods exist to measure P300 amplitude levels, base-to-peak (BP) and peak-to-peak (PP). Generally speaking, the BP method consists of measuring the amplitude levels between the pre-stimulus presentation baseline to the midpoint of the maximally positive segment. The PP is slightly different in that the algorithm "searches for the maximally negative 100ms segment between P300 latency and 1300ms and then subtracts the average absolute amplitude of that segment from that of the maximally positive segment described above." (Deng, Rosenfeld, Ward, & Labkovsky, 2016, p. 4). The preferred method is PP since it is reported to yield on average 20% superior detection relative to BP (Soskins et al., 2001).

CTP Laboratory Research to Date

To the best of our knowledge there is no known field experiment involving the CTP, and all but three studies, excluding those presented in this dissertation, were performed in Dr. Rosenfeld's laboratory at Northwestern University. The Hungarian-Norwegian team of Lukacs et al. (2016) have thus far conducted the only true independent validation investigation of the CTP. The other semi-independent research came from two Chinese-American teams of Deng, Rosenfeld, Ward, and Labkovsky (2016) and Lu et al. (2017). The very small number of genuinely independent examinations of the CTP alone demands greater scrutiny from external researchers.

The type of stimuli used to validate the CTP can be broken down into four categories, words (i.e., participant's surname or hometown, experimenter's first name, verbal depiction of an object), numbers (i.e., participant's date of birth, social security number, local telephone area code), and pictures (i.e., pictorial illustration of an object or a scene).

The experiments in connection with the CTP conducted thus far have either combined some of the stimuli described above with an additional experimental feature such as an implicit autobiographical association test (IAAT), a CM, some feedback mechanism, or simply tested a variety of different stimuli or modality (i.e. auditory). We limit the review here to studies relevant to the experiments that we conducted.

Verbal Stimuli Only

Words have been used the most often in CTP research with varying degrees of success and were displayed at study and at test in different forms of modality (i.e. verbally or auditorily). Rosenfeld et al. (2008) first tested the CTP with verbally presented autobiographical information (e.g., mother's first name, family surnames, and home towns) in different colors, with and without a CM, over the course of three weeks where each week coincided with a respective block of trials (studies that investigated CM are covered in a separate section below). No CM was applied in the first and third week, and the latter week also served as a replication study of week one. The original design of the CTP did not use numbers for target and nontarget stimuli as described earlier. Instead the verbal stimuli were presented in one of five colors (Green, Red, Blue, Yellow, and Purple). Green was the color assigned to target stimuli and the other colors defined the non-targets. Probe and irrelevants were presented in White (presumably on a dark colored background although unspecified in the method section). The authors stated that probe and irrelevant items were also presented in any of the five colors and could reoccur as targets and non-targets. There is no explanation offered for the incorporation of this Stroop-like feature and the stimuli portion of the method section is confusing. Nevertheless, Rosenfeld, Labkovsky, Winograd, Lui, Vandenboom, and Chedid (2008) reported near to perfect individual classification rates for both weeks that no CM were used. Hit rates of 92% (11/12) were obtained with one set of search windows of 500-800ms for the positive P300 peak, and 800-1300ms for the subsequent negative peak, and the rate improved to 100% (12/12) when the look windows were individually tailored at 500-700-1600ms. The authors did not provide any detail on how the search figures were determined to measure both peaks.

The team of Lukacs et al. (2016), performed a near replication of Rosenfeld et al. (2008) but used the latest, and more conventional form of the CTP, and a larger sample size (n = 66). Family names (in White font over a Black background) were used as probes in this investigation except for the control group where none of the surnames presented were relevant to the participants. The irrelevant items came from a list of 20 Hungarian family names and verified with each participant for non-pertinence or saliency. The search windows were 500-800-1300 ms but no description is offered as to how researchers arrived at those figures. The P300 probe

versus irrelevant difference was significant and near perfect hit rates were observed for the control group (13/14) and simply guilty participants (14/15).[7]

Autobiographical data is deeply memorized and decidedly not an appropriate standard for CIT field application. A concern often raised in CTP studies is the need for ecological validity (Rosenfeld, Hu, Labkovsky, Meixner, & Winograd, 2013; Lukacs et al., 2016). To date no field studies have been conducted with the CTP, but several looked at the effectiveness of the CTP in the context of a mock crime and verbal stimuli. Winograd and Rosenfeld (2011) explored whether the CTP was sensitive enough to detect information acquired only during the commission of a mock crime, in this case the theft of a ring. Participants were asked to attend a mailbox in the Psychology department, locate an envelope in Dr. Rosenfeld's mailbox, steal the item inside it, and return to the laboratory for further instructions. They were never told that the envelope contained a ring. Control subjects were asked to walk down to the office and come back without doing anything else. Only the episodic memory of committing the act was expected to be sufficient enough for encoding the event. The word "RING" was the probe and irrelevant items were assorted names of jewelry items (e.g. watch, earring, bracelet, etc.). A "standard look window" (p. 157) of 300-700-1500 ms was applied but neither is the term "standard" defined nor the computations to arrive at those figures described. The individual diagnostic rate for guilty subjects was 83% (10/12) and innocent controls was 92% (11/12).

Meixner and Rosenfeld (2011) employed a mock terrorism scenario in which subjects in the guilty group were given a briefing document describing how they were to conduct an attack on the United States in their role of a terrorist agent. Three blocks of verbal stimuli were used. One block consisted of types of bombs, another with names of cities, and a third one with possible months during which the attack was to occur. Participants were then asked to compose a letter to their superior in the terrorist organization in which they outlined the choices they had made from the list of bombs, cities, and dates. Participants in the control condition completed a similar task planning a vacation. The findings in Meixner and Rosenfeld (2011) were impressive with perfect accuracy in both guilty and innocent groups. However, a few details were not reported. As in Winograd and Rosenfeld (2011) the method used to identify the search windows

[7] Labelling a group as simply guilty (SG) in CIT experiments is often used even if mock crimes are not used as part of the paradigm. When no mock crime is used this label simply refers to an experimental condition where recognizing the probe is understood as a proxy to recognizing a crime relevant item.

was not described. Another issue was that the data from all three blocks were averaged into one. Consequently, it is impossible to know if one or more type of stimuli was recalled and detected better than others. No details are offered on the letter each participant, guilty or innocent, put together. It does not appear that they were coded for a semantic assessment of how each stimulus was incorporated in their document, although stimuli were verified that they had no personal relevance to participants. As noted in later discussion, levels of processing influence the nature in which details are encoded and later retrieved.

One important weakness of CIT is the amount of detail made available to the examinee, through information leakage (e.g. media coverage, inappropriate disclosure during an interview), prior to the test. As mentioned earlier, if a person encodes a crime detail through others means than having taken part in the actual crime, the likelihood of a false positive result is real. Winograd and Rosenfeld (2014) utilized the same scenario as in Winograd and Rosenfeld (2011) to test the CTP's sensitivity to such likely occurrence. Four experimental conditions were manipulated (the detection rates appear in brackets): a group of guilty-informed subjects were instructed to steal and did steal a "ring", (100% - 13/13); a group of guilty-naïve volunteers were instructed to steal and did steal an "item", (79% - 11/14); a group of innocent-informed participants were asked to steal a "ring" but did not take it, (31% - 4/13); and finally a group of innocent-naïve persons were told to steal an "item" but they did not take it, (86% - 12/14). Clearly, revealing crime details to examinees enhances sensitivity rates, such that 69% of innocent-informed participants were incorrectly classified as guilty (false positives). Of note here is that the confidence interval applied in this instance was 80% instead of the usual 90%. The detection rate might have been lower with the higher confidence level. A confidence level is applied during bootstrap analyses for individual classification. It indicates the certainty level that the number of each bootstrapped iteration, out of 100 for example, the probe peak-to-peak amplitude difference exceeds that of irrelevants. Nonetheless, the experiment demonstrates the vulnerability of the CTP in cases where information is improperly leaked to examinees ahead of time.

Meixner and Rosenfeld (2014) examined the performance of the CTP in relation to incidentally acquired memories. In this study, subjects wore a body camera and recorded their activities for several hours. A series of stimuli were extracted from those recordings and verbally shown in later testing with findings of perfect accuracy. However, as in Meixner and Rosenfeld

(2011), some reporting issues need to be highlighted. Whereas the confidence interval was .9 in Meixner and Rosenfeld (2011), it was not reported in the bootstrapped calculations in Meixner and Rosenfeld (2014). They claimed to have perfectly discriminated all 24 participants (12 in each group). A visual inspection of their individual subject's bootstrap data (at table 2, page 6) raises a doubt about the perfection claim. First, the data presented represents an average of all three blocks of stimuli, and if a .9 decision criterion was applied to both Iall and Imax tests (based on the number of iterations above 900 out of a possible maximum score of 1,000), the detection scores for participants in the knowledgeable group would have been 75% (9/12) and 100% (12/12) in the nonknowledgeable group. With the Imax method the detection rates would have been 25% (4/12) and 100% (12/12) respectively. Second, on closer scrutiny of the individual block data (at Fig. 7, page 8), a .9 confidence interval would place some individual data in false positive and false negative territory. For instance, where some individuals may have been ruled as knowledgeable on one or more block of information but classified as nonknowledgeable on the other(s). This possibility is applicable to about six out of 24 (25%) participants. Nevertheless, the probe-irrelevant mean amplitude difference was significantly larger in the knowledgeable group relative to the nonknowledgeable group.

Two experiments tested the CTP's performance with verbal stimuli (i.e. the participant's hometown) in the verbal and auditory modalities (Rosenfeld, Ward, Frigo, Drapekin, & Labkovsky, 2015; Deng, Rosenfeld, Ward, & Labkovsky, 2016). In Rosenfeld, Ward, Frigo, Drapekin, and Labkovsky (2015) 10 participants were visually or auditorily (about 72 dB) presented with the name of their respective hometowns (probe) where the modality alternated from verbal to audio. The irrelevants were the names of other cities (i.e. Atlanta, Buffalo, Orlando, Pittsburgh, Stockton, & Wichita). However, the target-nontargets were only presented visually. Their findings point to a cross-over interaction where probes P300s were larger for the visual than auditory modality, and irrelevants P300s were larger for the auditory modality. The overall average number of bootstrapped iterations where probe amplitudes were greater than irrelevants was 83% in the audio modality and 98% in the visual modality, but in terms of individual diagnostic effects, they were able to correctly identify 9/10 participants in the visual modality and 6/10 in the auditory modality. Deng, Rosenfeld, Ward, and Labkovsky (2016) sought to replicate the previous study in addition to verify whether the presentation of target-nontargets in both modalities also made a difference. This last group conducted two experiments

where target-nontargets were always presented auditorily (Exp 1) and where target-nontargets were simultaneously auditory and visual (Exp 2). In experiment 1, the individual detection rates were about 80% and 84% in auditory and visual modalities respectively, while in experiment 2, detection rates were about 84% and 95% in favor of the visual modality. However, no individual diagnostics scores were reported. In sum, those two experiments indicated that the preferred modality in CTP experiments, relative to auditory, is the visual presentation of stimuli.

Pictorial Stimuli Only & Pictorial-Word Stimuli

Pictorial stimuli have not been studied with the CTP technique as extensively as words but have generally produced better results. Labkovsky and Rosenfeld (2014) introduced a novel version of the CTP by combining the presentation of two probes, words (the name "Meixner") and images (a USB drive) within the same presentation. The verbal probe/irrelevant/target stimuli were displayed in the CTP fashion, without the typical 11111, 22222, 33333, etc. targets and non-targets, and the pictorial probe/irrelevant/target stimuli were displayed using the three-stimuli format.[8] A typical CTP presentation of one block of stimuli generally lasts about 15 minutes. This novel conception was believed to be more efficient, as opposed to the more time-consuming option of conducting two successive tests with the conventional CTP, and to alleviate possible fatigue concerns from examinees. Participants performed a similar mock crime as illustrated above in Winograd and Rosenfeld (2011). Individual diagnostic rates of 73% (11/15) and 100% (14/14) were observed with guilty and innocent groups respectively in relation to verbal stimuli. Diagnostic rates for pictorial stimuli were higher at 93% for both guilty and innocent groups. These figures were based on a .9 confidence level.

Rosenfeld, Ward, Thai, and Labkovsky (2015) compared both modalities in a between-group design but using the conventional CTP. Their findings supported the superiority of pictorial stimuli, based on the mean group amplitude difference between probe and irrelevant items, but no individual diagnostic values were reported. In Rosenfeld, Ozsan, and Ward (2017), participants viewed videos of a mock crime from the visual perspective of the person committing the crime (a gloved hand shown opening a drawer and reaching inside to steal a precious object). One group was instructed to imagine the situation as if they were witnessing the event while the

[8] The three-stimuli protocol involves the presentation of a probe, an assortment of irrelevant items, and a target stimulus to maintain attention. It is the precursor to the four-stimuli CTP.

other group was asked to visualize themselves as the author of the crime. Interestingly, both groups showed P300 effects although larger amplitudes were noted for the suspect group. In another investigation looking into financial incentive as a potential motivation factor, P300 amplitude probe-irrelevant difference from guilty volunteers who committed a mock crime (stealing a watch or a bracelet from a desk drawer) appeared moderate (Rosenfeld, Sitar, Wasserman, & Ward, 2018). A visual inspection of a line graph (Fig. 3f, p. 46) shows bootstraps iterations for the guilty group at approximately 76%, which means that about 76 iterations out of 100 of probe-irrelevant differences were significantly larger. There was no control group involved in this study as the guilty-innocent difference was not one of the experimental objectives. Consequently, it is unknown how innocent controls would have reacted. Instead of reporting the confidence level, the authors used Bayes factors (BF) to confirm the likelihood of retaining the null hypothesis instead of the alternative hypothesis (Rosenfeld, Sitar, Wasserman, & Ward, 2018). In this case a BF of 3.52 favoring the null hypothesis was obtained. According to Jarosz and Wiley (2014) (see table 4), a BF of 3 to 10 is considered substantial, while a BF > 150 is decisive. Individual classification scores were not computed in none of these experiments.

The research of Lu et al. (2017) is a semi-independent study of the CTP to make use of pictorial stimuli. Thirty-six volunteers were equally divided into a group of guilty participants acting alone and another group of guilty persons acting collaboratively. Individual guilty subjects enacted a conventional mock crime of stealing a ring from a jewelry case in a nearby office and hid the object in a secret place outside the room. A collaborative team of two guilty individuals performed the same mock crime and were encouraged to engage in low conservations as they committed the crime. Participants of both groups were then individually tested immediately thereafter. No innocent control group was recruited but seven randomized values from a "standard normal distribution" were drawn, with one of them acting as the probe, to generate a simulated data set for an innocent group. This process was replicated 16 times to simulate a group of 16 innocent subjects. The origin of this "standard normal distribution" is not specified in the publication. Individual diagnostic rates were not as impressive as previous claims of perfection made in other Rosenfeld-led research. Bootstrap analyses showed a detection rate of 75% (12/16) in the individual group compared to 25% (4/16) in the collaborative group. The collaborative group was undistinguishable from the simulated innocent group, suggesting an important vulnerability of the CTP to mistakenly identify a collaborative perpetrator for an

innocent examinee. A more realistic comparison with true innocent controls would likely settle the matter more convincingly.

Countermeasures

Any forensic instrument is only as good as its ability to resist to CMs. The polygraph is widely reported to be vulnerable to physical countermeasures (National Research Council, 2003). An examinee intent on influencing the outcome of a polygraphic test by producing physiological responses may engage in tongue biting or pressing toes against the floor to lead the examiner to conclude of their truthfulness or an inconclusive result (Vrij, 2008). Other anecdotal CMs with the polygraph include pressing toes against thumb tacks place inside the shoes or tensing the anal sphincter muscle while answering crucial questions. Polygraph examiners have adapted their techniques by asking examinees to remove their shoes for the test and by adding an additional signal (a sensor pad on the seat portion of the examinee's chair) to detect muscular activity when the person is seated. Lie detection tests with fMRI are just as vulnerable. Another CM could be when suspects prepare their lies in advance. Ganis, Kosslyn, Stose, Thompson, and Yurgelun-Todd (2003) first tasked participants to record their most memorable work and vacation experiences and then asked them to develop alternative scenarios built around these momentous experiences. They examined their neural activity when they either responded with memorized or spontaneous lies. They found different neural activation patterns between spontaneous and memorized lies but they did not attempt to detect the truthtellers from liars. Ganis, Rosenfeld, Meixner, Kievit, and Schendan (2011) reported a success rate in detecting deception of 100% when no CM was employed but it dropped to only 33% once right-handed participants made indiscernible hand digits movement resulting in most cases to be classified as false negatives (i.e. ruled deceptive when honest in reality). Rosenfeld et al. (2008) also tested physical CMs in EEG-based CIT. As described above, the experiment involved three blocks of trials spread over three weeks. In week 2, participants were asked to apply anyone of four CMs which implicated the undetectable movement of fingers or toes or imagining being slapped in the face by the examiner. Their findings showed a decrease in CTP performance ranging from detection rates of 90% (9/10) at .9 confidence level, 82% (9/11) at .95 confidence level when using the Iall method, and 70% (7/10) at .9 confidence level, 73% (8/11) at the .95 confidence

level when using the Imax method. However, CM use was discernible through elevated reaction time (RT) scores.

Examinees may also employ mental CMs such as thinking about something or someone specific (i.e. a frightening event, counting backwards) (Vrij, 2008), inhibiting the stimuli they are anticipated to recognize (Hu, Bergström, Bodenhausen, & Rosenfeld, 2015; Ward & Rosenfeld, 2017). In two experiments, Meixner, Haynes, Winograd, Brown, and Rosenfeld (2009) investigated the role of task demand on CTP performance by forcing the subject to make countermeasure-like responses to stimuli. The probes for each experiment was the participant's own birthdate assorted with four irrelevant dates. In experiment 1, participants in the CM condition were instructed to attempt to increase the salience of two irrelevant dates by silently saying their first name as they saw one date and to say their last name as they saw the other date, all the while pressing the same button on the response box with the same digit of the left hand. The same trial structure was applied in experiment 2, with the difference that different dates were assigned to different button box presses by either the middle finger or the index finger of the left hand. Using the Iall method detection rates decreased substantially in the CM condition of both experiments, 5/11 (Exp 1) and 10/16 (Exp 2).

Meixner and Rosenfeld (2010) examined the effect of response omission to designated stimuli as a possible CM technique. Participants were shown five dates as stimuli with the probe being their respective birth dates assorted with four irrelevant dates and instructed to press the left hand five-button box where each button was assigned a date. In one condition all stimuli required a button press while in others the instructions were conditional button presses depending whether the date was a probe or an irrelevant. Their detection rates ranged from 10/12 (Guilty no omit), 9/12 (Innocent omit irrelevant), and 12/12 (Guilty omit probe), suggesting an *omit* effect when the probe is the only uncountered item.

In other variations of possible mental CM studies, Rosenfeld and Labkovsky (2010) exposed participant's date of birth as probes and tested the efficiency of imagining one's first and last name as a possible mental CM. Detection rates were 100% (12/12) using the Iall method and 92% (11/12) when using the Imax method. Sokolovsky, Rothenberg, Labkovsky, Meixner, and Rosenfeld (2011) looked at sequential versus simultaneous mental CMs (i.e. either making a button press, in response to seeing the probe birthdate, after silently saying the first or last name or pressing at the same time). Detection rates were similar across all groups (10/12 for simply

guilty, 11/12 for sequential CM, and 11/13 for simultaneous CM). RT scores enabled investigators to detect CM use in serial responders but not for the simultaneous ones, and RT numbers for simultaneous participants were indistinguishable from controls.

In an attempt to investigate the ability of RT measures as a possible CM detector, Hu, Hegeman, Landry, and Rosenfeld (2012) increased the number of irrelevants to eight and manipulated the CM to irrelevants ratio from 25%, 50%, to 75%. In other words, participants were asked to silently say their first or last name to two, four and six designated names of towns (the probe being their hometown). Their findings showed that individual detection rates for guilty subjects decreased as the number of CMs increased, from 92% (2-CM), 83% (4-CM), to 71% (6-CM), but elevated RT measures was found to be an effective solution to detect simultaneous use of CM in the 50% and 75% proportion groups. Reaction times were also studied in relation to the number of CMs applied to irrelevants (Labkovsky & Rosenfeld, 2012). In what they refer as a sequential CM (e.g. a button press on one of the five button press box is made only after the participant observes the to-be-countered irrelevant item (dates) and then silently thinks of one of the CM to apply), participants were instructed to perform a mental CM to one of five stimuli (i.e. CM1 = last name, CM2-3-4-5 = names of meaningful people, mother, father, siblings, friends). The detection rates of the probe birthdate varied from 92% (CM1-4-5) to 100% (CM2-3).

Three studies were performed in relation to memory suppression as another possible CM. First, in Hu, Bergström, Bodenhausen, and Rosenfeld (2015) guilty participants executed the usual stealing-the-ring-from-a-psychology-staff-member's-mailbox mock crime and were shown the verbal probe 'ring' and non-pertinent jewelry items. A standard guilty group was not given any CM instruction whereas the suppressed-guilt group was instructed to dismiss the memory of the lab-based crime and not to allow it to come to mind during the test. They were also asked not to generate self-distracting thoughts. Evidence of probe and irrelevant P300 amplitude reduction but larger LPN curves among the suppressed-guilt group were observed. Participants in the standard guilty group had distinguishable P300 amplitude probe-irrelevant differences compared to the suppressed-guilty group, but the latter group exhibited large enough LPNs to reveal their guilt. No individual classification data was reported. This may point however to two possible components involved in the recollection process where P300 supports the idea of conscious recall and the dissociable LPN is linked to response-monitoring processes, but more research is

required in relation to LPN. Rosenfeld, Ward, Drapekin, Labkovsky and Tullman (2017) replicated the work of Hu, Bergström, Bodenhausen, and Rosenfeld (2015) with the given name as a probe and found opposite results. P300 amplitude effects, as measured with the PP method, were larger for guilty subjects attempting to suppress their episodic memory. Ward and Rosenfeld (2017) replicated the Hu, Bergström, Bodenhausen, and Rosenfeld (2015) experiment once again by modifying the target to nontarget ratio from 50-50, as had been used in the previous two studies, to 20-80. It was hypothesized that the latter ratio would require less response switching, and less task demand, and therefore free up cognitive resources. Although the authors claim detection accuracy rate of 90% or higher for both guilty groups (standard and suppression), no individual classification data was provided.

Levels of Processing

A line of investigation explored in the experiments outlined in this manuscript was the influence of processing depth on memory performance, specifically with the CTP. As described above Meixner and Rosenfeld (2014) were able to discriminate perfectly 12 knowledgeable and 12 nonknowledgeable subjects in their P300-CIT based experiment using the CTP. To recap, participants were asked to wear a video-recording device for several hours as they walked about their daily activities around the university campus. Upon returning to the laboratory the next day, participants in the knowledgeable group were shown three verbal probes (e.g. name of friend, class attended, and location of lunch), one in each block, extracted from their own video footage, and assorted with five irrelevant word items. Nonknowledgeable subjects viewed probe items unrelated to their video recording. However, the results obtained from the detection of incidentally acquired memories of real events did not permit to conclude the origin of this perfect outcome. Perhaps the precise hit rate was simply attributable to the CTP's greater methodological sensitivity or to the encoding quality during the creation of episodic memories which may have somehow improved individual performance. Of relevance here is the contribution of Zinchenko (as reported in Craik & Lockhart, 2008) who postulated a central role for physical activity in memory encoding and retrieval. "]M[aterial is better remembered when it relates to the target of action rather than to peripheral conditions." (Zinchenko cited in Craik & Lockhart, 2008, p. 56). The possibility that physical activity aided the participants' memory performance in Meixner and Rosenfeld (2014) cannot be discounted.

It is fairly well established now that memory can be sorted into levels of storage and that durability of memorable traces are a positive function of depth of processing (Craik & Lockhart, 1972; Craik & Tulving, 1975). Memories can be preserved in three repositories: sensory registers, short-term memory (STM), and long-term memory (LTM) (Craik & Lockhart, 1972). The first category is characterized mainly by its detachment from attentional demand, or preattentive sensory stores, its fairly large capacity but rapid decay with a trace duration of approximately ¼ - 2 seconds (Craik & Lockhart, 1972). The principal features of the second class are its requirement on continued attention, its limited capacity but slower rate of forgetting, and a trace duration of up to 30 seconds (Craik & Lockhart, 1972). LTM is predominantly distinguished from the previous two stores by the operations of rehearsal, repetition and organization of incoming information, its endless capacity, a very slow declining rate, and a trace duration ranging from minutes to years (Craik & Lockhart, 1972). A substantial body of documented research illustrates that enduring properties of memories are influenced by the degree of semantic involvement (Craik & Tulving, 1975). Through a series of 10 experiments, Craik & Tulving (1975) confirmed the framework proposed earlier by Craik & Lockhart (1972). In short, deeper encodings (deep processing), relative to superficial aspects of words (shallow processing), are the product of greater semantic involvement, such as the richness, elaborateness, and the spread or number of encoding features (Craik & Tulving, 1975). In other words, verbal stimuli are more likely to be remembered when they are integrated into a meaningful and complex sentence involving special typescripted, rhyming pluri-syllabic words (Craik & Tulving, 1975) or as capitalized words at the end of normal (deep) rather than superficial (shallow) sentences[9] (Loaiza, McCabe, Youngblood, Rose, & Myerson, 2011).

Research about levels of processing (LOP) and their influence on encoding and retrieval processes is fairly consistent across studies and modalities. Deeper LOP aid at the encoding stage and benefit recollection performance compared to surface processing (Gallo, Meadow, Johnson, & Foster, 2008) regardless if attentional levels at encoding are disrupted (Knott & Dewhurst, 2009), distracted (Loaiza, McCabe, Youngblood, Rose, & Myerson, 2011) or if pictorial stimuli are used (Marzi & Viggiano, 2010). For example, in studies by Gallo, Meadow, Johnson and

[9] Examples of deep phrases are “The brother of one of your parents is an UNCLE” or “A tool for making clothes is a sewing MACHINE”, and shallow sentences are “A word made up of five letters is UNCLE” or “There are three different vowels in the word MACHINE”.

Foster (2008) that investigated the distinctiveness heuristic, findings of fewer errors on recollective performance in the deep task (deciding whether the word shown is pleasant) relative to a shallow task (deciding whether the word contains the letter 'e'), were consistent despite shallow items being presented more often than deep items or reverse-cued[10]. Only when words were transcribed before testing that LOP effects disappeared. The transcription of words increased their distinctiveness as well as their recollective accuracy, and by the same token reduced false recognition errors. However, it is not necessarily the LOP alone that contribute to greater recall since qualitative memory trumps quantitative differences in retrieval memory processes of words (Gallo, Meadow, Johnson, & Foster, 2008). Repeating shallow words may have increased their familiarity but it did not eliminate false recognition effects. "]T[he most effective way to encode items for subsequent recall or recognition is to associate each item with information from pre-existing knowledge that can later provide a large number of unique features to retrieve." (Gallo, Meadow, Johnson, & Foster, 2008, p.1109). Marzi and Viggiano (2010) obtained similar results in a facial recognition investigation where LOP were manipulated. ERP amplitudes, as indexed by the N170 component, were larger with deeply encoded faces. Viewed from the recollection and familiarity dichotomy, it appears that "]d[eep encoding conditions therefore increase recollective retrieval, with shallow encoding conditions leading to recognition based familiarity" (Knott & Dewhurst, 2009, p. 1049) in the case of verbal stimuli, although Marzi and Viggiano (2010) found that deep encoding of pictorial stimuli involved both processes with a fronto-parietal distribution.

Four studies examined LOP in a context more closely related to the focus of this research. Ferlazzo, Conte and Gentilomo (1993) presented semantically related (deep processing) and rhyming word pairs (shallow condition) to nine volunteers in an old/new paradigm. They found enhanced P300 amplitudes for deeply processed items compared to shallowly processed ones, a clear effect of encoding instructions at the recognition phase, but RT data analysis failed to show any significant relationships. Van Hoof and Golden (2002) tested the effectiveness of ERP-based memory detection of learned words versus rote repetition of words in an oddball paradigm. This design was not a CIT in the classical sense. Volunteers were instructed to learn a list of neutral non-semantically related words (learned targets) over an unlimited amount of time and later

[10] Reverse-cuing refers to a procedure where participants are presented with a study word before being prompted to make a deep or shallow encoding judgment on that word.

tested on recall. They studied another list and were tested on both lists which created learned nontargets, repeated targets and new words. Behavioral key presses, RT and EEG were recorded. Significant P300 were observed for learned targets and learned nontargets and not for repeated and new words. The takeaway message from this research is that ERP-based memory assessment is likely best effective in situations where words are intentionally learned and least successful with words that were formed out of weak memory traces. No individual hit rates were analyzed in this study. Seymour and Fraynt (2009) used a three-stimuli protocol CIT to measure RT in the presentation of two-worded phrases classified in six sub-groups (e.g. names – "Phil Jenks", street names – "Perch street", file description – "Rain File", articles of clothing – "Blue Coat", and operation names – "Op Horse"). Each sub-group was further organized according to the three-stimuli protocol such that six items (one from each category) were randomly defined as probe items, target items and irrelevant items. Participants in the shallow condition were asked to read a short mock newspaper article about a fictitious theft on campus and were given three minutes to study it whereupon they had to paraphrase it later. Subjects in the deep condition were instructed to complete four tasks (picture matching, word jumble, handwriting, and word shouting) designed to increase exposure duration, attention and to promote richly encoded multimodal memory representations. Both groups were further assigned to three delay conditions: 10 min, 24 hours, or 1 week. General accuracy rates of 93% in classification of familiar probes versus unfamiliar-probes were observed regardless of probe elaboration or delay. The classification accuracy dropped slightly to 90% after one week in the case of well-encoded probes but decreased down to chance level for poorly memorized items. Two methodological issues are worth mentioning here. First, Seymour and Fraynt (2009) used reaction time as a proxy for cognitive processing. This behavioral measure of neural treatment of memory processes is less than ideal compared to EEG. Second, they used a three-part algorithm, made up of a distribution analysis based on three tests (Kolmogorov-Smirnov Test, *F*-test, and Fisher's exact test) to tease apart probe and irrelevant items in their individual classification instead of the bootstrapping method which is considered the diagnostic method of choice in P300-CIT based psychophysiological research (Rosenfeld & Donchin, 2015).

Gamer and Berti (2012) probed the issue further where central and peripheral details were compared in a mock crime paradigm. Central items were viewed as deep processing because they would be handled by the author of the crime (e.g. theft of an object) during the execution of

the crime. In contrast, peripheral features were defined as shallow encoding since they would be unrelated, yet present, to the execution of the crime (e.g. lamp on a desk, framed painting on a wall). Gamer and Berti (2012) conducted their experiment according to the three-stimuli CIT protocol and measured EEG and skin conductance response (SCR). All participants were tested one week after the execution of the mock crime. Consistent with previous research, they reported high recognition rates of centrally and peripherally processed items as indexed by the P300 and electrodermal reactions, and longer response time for probes relative to irrelevants (Gamer & Berti, 2012). Individual hit rates were not reported. In sum, this means that P300, at Pz, and SCR are two measures relatively resistant to varying LOP.

The Proposed Research

The experiments reported in the following chapters represent three objectives of this research. In chapter 2, in light of the fact there exists only one truly independent replication of the CTP (Lukacs et al., 2016), we set out to add to that unique study and verify the CTP's effectiveness in a nearly identical paradigm with autobiographical data. We chose to deviate from the European team but only with respect to the CM. The counting-backward-CM we tested had never, to our knowledge, been applied to the CTP.

In chapter 3, another objective of this dissertation was to scrutinize the roles of LOP and stimuli type (pictorial versus verbal) in the CTP's performance of a mock theft. We asked if the CTP was sensitive enough to detect participants involved in the physical manipulation alone of a probe item, or its manual handling coupled with a semantic task, the latter aimed at affording a deeper encoding experience. And whether that sensitivity was any different when the probe item was displayed as a word or as its pictorial referent.

Third, in chapter 4, the final aim of the research was to verify the CTP's performance in a study that combined together a live episodic event (i.e. mock terrorism scenario), a face to face encounter with an accomplice male, the physical manipulation of components of a mock bomb in the context of a mock crime scene, the pictorial presentation of three types of stimuli (i.e. faces, crime scenes, and explosive devices) in sequential blocks, and the independent replication of the memory suppression CM from Hu et al. (2015).

CHAPTER 2: AN INDEPENDENT VALIDATION OF THE EEG-BASED COMPLEX TRIAL PROTOCOL WITH AUTOBIOGRAPHICAL DATA AND CORROBORATION OF ITS RESISTANCE TO COUNTERMEASURE

Abstract

This experiment is the second independent validation of the Complex Trial Protocol (CTP), a P300-based Concealed Information Test (CIT). The theoretical underpinnings of the CIT in the context of law enforcement usage are considered to be sound. The CTP is said to effectively discriminate individuals who recognize novel and meaningful stimuli and to be countermeasure resistant. Forty-five undergraduate students were assigned to three groups and instructed to perform a computer task using autobiographical data. P300 peak-to-peak amplitude differences between probe (surname) and irrelevant (patronymic foils) items accurately identified 100% (14/14) of Innocent Controls (IC), 94% (15/16) of Simply Guilty (SG) participants, and 93% (14/15) of GCM subjects who were asked to counter the stimuli with the simultaneous performance of mathematical operations. Increased levels of combined cognitive and behavioral errors significantly detected GCM from IC and SG individuals. Longer Reaction Time (RT) was only significant between GCM and IC groups. Implications for forensic issues are also discussed.

Keywords: Concealed Information Test, ERP-based CIT, P300-based CIT, Complex Trial Protocol

Introduction

The seeds of the Guilty Knowledge Test (GKT) were sown in 1932 with the kidnapping of 20-month-old Charles Lindberg Jr., son of Anne Lindbergh and famed aviator Charles Lindberg. William Marston, a student of Harvard psychology professor Hugo Münsterberg, approached accused Bruno Hauptmann's defense counsel with the proposition to test Hauptmann's claims of innocence of the Lindberg crime (Lykken, 1998). Marston was of the view that details of the abduction could be stored in the real kidnapper's brain, and with the aid of his newfound lie detector machine he could confirm or disprove Hauptmann's culpability (Lykken, 1998) based on a discontinuous systolic blood pressure test he had developed in 1915 (Synnott, Dietzel & Ioannou, 2015). His efforts were rebuffed, Hauptmann was found guilty of the crime and put to death in the electric chair. The polygraph went through a series of developmental phases in the

decades that followed, ultimately heralded, by the middle of the 20th century, as the instrument of choice by law enforcement to detect deception (Synnott, Dietzel & Ioannou, 2015).

Long overshadowed by the more popular and yet controversial Comparison Question Test (CQT)[11], it was not until 1958–1959 that a more scientifically sound test would be developed out of serendipity and ignorance of the field of lie detection (Lykken, 1998). David Lykken, a clinical psychologist and neuropsychiatrist, agreed to supervise two young medical students over the Summer break, and further consented to the student's inquisitive inclination to perform a lie detection experiment with newly acquired polygraphic equipment and in relation to a subject he knew nothing about (Lykken, 1998). The group recruited an undergraduate sample who performed a mock theft and murder. As Lykken and his team of assistants worked to devise an interrogation strategy, they came to believe that lie detection was likely impossible, and as a result they turned their interest to finding evidence of guilt rather than deception (Lykken, 1998). They developed a series of questions to target what their pretend thieves and murderers had to know about their respective fake crimes and included equally plausible alternatives. They measured only one physiological signal, skin conductance, for simplicity sake. Their experiment was a great success, as 48 out 50 innocent suspects, and 44 out 50 guilty suspects were correctly classified respectively, and the GKT was born (Lykken, 1998). This test is now known by its more contemporary name, the Concealed Information Test (CIT).

EEG-based CIT experiments first appeared in academic circles with research from Rosenfeld in 1986 (Rosenfeld, Nasman, Whalen, Cantwell, & Mazzeri, 1987). After testing a few different methodologies, Rosenfeld and his colleagues settled on a four-stimuli protocol based on the traditional P300 oddball paradigm and named it the Complex Trial Protocol (CTP) (Rosenfeld et al., 2008).

The CTP involves the presentation of two types of stimuli, those associated with the crime under investigation and others meant as attention grabbing. The first type consists of either crime relevant items, only known to the offender and the authorities called '*probe'* (e.g. the murder weapon, the crime scene, the victim's face or wound patterns, etc.), or an assortment of equally plausible alternatives called '*irrelevants'*. For example, in the case of a murder investigation the probe could be the weapon used in the homicide (i.e. pistol), and irrelevants could be alternatives

[11] The CQT was developed by John E. Reid in 1947 who went on to establish the first polygraph school in the USA. For a more detailed history of the polygraph see Synnott, Dietzel, and Ioannou (2015).

such as a baseball bat, an axe, a rifle, etc. The second type of stimuli forces attention of examinees onto the computer monitor. They conventionally comprise of strings of numbers (i.e. 11111) as the *'target'*, and *'non-targets'* (i.e. 22222, 33333, 44444, and 55555) and require button presses from a separate computer mouse. The CTP separates both tasks into an implicit probe/irrelevant discriminatory judgement from one mouse, and an explicit target/non-target forced attention decision from another mouse (Rosenfeld et al., 2008).

The majority of experiments using the CTP have been, and continue to be, generated from the Rosenfeld laboratory. To our knowledge, only one experiment to date has independently validated the CTP. Lukacs et al. (2016) undertook to test the effectiveness of the CTP using autobiographical identifiers as well as investigating its vulnerabilities to countermeasures. Their findings corroborated claims by Rosenfeld-led teams that the CTP accurately classifies participants as guilty or innocent (based on the recognition of autobiographical data) but found the CTP to be more vulnerable to countermeasure than previously documented. As in several of Rosenfeld-led studies with given or last names (Rosenfeld, Ward, Drapekin, Labkovsky, & Tullman, 2017; Rosenfeld et al., 2008; Rosenfeld J. P., Labkovsky, Davydova, Ward, & Rosenfeld L., 2017), Lukacs et al. (2016), presented their participants with their respective family name as probes (except for the control group) assorted with Hungarian last names as non-critical alternatives. Their findings were consistent with those of Rosenfeld et al. (2008). Lukacs et al. (2016) respectively identified 14/15 (93%) and 13/14 (93%) participants in their 'simply guilty' and 'innocent control' groups at a confidence level of 90% relative to detection rates of 92% (11/12) at the same confidence level in the simply guilty condition in the Rosenfeld led study.

In the current study, we sought to replicate Lukacs et al. (2016) with surnames, and by extension the Rosenfeld-led experiments into autobiographical data as well. Consequently, we expected our participants in the simply guilty (SG) condition to show a significantly elevated P300 when shown their family names relative to irrelevant family names (Hypothesis 1).

Another claim of the CTP is that it is believed to be countermeasure (CM) resistant (Rosenfeld et al., 2008). CMs are most often mental actions taken by an examinee to defeat a test, by either increasing the saliency of a non-critical alternative stimulus or decreasing the prominence of the probe item. A successful CM would seek to make a true positive participant (guilty condition) appear as a false negative (guilty condition but masking as an innocent

control). In terms of an EEG-based CIT, CM examples include the use memory inhibition (Hu, Bergstrom, Bodenhausen, & Rosenfeld, 2015), participants imagining their own first name, last name, or names of meaningful people (mother, father, siblings or friends) (Rosenfeld & Labkovsky, 2010; Labkovsky, & Rosenfeld, 2012; Hu, Hegeman, Landry, & Rosenfeld, 2012), silently and mentally say their first and last names either sequentially or simultaneously with their birthday as the probe (Meixner, Haynes, Winograd, Brown & Rosenfeld, 2009; Sokolovsky, Rothenberg, Labkovsky, Meixner, & Rosenfeld, 2011), omitting to press the button only upon presentation of the probe item (birthdate) (Meixner & Rosenfeld, 2010).

Rosenfeld-led teams of researchers have published nearly 30 papers to date of laboratory-based experiments involving the CTP. Of those publications that dealt with CMs the authors have repeatedly maintained the CTP's ability to resist CMs and/or detect CM users (Rosenfeld et al., 2008; Meixner & Rosenfeld, 2010; Rosenfeld & Labkovsky, 2010; Sokolovsky, Rothenberg, Labkovsky, Meixner, & Rosenfeld, 2011; Hu, Hegeman, Landry, & Rosenfeld, 2012; Labkovsky & Rosenfeld, 2012; Hu, Bergström, Galen, Bodenhausen, & Rosenfeld, 2015; Ward & Rosenfeld, 2017; Rosenfeld, Ward, Drapekin, Labkovsky, & Tullman, 2017). There is one exception. Findings from Meixner, Haynes, Winograd, Brown, and Rosenfeld (2009) suggested that a high level of task demand associated with a mentally demanding CM (i.e. participants silently said their first name when presented with a date, and silently said their family name when they saw another date) caused reduced hit rates ranging from a hit rate of 36% (4/11) for CM users in one experiment using random responses to a hit rate of 44% (7/16) for CM users in another experiment using assigned finger-stimuli button responses. In all other experiments where CMs were tested, Rosenfeld and colleagues have claimed that CMs did not prevent accurate guilt judgment with hit rates ranging from 87% to 100%. Ben-Shakhar (2011), however, recommended that studies implicating CMs ought to be replicated in other laboratories before practical implications were made about the CTP.

We sought to accomplish just that with a mentally demanding CM. Lukacs et al. (2016) tested two CMs, silently saying the given name of close relatives or short one or two syllabic words (i.e. up/down, dog/cat) simultaneously or immediately after the button response. Their findings indicated that the short-word CM (12/15, 80%, AUC = 0.943) was somewhat more effective than the given-name CM (14/16, 87.5%, AUC = 0.929), despite not being significantly different from each other, but when compared to hit rates of 83% (10/12) from Rosenfeld et al.

(2008) the CMs in Lukacs et al. (2016) appear to be as equally ineffective. However, the CMs employed in the Rosenfeld-led experiment were physical and cognitive in nature (i.e. increase pressure of a digit finger on the leg, wiggling of toe inside the show, and imagining being slapped in the face by the operator), whereas the CMs in Lukacs et al. (2016) were mental.

In their initial report about the CTP, authors expected to see elevated P300 for probes relative to irrelevant stimuli even with subjects using a preferred CM method (Rosenfeld et al., 2008). However, a concern was identified in Meixner and Rosenfeld (2010) about the possibility of giving the CTP examiner an unfair advantage. In fact, an 'omit effect' had been confirmed when the CM was applied only to irrelevants but not the probe stimulus. In other words, the probe item gained saliency by virtue of being the only stimulus that was not countered. Rosenfeld and Labkovsky (2010) verified this possibility by asking a sub-group of participants to perform mental CMs to two of four irrelevant items. Their hit rate for the CM group was 100% (12/12) leading them to suggest that the omit effect was mediated by the P300 oddball effect. An elevation of the P300 as an indication of recognition appears to be unaffected by CMs, at least those employed in these experiments.

It has been proposed that lying is cognitively demanding and purposely imposing mentally taxing interventions would result in more pronounced behavioral displays of cognitive load (Vrij, Fisher, Mann & Leal, 2008). Furthermore, as mentioned above, mental task demand was also a critical factor in reducing test effectiveness (Meixner, Haynes, Winograd, Brown, & Rosenfeld, 2009). We posited that silently conducting mental arithmetic operations during the CTP trials would take away cognitive resources from accurately following instructions required by the CTP (i.e. button presses when shown attention loaded target items), and relative to a control condition, manifest itself in either longer reaction times (Suchotzki, Verschuere, Bockstaele, Ben-Shakhar, & Crombez, 2017; see also Rosenfeld et al., 2008) and/or an increase in error rates on button presses. We also predicated our expectations that cognitive efforts expended for the demanding task would take away mental resources needed for following instructions about button presses on target-non-target attention-focus stimuli. If the CTP is as resistant to CMs as claimed by Rosenfeld and colleagues, we hypothesized that participants using a CM would be detected by their longer reaction time and/or their increase in committing button press errors (Hypothesis 2). In a real-life application of an EEG-based CIT with the CTP, persons employing a CM would be

ruled as non-cooperative and their data not analyzed. We made an exception for this field rule in our laboratory-based experiment.

Method

Participants.

A power analysis using G*Power (Faul, Erdfelder, Lang & Buchner, 2007) estimated a sample group of about 42 (14 per group) was necessary to enable the detection of an effect size of f(V) = 0.7, or approximately 0.3 Cohen's *d*, at an alpha of .05. A total of 46 participants were recruited for this experiment. With the exception of one person, none reported being color blind, nor suffering from or diagnosed with a major psychological disorder (i.e. schizophrenia). The lone individual whose data were excluded declared to be taking medication for severe depression with anxiety issues. This left the data from 45 participants (13 males) for analysis. The mean age was 22.0 years (*SD* = 3.5) ranging from 18 to 34 years old. All were undergraduate students from Concordia University's psychology department and were offered a course credit for their participation. All had normal or corrected-to-normal vision and were fluent in English.

This research was authorized by Concordia University's ethics committee (certificate #30006647). All participants signed a written consent form prior to commencing the experiment. This document clearly explained the purpose of the research, the general procedure, the risks and benefits, and the conditions of participation, which included a confidentiality commitment from the experimenters.

Research design.

Following the CTP protocol as outlined in Rosenfeld et al. (2008) and described in the following sections, a 3 x 3 x 2, mixed between-within subject factorial design was used with groups (innocent-control, simply guilty, and guilty-countermeasure) as the between-subject variable, and sites (Fz, Cz, Pz), and stimuli (probe and irrelevant) consisting of the within-subject variables. The main dependent variable was the P300 cerebral amplitude (in μV) at sites Fz, Cz, Pz.

Procedure.

Volunteer participants were first asked to read and sign a written consent and to complete a demographic data sheet. They were then randomly assigned to one of three groups, Innocent-Control (IC) (n = 14), Simply-Guilty (SG) (n = 16), and Guilty-Countermeasure (GCM) (n = 15).

Individuals in the SG group were handed out a scripted scenario that briefly outlined details of a burglary that had occurred at the university, and that upon obtaining a list of students' family names the focus of the police investigation had now turned towards the participant (see Appendix A). They were invited immediately thereafter to enter the room for testing.

Persons assigned to the GCM condition were given the same instructions as the SG group. However, prior to testing they were informed that they would be using a CM which consisted in having the participant to silently count backwards in 3 from 1,000 during the entire presentation. They were given a few minutes to practice and once they indicated their readiness to resume, they were allowed to begin the experiment.

Since no actual mock crime took place, guilt was simulated in the SG and GCM groups by virtue of presenting these participants with their respective family name as the probe stimulus. An elevated P300 amplitude difference between probe and irrelevant items from those in the SG and GCM groups was considered as an indication of possession of guilty knowledge.

Candidates of the IC group were handed the same scripted theft scenario as the other two groups prior to testing, but the probe stimulus was replaced with an irrelevant family name. The actual test lasted approximately 15 minutes irrespective of the condition, and the entire testing session carried on for about 60 minutes.

Trial structure and testing procedure.

Typically, the CTP involves the presentation of four types of stimuli on a computer monitor: a *probe* (the concealed item known only to the author of a crime and the authorities), *irrelevant items* (an assortment of similar stimuli acting as fillers to the probe item), a *target item* (a string of numbers, typically 11111), and a series of four *non-target items* (a series of strings of numbers, typically from 22222, … to 55555). Following a baseline of 100ms of recorded pre-stimulus brain activity, the stimuli, regardless of their type, are always presented for 300ms at the center of the computer screen.

In the CTP, a trial consists in the presentation of two stimuli. The pair is always made up first of a probe or an irrelevant item, randomly followed by a target or non-target item, and separated by a fixation cross. The button press response from one mouse to the first stimulus is intended to confirm that the participant has implicitly seen the stimulus in question (i.e. the "I saw it" response), while the conditional button presses from a second mouse in response to the second stimulus is meant to confirm the participant's explicit attention to the stimuli presentation

on the computer screen (Rosenfeld et al., 2008). The inter-stimulus interval varied from 1600 ms (from the random presentation of item 1 to item 2) to 2700 ms (from the random presentation of T/NT to the next T/NT).

Investigators instructed participants to press the right button from a mouse on the left as fast as they could each time they saw the first stimulus, a surname (see Appendix B). They were also told to immediately press either the right or left button from another mouse on the right when they saw the second stimulus. If it was the target item 11111, they had to press the right button, and if it was one of any non-target items (i.e. 22222, 33333, 44444, or 55555), they had to press the left button on that same mouse. In keeping with Rosenfeld's practice, participants who normally commit more than 20% of button press errors on either stimulus are usually excluded, but in this case they were not because button press errors was a dependent variable to measure the efficacy of a mental CM. These mistakes were called behavioral errors.

Finally, the CTP investigator usually interrupts the experiment periodically to seek the identity of the last stimulus seen by the participant. According to Rosenfeld et al., (2008), this step further ensures that the participant is fully attentive, cooperative and not employing any countermeasure. However, the literature is quite unclear in terms of the number of interruptions. A review indicates that the number of pop quizzes varies from a little as 7–12 trials (Rosenfeld, Ward, Frigo, Drapekin, & Labkovsky, 2015) to as high as 50–70 trials (Deng, Rosenfeld, Ward, & Labkovsky, 2016). One CTP experiment did not report its number of inquisitive pauses (Meixner & Rosenfeld, 2010), while some others remained vague about it, such as stating that the number of interrogative stops were made every few trials (Dietrich, Hu, & Rosenfeld, 2014), that participants' attention was monitored but not maintained (Labkovsky & Rosenfeld, 2014), that surprise tests were conducted at unpredictable times (Labkovsky & Rosenfeld, 2012; Rosenfeld, Ward, Drapekin, Labkovsky, & Tullman, 2017), or that the procedure was occasionally paused (Lu et al., 2017).

No finding has ever been offered to date as to the optimal number of interruptions. We took a middle of the road approach and quizzed our participants about every 40 trials, ranging from 38 to 50 trials (M = 43.0), for a total of nine pauses over 374 stimuli presentations. As this step could be interpreted as a reinforcement procedure, care was taken to ensure that a pause coincided with all stimuli at least once. Since there were seven items, six irrelevant and one probe, and nine pauses (including one practice pause), this allowed for two items to be reinforced

twice. As a precaution against a possible reinforcement of the probe stimulus, investigators chose two irrelevant items for the extra interruptions. As in all of Rosenfeld's experiments, participants were informed prior to testing that they would be questioned periodically on the last stimulus seen, and that more than two errors, or a 25% error rate, would lead to their data being set aside. This type of mistake was called a cognitive error. Participants were not informed of a practice session, but investigators edited out the first 20 presentations which included a surprised test. To summarize, participants were shown a total of 374 presentations but only 354 were kept for analysis.

Stimuli.

As in Winograd and Rosenfeld (2011) and Lukacs and colleagues (2016), we used one probe and six irrelevant items in our experiment. The stimuli consisted of one block of family names where the probe was the participant's last name. Irrelevant stimuli were selected from a list of patronymic equivalents in keeping with Lykken's (1998) plausibility criterion that irrelevant items ought to be "equally plausible alternatives" (p. 39). Prior to being tested all participants were shown a list of 20 surnames selected from the website www.mongabay.com listing the most to least common cognomens in America according to the 1990 US census (See Appendix C). They were then asked to point out any of the names that were meaningful to them (e.g. matching the name of close relative, friend, etc.) or strikingly unique such that it stood out too much, and those were excluded. A total of six family names were retained from the remainder of the list.

Stimuli presentation was done through PsyTask and displayed on a 55cm HP Compact (LA2206x) flat monitor with a 1280 x 1024 resolution in a dimly lit room. The average stimuli size was 6.77cm x 6.93cm. At a viewing distance of 63.5cm (measured from the participant's right eye to the center of the screen) the average stimulus subtended 5.5° x 6.1° of visual angle. The viewing distance from the participant's nasion to the fixation cross at the center of the monitor was 61cm. All items were presented in Black on a White background surrounded by a wide Black edge.

EEG data acquisition.

The software EEG Studio was used to collect the data, which was recorded with a Mitsar amplifier, model 201 (Mitsar company, St-Petersburg, Russia) sampling at 500 Hz, and seven conductive gel filled Ag/AgCl electrodes. The ground electrode was placed on the forehead

above the corrugator muscles. The electrooculogram (EOG) electrode was placed approximately one cm above the center of the left eyebrow. Three electrodes were attached to the scalp midline at sites Fz, Cz, Pz and referenced to linked mastoids. In accordance with the International 10–20 system, and prior to being tested, the distance between the inion and the nasion for each participant was measured such that the Cz electrode was consistently placed at the 50% mark on the scalp. Participants were asked to refrain from making head and upper torso movements, speaking or fidgeting in their seat, and to keep their feet flat on the floor during the test. Impedance between the scalp and electrodes was kept at below 5 KΩ. Signals were passed through the amplifier with a 30 Hz low cut filter, a 0.16 Hz high pass filter, a notch of 55–65 Hz, and a gain of 70 μV.

Offline analysis was conducted with WinEEG software (version 2.103.70, 2014). Eyeblink artifacts were corrected according to Semlitsch, Anderer, Schuster, and Presslich (1986), and all EEG and EOG segments with an amplitude over +/- 70 μV were removed from analysis.

Analysis Methods.

P300 amplitude and latency.

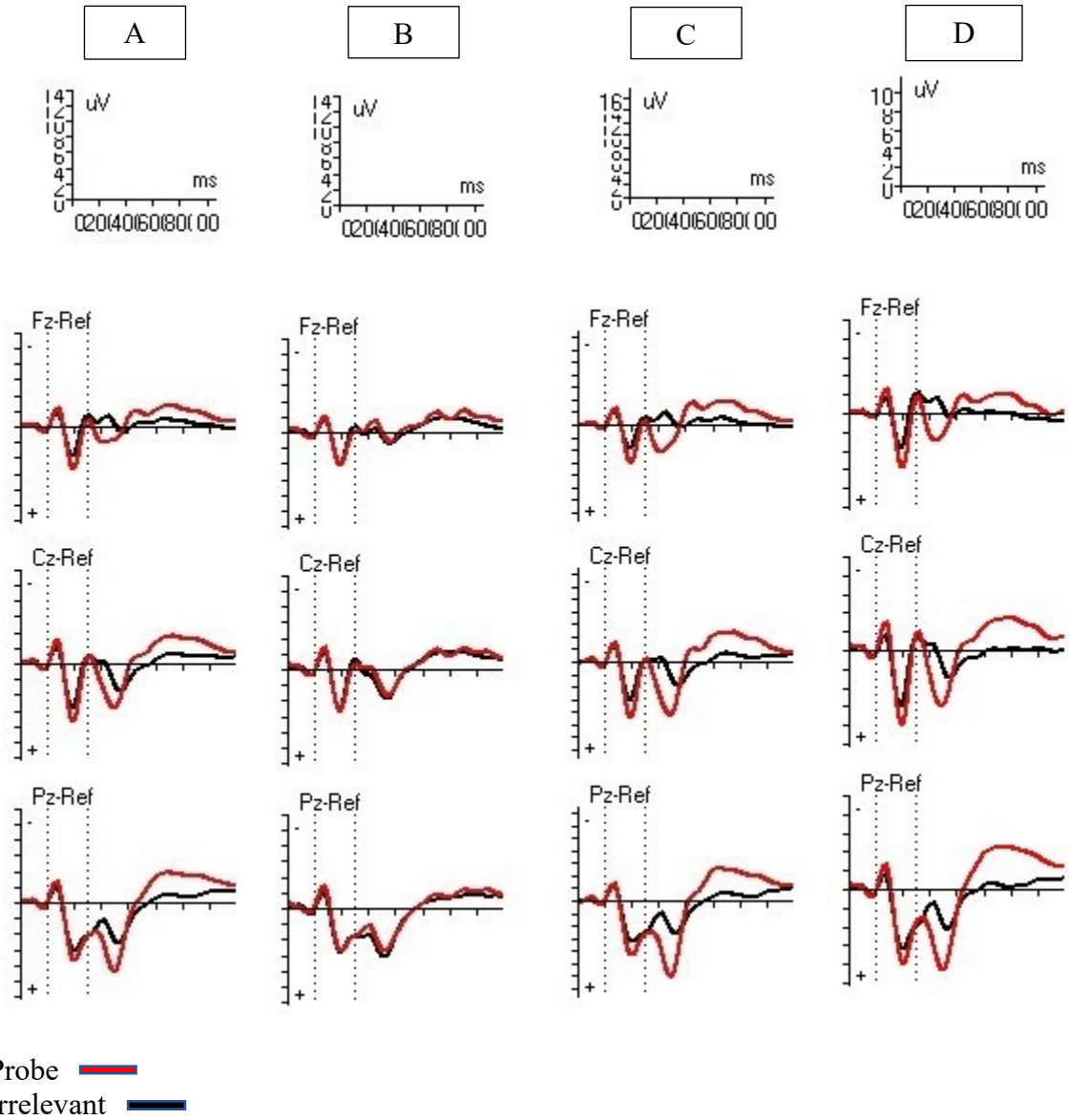
The peak-to-peak (p-p) method of analyzing P300 amplitude was used as it has been found to be superior to the base-to-peak (b-p) method (Soskins, Rosenfeld, & Niendam, 2001), and more sensitive in concealed information detection (Rosenfeld, 2011). All three sites were analyzed but final analyses were based on the Pz site since it has been found to produce the largest amplitudes (Rosenfeld, 2011) and is the most often site reported in the literature.

Grand averages (see Fig. 1) were calculated with all groups and conditions according to Keil et al. (2014). Search windows were established to be 352ms, 664ms, and 1300ms from the probe ERP. Based on the grand mean of the probe curve, T1 (352ms) was the point where the curve began its downward trajectory post-stimuli exposure[12], T2 (664ms) was determined to be the point where the curve re-intersected the X axis past the most positive segment, and T3 (1300ms) was selected arbitrarily as the point where the algorithm stopped searching for any waveform. We utilized an in-house, non-commercially available software, supplied by Rosenfeld (personal communication, May 2015), whose algorithm then searched for the mid-point of the most positive 100ms average potential segment (also called the P300 latency) in the 352–664ms

[12] The polarity for this experiment was purposely inverted with positive amplitudes below the Y axis.

look window, and then subtracted the average of the mid-point of the most negative 100ms segment amplitude found between the 664–1300ms window. The subsequent value after subtraction was defined as the P300 p-p amplitude. This program operates on the Matlab platform.

Figure 1. Grand Averages All Groups Combined (A), Innocent Control Group (B), Simply Guilty Group (C), and Guilty Countermeasure Group (D)



Group statistical analyses.

A series of repeated measures analysis of variance (ANOVA) were conducted for group analysis. Where necessary, post-tests were conducted using a Bonferroni correction.

Individual diagnostics.

The preferred statistical method for individual diagnostics in EEG-based CIT research is bootstrapping (Rosenfeld, 2011; Rosenfeld and Donchin, 2015). This technique permits the random resampling with replacement (n-1) of an EEG single sweep data distribution instead of repeating the identical test many times over with the same individual. An average amplitude can then be calculated by bootstrapping a set of P300 probe waveforms for each participant. The same procedure is then applied to a corresponding set of waveforms for irrelevant items. The irrelevant P300 amplitude mean is then subtracted from the probe mean and subjected to multitude iterations. In our case, we elected on 100 iterations as recommended by Rosenfeld, Ward, Meijer, and Yukhnenko (2017).

Three dependent variables were used here as in Lu et al., (2017), and Rosenfeld, Ward, Frigo, Drapekin, and Labovsky (2015), but for each electrode site. The first was the P300 p-p amplitude difference in microvolts (labelled as Dx) between probe and irrelevant items of participants in each condition. The second variable was the difference in averaged bootstrapped p-p values of the probe and irrelevant items. The third was the greater number, out of 100, of p-p bootstrapped iterations between probe and irrelevant items where an iteration achieved a confidence value over the 0.9 criterion. Although this criterion level is considered traditional and somewhat arbitrary (Rosenfeld, 2011), it has been used in many P300 studies from the Rosenfeld laboratory, and it has been found to be discriminating effectively at the individual level (Meixner & Rosenfeld, 2010). In other words, for a participant to be classified as *knowledgeable*, at least 90 out 100 bootstrapped P300 p-p iterations for the probe had to be greater than 0.9 than that of the p-p P300 iterations for irrelevants.

Receiver Operating Characteristic (ROC) Analysis.

We used a ROC analysis to further substantiate the CTP's effectiveness in determining knowledgeable participants from non-knowledgeable ones. The resulting values of sensitivity (the true-positive rate) and 1-specificity (the true-negative rate) from a ROC analysis is the Area Under the Curve (AUC). The AUC permits the assessment that a CIT investigator "will correctly identify the positive case when presented with a randomly chosen pair of cases in which one case

is positive and one case is negative." (Eng, 2005, p. 910). Finally, the bootstrapped iterations values also served to calculate the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) analysis.

Post-test questionnaires.

At the conclusion of the test participants were asked to complete a post-test questionnaire designed according to their condition. These questionnaires served to validate certain information such as general instructions and CM compliance, and salience of names.

Results

Between-Groups Comparisons.

P300 p-p amplitudes.

Data analysis was performed to verify normality, skewness, and kurtosis. All data were normally distributed. Outliers were analyzed both globally and within groups and none were found as all participants' data was within +3 / -3 Z score.

A series of mixed ANOVA were performed in SPSS (version 25) in relation to P300 p-p amplitudes, the dependent variable. First, a 3 groups (IC, SG and GCM) by 3 sites (Fz, Cz, and Pz) by 2 stimuli (probe and irrelevant) mixed ANOVA revealed a significant sites x stimuli interaction (F (2, 84) = 5.53, p = .006, ηp^2 = .12) (see Fig. 2). The total mean amplitude at the Pz site (M = 14.93, SD = 7.41) was significantly greater than that of sites Fz (M = 8.14, SD = 5.36) and Cz (M = 11.80, SD = 6.70). This accords with previous literature and as a result, we conducted the remainder of our analyses based on data from the Pz site only.

Figure 2. Sites by Stimuli Interaction

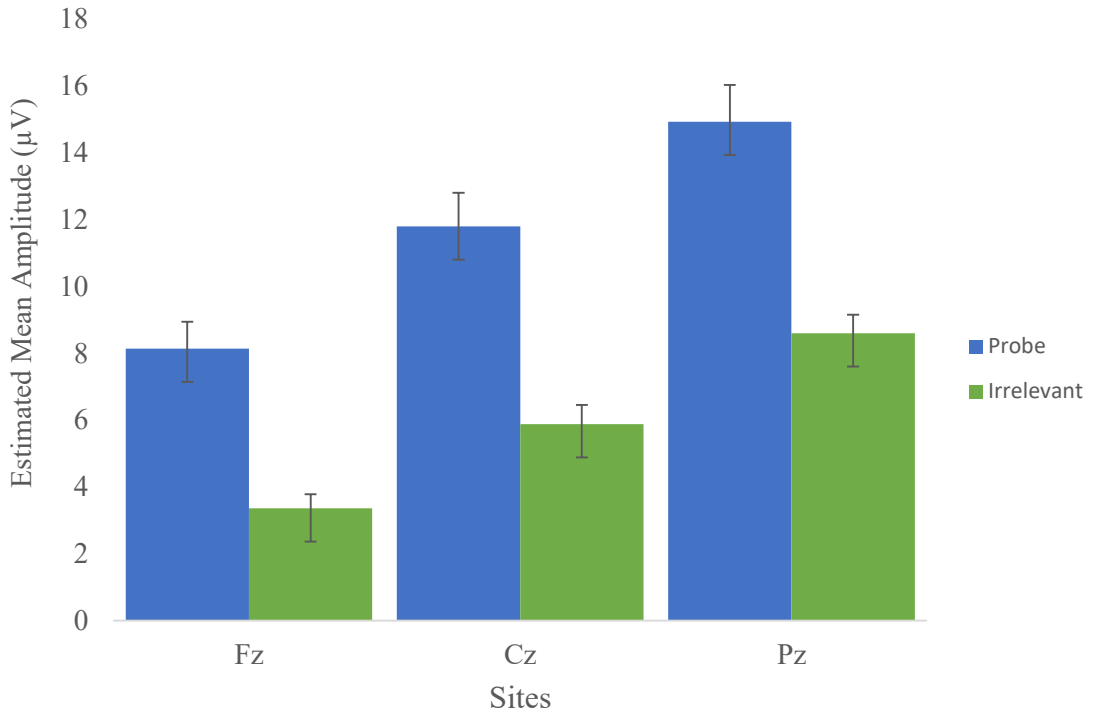


Figure 2. Mean P300 amplitudes expressed in microvolts (µV) by site for innocent, simply guilty and countermeasure groups combined.

Note. Error bars represent standard error of the mean.

Second, a stimuli-by-groups mixed ANOVA on the Pz site revealed a significant main effect of stimuli, $F(1, 42) = 92.35$, p = .000, $\eta p^2 = .69$. This main effect was qualified by a significant stimuli by group interaction, $F(2, 42) = 21.56$, $p = .000$, $\eta p^2 = .51$ (Fig. 3). Post-hoc analysis indicated that P300 values in the SG group were notably greater for the probe (M = 19.22, SD = 9.05) than for irrelevant (M = 8.85, SD = 4.32) stimuli, $p = .000$, $\eta p^2 = .69$. Similar results were obtained for the GCM group in that amplitude levels were higher for probe stimulus (M = 14.78, SD = 5.0) than irrelevant items (M = 7.20, SD = 2.81), $p = .000$, $\eta p^2 = .53$. No significant difference was observed between stimuli in the IC group ($p = .75$). While post-hoc comparisons indicated greater P300 mean amplitudes for the probe item in the SG group than in the IC group (M = 10.18, SD = 4.18, $p = .002$), probe levels failed to reach significance in the GCM group ($p = .198$). Probe amplitude difference between the SG and GCM groups were also not meaningful ($p = .200$).

Figure 3. Mean P300 Amplitudes in Microvolts by Group and by Stimulus Type

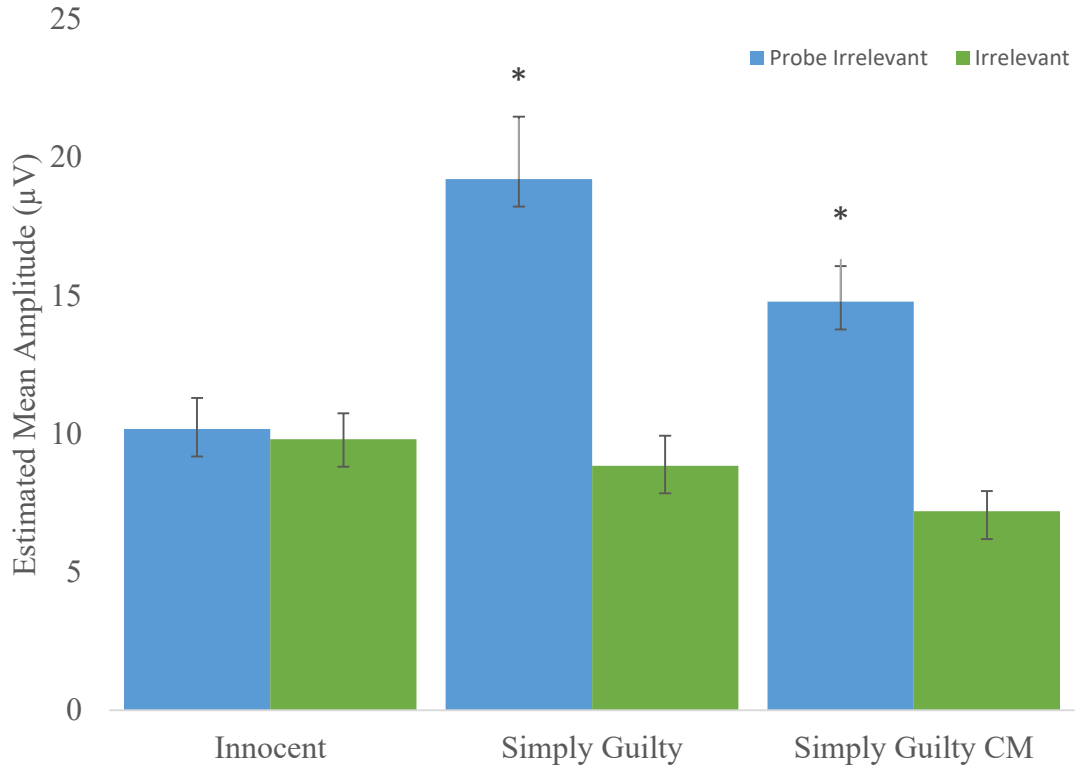


Figure 3. Mean P300 amplitudes expressed in microvolts by group for probe and irrelevant items.

Note. Error bars represent standard error.

* $p < .05$

To get a clearer view of the behaviors of the SG and the GCM group compared to the IC, we ran 3 separate repeated measure ANOVAs comparing both guilty groups to the innocent one. When the IC and GCM groups were compared, a significant Group by Stimuli interaction was found ($F(1, 27) = 47.95$, $p = .000$, $\eta p^2 = .64$). Pairwise comparisons indicated that the mean amplitude p-p difference for probes in the IC group ($M_{pr} = 10.18$, $SE_{pr} = 1.24$) and GCM group ($M_{pr} = 14.78$, $SE_{pr} = 1.19$) was significantly different ($p = .012$) as was the mean P300 p-p difference for irrelevants in the IC group ($M_{Iall} = 9.80$, $SE_{Iall} = 0.85$) and those of the GCM group ($M_{Iall} = 7.21$, $SE_{Iall} = 0.82$) ($p = .036$). When the IC and SG groups were contrasted, the mean amplitude of probes in the IC group ($M_{pr} = 10.18$, $SE_{pr} = 1.93$) was significantly lower ($p = .002$) than their counterparts in the GCM group ($M_{pr} = 19.22$, $SE_{pr} = 1.80$). The mean difference for irrelevants between the IC group ($M_{Iall} = 9.81$, $SE_{Iall} = 1.06$) and the SG group ($M_{Iall} = 8.85$, SE_{Iall}

= 0.99) did not reach significance (p = .516). When compared to the SG group, no differences emerged with the GCM.

Total errors.

A one-way between-subjects ANOVA was performed to compare error rates by group resulting in findings of a significant main effect of combined cognitive and behavioral errors $F(2, 42) = 13.32$, p = .000, ηp^2 = .39 (Fig. 4). Pairwise comparisons further confirmed that, relative to participants in the IC (M = 11.14, SD = 10.41, p = .002) and SG (M = 5.81, SD = 3.89, p = .000) groups, those in the GCM group, committed significantly more button press errors (behavioral errors) and failed memory checks (cognitive errors) (M = 28.07, SD = 18.63).

Figure 4. Mean Differences Between Groups for Total Errors

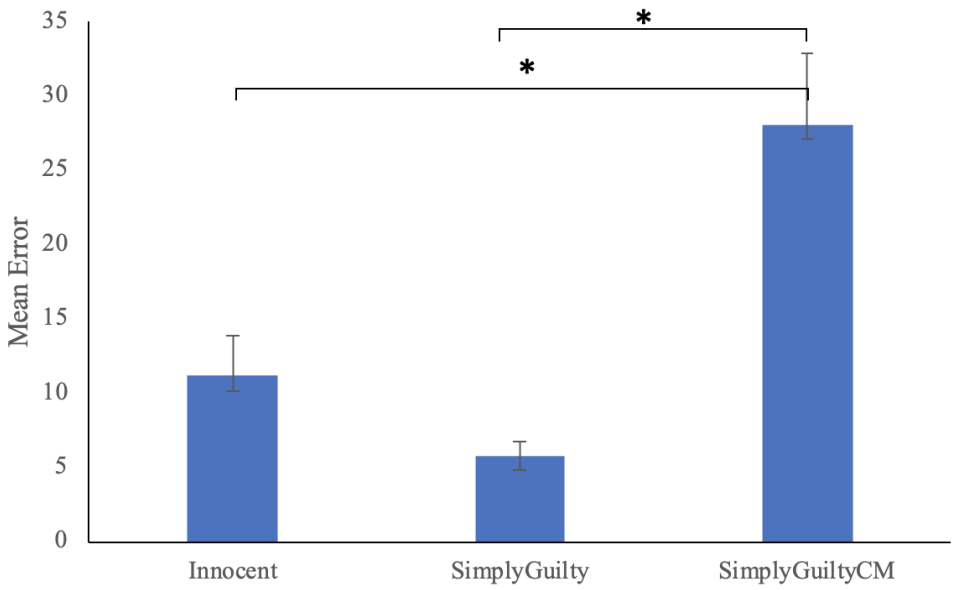


Figure 4. Mean differences between groups for total errors.

Note. Error bars represent standard error.

* $p < .05$

Reaction time.

A stimuli-by-group mixed ANOVA was executed to determine the effect of reaction time (RT) (Fig. 5 and 6). Findings from this analysis revealed a significant main effect of stimuli, (F *(1, 42)* = 7.29, p = .010, ηp^2 = .15). The mean RT for the probe (M = 514.84, SD = 124.03) was significantly longer (M_{diff} = 17.28, SE = 6.40, F (1, 42) = 7.29, p = .010, ηp^2 = .15) across groups than irrelevant (M = 496.84, SD = 126.36) items. The group by stimuli interaction narrowly missed significance level *(F (1, 42)* = 3.11, p = .055, ηp^2 = .13). However, since we had predicted longer RT with the GCM group we conducted post-hoc analyses. In effect, the GCM participants (M = 570.03, SE = 29.77) had significantly longer RT than those in the IC group (M = 448.57*,* SE = 30.81; M_{diff} = *121.46, SE* = 42.84*,* p = .021, ηp^2 = .16), but not the SG group. Furthermore, only the SG group showed a significant RT difference between the probe (M = 515.50, SE = 29.17) and irrelevant (M = 476.06, SE = 29.46) items (M_{diff} = 39.44, SE = 10.72, p = .001, ηp^2 = .24).

Figure 5. Mean Difference Between Groups for Reaction Time

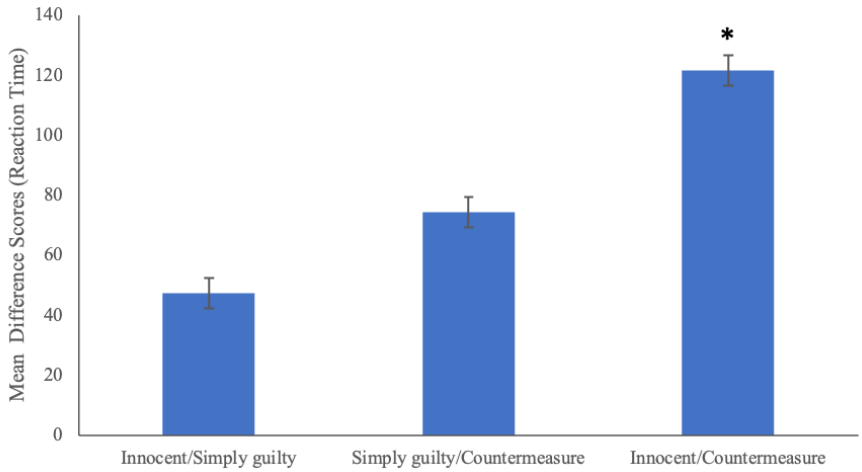


Figure 5. Mean difference between groups for reaction time.

Note. Error bars represent standard error.

* $p < .05$

Figure 6. Mean Reaction Time Between Groups for Probe and Irrelevant Items

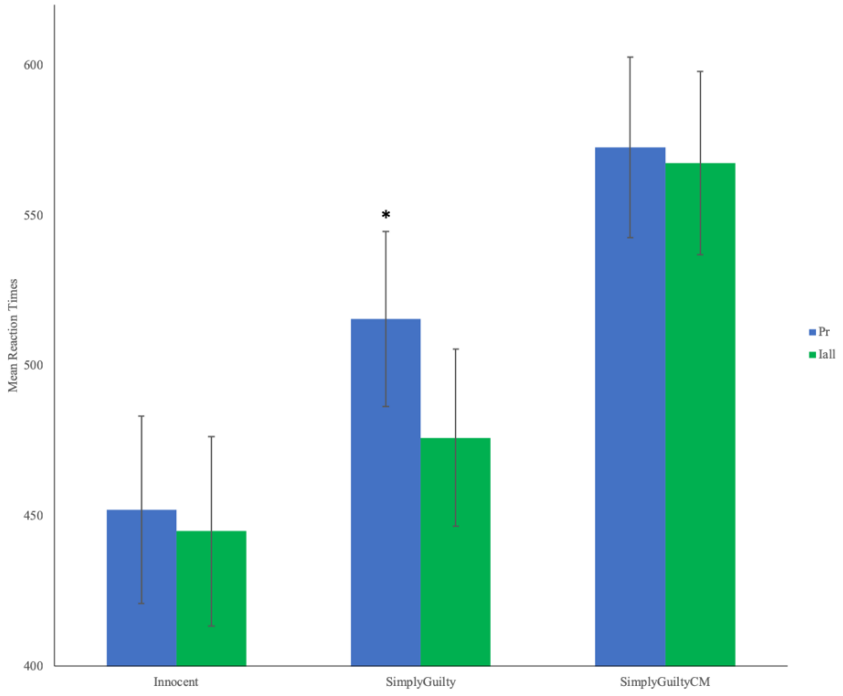


Figure 6. Mean reaction time between groups for probe and irrelevant items.

Note. Error bars represent standard error.

* $p < .05$

Individual Classification.

ROC Curves.

Receiver Operating Characteristic (ROC) analysis is a diagnostic tool used widely in various fields (e.g. medical imaging, weather forecasting, materials testing) (Gronlund, Wixted, & Mickes, 2014) and increasingly in recognition memory research (Yonelinas & Parks, 2007). According to these last authors, a ROC is the function of proportion between rates of a correct hit (i.e. accurately identifying target stimuli) and false alarm (i.e. incorrectly recognizing lure items) across a variation of conditions (2007). In terms of applied forensics Gronlund, Wixted, and Mickes (2014) argued the superior discriminability performance of ROC curves in evaluating eyewitness identification. In brief, a ROC curve plots the combination of sensitivity (true-positive rate) and specificity (1-specificity or false-positive rate) data points across a series of cutoff values (Tripepi, Jager, Dekker, & Zoccali, 2009). The overall discrimination performance output is quantified by computing the area under the curve (AUC), with values ranging from 0.5 (chance level discrimination) to 1 (perfect discrimination). "A test has (at least some) discriminatory power if the 95% confidence interval of the AUC does not include 0.50." (Tripepi, et al., 2009, p. 253). As mentioned above ROC analyses were based on the number, out of 100, of bootstrapped iterations in which the bootstrapped probe P300 exceeded the bootstrapped irrelevant P300, over 90% of the time. ROC analysis revealed sensitivity diagnostic rates at the Pz site of 100% (*AUC* = 1.000, *SE* = .000, *p* = .000, 95% CI: 1.000–1.000) for subjects in the IC group compared with both guilty groups (SG and GCM); 100% (*AUC* = 1.000, *SE* = .000, *p* = .000, 95% CI: 1.000–1.000) when contrasting the IC group to the SG group; and 100% (*AUC* = 1.000, *SE* = .000, *p* = .000, 95% CI: 1.000–1.000) for an IC to GCM group comparison, for an overall average rate of 100%. According to Meyers, Gamst, and Guarino (2006), this discriminative ability is excellent.

Bootstrapping.

In addition to ROC curves for individualized diagnostic, we utilized the bootstrapping technique as described above to predict participants' group membership. We were able to identify each participant in the IC group (14/14; 100%) as true negatives, 15/16 (93.8%) subjects in the SG group, and 14/15 (93.3%) individuals in the GCM group as true positives (Table 1) at a .9 confidence level. The individual bootstrap scores for each participant are listed in Table 2. We further calculated Grier's *A*' (Grier, 1971) for both SG and GCM groups and obtained values of

.984 and .983 respectively. Grier's parameter determines the overall discriminative ability between hit rates (true positives) and false alarms (false positives) of an instrument, in this case the CTP. A value of 1.0 represents perfect discrimination, and 0.5 corresponds to chance discrimination.

Table 1: Individual Classification Rates

	IC	SG	GCM
Correct	(14/14) 100%	(15/16) 94%	(14/15) 93%
Incorrect	(0/14) 0%	(1/16) 6%	(1/15) 7%

Table 2: Probe vs. Irrelevants Bootstrap Results at Pz for each Participants

No of Participant	P vs. Iall		
	IC	SG	GCM
1	48	100	99
2	81	82	100
3	74	100	100
4	62	100	100
5	70	100	100
6	32	100	100
7	67	100	100
8	24	100	99
9	42	100	100
10	43	100	99
11	76	100	88
12	78	100	100
13	49	100	100
14	21	100	96
15		100	100
16		100	
Mean	**55**	**99**	**99**

Post-test questionnaire.

In general, subjects from both guilty groups (SG & GCM) (n = 31) rated the burglary scenario as somewhat realistic (*M* = 2.94, *SD* = 1.41) based on a 7-point Likert scale ranging from 0 (not realistic at all) to 6 (extremely realistic). They further expressed that their family name came to mind extremely automatically (*M* = 5.61, *SD* = .56), based on a similar scale with anchors of 0 (not automatic at all) to 6 (extremely automatic). Out of a possible range of 0 (not motivated at all) to 6 (extremely motivated), their motivation to beat the test (prove their innocence) was relatively high (*M* = 4.48, *SD* = 1.41) considering that half of them (the SG group) were not instructed to follow any CM at all.

For their part, GCM participants followed the CM instructions more often than not (*M* = 4.00, *SD* = 1.25) out of a possible maximum score of 6 (always followed the instructions), they were less than somewhat confident in 'beating the test' (*M* = 2.73, *SD* = 1.16) relative to an anchor of 3 (somewhat confident), and they found the counting backward strategy very difficult based on a 7-point Likert scale with anchors of 0 (extremely easy) to 6 (extremely difficult), (*M* = 5.33, *SD* = .816).

Discussion

The CTP is an EEG-based CIT from which one can assess the memorability of encoded stimuli. Our principal objective here was to replicate and extend on the work of Lukacs et al. (2016), and by the same token, add to the very limited number of independent investigations to the body of work already done on the CTP. This experiment is the second truly unconnected study of the CTP to be done outside the Rosenfeld laboratory. Rosenfeld has collaborated as a secondary author with other researchers from other laboratories in the past in studies involving acoustic vs visual stimuli (Deng, Rosenfeld, Ward, & Labkovsky, 2016) or pictorial stimuli (Lu et al., 2017).

We initially postulated that participants in the SG group would show significantly elevated P300 amplitudes upon seeing their surnames compared with other North American last names. Our findings are in line with our first hypothesis. Simply guilty participants showed significantly greater P300 amplitudes at the Pz site than those in the innocent condition, allowing us to conclude that persons in the first group were in possession of guilty knowledge, compared to those of the second group who were unknowledgeable. These results accord with other experiments with autobiographical data (i.e. participant's surnames), conducted by Rosenfeld-led

teams (Rosenfeld et al., 2008; Rosenfeld, Ward, Drapekin, Labkovsky, & Tullman, 2017; Rosenfeld J. P., Labkovsky, Davydova, Ward, & Rosenfeld L., 2017) and other independent researchers (Lukacs et al., 2016).

Participants applying a CM also displayed meaningfully higher neural activity than innocent subjects when seeing their family name, relative to non-pertinent alternatives. We expected that GCM participants would have longer RT and that they would make more errors, thus allowing us to identify them. We note the CTP's performance in accurately identifying nearly all (14 out of 15) individuals in the GCM group despite the application of a robust mental CM. One particular posttest question dealing with the perceived level of difficulty of the CM was useful for us. Indeed, individuals in the GCM group rated the CM as very difficult (M = 5.33, SD = .816) to implement, implying that the counting backward strategy was cognitively demanding. Moreover, and relative to innocent controls, we were able to identify those same persons as they were making use of the CM in question from their longer reaction time and excessive number of total errors (behavioral and cognitive combined) thereby confirming our second hypothesis.

These results, combined with a perfect detection rate of true negatives (14/14), constitute additional evidence of the CTP's sensitivity to detect memorable stimuli, in this case a person's own family name, despite the usage of a realistic mental CM. The CM selected for this study was recommended by former polygraph examiners George Maschke and Gino Scalabrini (2005), of antipolygraph.org, and confirmed as effective in the case of polygraphic testing (Honts, Raskin, & Kircher, 1994). Furthermore, our findings represent further evidence of the underlying mechanisms involved in memory recognition and their likely independence to other motor or cognitive processes involved in the simultaneous execution of mental mathematical calculations and behavioral button presses. The CM employed here appears to have had little influence on P300 amplitude probe levels and RT. Neither the neural activity (p = .200) nor the RT (p = .241) produced significant differences between the SG and GCM groups. Either we underestimated the CM's ability to interfere in the memory recognition processes or, stated another way, the CTP is indeed resistant to this particular CM. However, one should not lose sight of the fact that some CMs tend to reduce P300 values, such as those involving a highly demanding task (Meixner, Haynes, Winograd, Brown, & Rosenfeld, 2009) or others aimed at suppressing unwanted memories (Hu, Bergstrom, Bodenhausen, & Rosenfeld, 2015). Be that as it may, the CMs that

have been tested so far have not reduced P300 levels to the point of effectively interfering with the identification of guilty individuals.

As anticipated, and as expressed by our combined guilty persons (n = 31) that their last name came to mind effortlessly, autobiographical data make for highly salient stimuli, especially family names. Participants in the SG and GCM were almost all correctly identified (29/31; 93.5%). Until further independent verification, the question remains open though with respect to whether the CTP's performance generalizes to other autobiographical data such as birthdate, mother's or father's first name, or the name of a participant's hometown or place of birth.

In a similar vein, does the ecological validity of the CTP and its excellent performance with personally rich stimuli extend to names of co-offenders and victims, when an accused person is tested? Rosenfeld, J.P., Labkovsky, Davydova, Ward, Rosenfeld, L. (2017) tested the CTP's performance with the experimenter's first name as the probe, but in a context of assessing the CTP with respect to financial incentives among malingerers. No individual data pertaining to classification was reported. In addition, the experimenter's first name (Elena) was mentioned only 2–3 times during the interaction with the participant. Although this degree of encoding of a stranger's first name is perhaps more realistic of a typical victim-offender short encounter, this kind of interaction is hardly representative of crimes involving co-conspirators. In such cases co-offenders would likely know each other from the past and have been in communication either face to face or through some other electronic means thereby likely increasing the memorability of an accomplice's first or last name. From a law enforcement perspective, it would be valuable to verify the CTP's effectiveness in these situations. Additional research is therefore necessary before claims of generalizability can be made.

This study is not without its limitations. The use of autobiographical information may not represent the most realistic types of probes, with the possible exception perhaps of cases implicating a victim's or an accomplice's surname. The results cannot be extrapolated to a real legal case. However, this study adds to the credibility of the rationale underlying the CTP as a potential and valuable investigative tool for law enforcement or national security investigators.

CHAPTER 3: EXAMINING LEVELS OF PROCESSING USING VERBAL & PICTORIAL STIMULI WITH THE COMPLEX TRIAL PROTOCOL IN A MOCK THEFT SCENARIO

Abstract

EEG-based Concealed Information Test (CIT) is a memory detection technique that is demonstrably valid and reliable with visual and verbal stimuli in laboratory experiments involving mock crimes. The performance of the Complex Trial Protocol (CTP) as a function of shallow versus deep levels of processing, during memory encoding of an event, has not been explored. Two experiments were conducted, one with verbal stimuli and the other with their pictorial referents. In both experiments participants from an undergraduate population were randomly assigned to three groups, Innocent Control (IC), Guilty Immediate Shallow Processing (GISP), and Guilty Immediate Deep Processing (GIDP). GISP and GIDP participants from both experiments underwent the same mock theft scenario and all three groups were later exposed to either a verbal (N = 41) or pictorial (N = 43) stimulus. In the word study, results showed that 14/14 (100%) of IC individuals, 2/13 (15%) of GISP persons, and 1/14 (7%) of GIDP subjects were accurately detected. In the image study, 14/14 (100%) of IC subjects, 8/14 (57%) of GISP persons, and 9/15 (60%) of GIDP individuals were correctly detected. LOP does not appear to have any bearing on CTP performance. Meaningfulness and realism of the mock crime scenario are discussed.

Keywords: Complex Trial Protocol, Concealed Information Test, P300-based CIT, Memory Detection, Levels of Processing.

Introduction

Police have many investigative tools at their disposal to collect evidence from the surface of, or within, a suspect's body such as latent fingerprints examinations, paraffin tests for gunshot residue left on the skin surface or clothing of a shooter, facial or vocal features, teeth impressions for bite marks, hair examination, DNA test on virtually any cellular component, and alcohol test. Investigators can draw incriminating or exculpatory evidence from these analyses, but all of them consist of extracting some kind of physical characteristics on the surface of, or from inside the human body. To date no instrument has been able to produce evidentiary clues, inculpatory

or exonerative, from the brain of a criminal suspect leading to his conviction or acquittal that is cost effective as well as scientifically sound. Notwithstanding the fact that such tests are cost prohibitive even with scientific and methodological advances, attempts at lie detection through fMRI remain unsuitable for field applications (Ganis, 2015) and brain imagery is vulnerable to countermeasures (Ganis, Meixner, Kievit, & Schendan, 2011). As for polygraphic results, they are mired in controversy over their validity and reliability (National Research Council, 2003). A potential candidate that has been attracting the attention of social scientists for half a century is the P300-based Concealed Information Test (CIT). First introduced in 1959 by David Lykken, this physiological technique was initially designed to identify individuals in mental possession of crime relevant information by tapping into a person's autonomic nervous system (ANS) (e.g. electrodermal activity) (Lykken, 1998). The CIT has since been expanded to collect signals directly from the central nervous system (CNS) (e.g. brainwave activity). One such method is a psychophysiological CIT protocol applicable to EEG instrumentation and better known as the Complex Trial Protocol (CTP) (Rosenfeld et al., 2008). The main characteristic of the CTP is its division of a conventional CIT into two tasks, (1) the presentation and implicit requirement of the examinee to respond to probes and neutral items with the operation of a single button press from one mouse irrespective of the stimulus shown, and (2) the presentation of attention grabbing items and the conditional button response from another mouse. The CTP is described in detail in the method section.

Notwithstanding the origin of the signal being detected, ANS or CNS, the fundamentals of a CIT remain the same; an individual is presented with two types of stimuli, criminally pertinent details, called *probes*, known only by the offender and the authorities, and plausible but neutral alternative items, called *irrelevants*. Because the probe is a relevant piece of information to the crime under investigation, and only known by its author and police, an inference of guilt could be drawn from a positive CIT by the trier of fact. The reaction generated from the rare probe presentation versus the more frequent exposure to irrelevants functions as an oddball paradigm. This involuntary cerebral manifestation is considered an index of memory recognition.

To be clear, memory detection is not synonymous to lie detection. It is not deception by omission either. In order for a mendacious statement to be considered a lie, the requisite intention to deceive must be present (Vrij, 2008). The underlying theory of an EEG-based CIT is the reflexive and non-deliberative manifestation of a cerebral component known as the P300.

However, and to an extent, the argument, that an ERP-based CIT could be used in the discovery of a duplicitous statement, could be made, but only in the context of an interviewing strategy. Consider the following example where a suspect submits to an EEG-based CIT immediately upon his detention for law breaking, and during which key details of the crime (e.g. the victim's face or name, the weapon used, unique characteristics of the crime scene, etc.) are shown. The suspect is then interviewed by police and denies knowledge of those critical details. The suspect's credibility could be challenged on the basis of significantly positive P300 CIT findings. On the other hand, an interview strategy where this sequence is not respected would not be useful since police investigators are likely to display pieces of evidence to the suspect during the course of the interview. In this scenario, a positive outcome of a P300 CIT would not be probative, as the suspect could claim later that the CIT results came from his encoding of the stimulus at the time of the interview, as opposed to processing those same stimuli during the commission of the crime. We are not aware of any memory detection test that could provide a temporal indication of when a stimulus was processed and stored, like a computer would do.

When a stimulus is perceived and then encoded into memory, a neural trace is created until it begins to degrade. The issue of memory decay as a limitation is certainly important here but it is beyond the scope of this article. While sensitive to the once ill-perceived notion that memory functions as would a camera (Schacter, 2001), it is the subsequent identification of that cerebral imprint stored somewhere in a person's cortices that is at the core of an EEG-based CIT. In particular, the P300-based CIT operates on the premise that a noticeable endogenous event related potential (ERP) occurs during an oddball paradigm sometime between 300ms to 900ms after the presentation of a rare and meaningful probe stimulus (Winograd & Rosenfeld, 2011). For instance, in a situation where the theft of a woman's purse is committed, the probe item could be the word '*purse*' and the irrelevants could be words like *laptop, wallet, phone*, etc. A significant positive ERP finding on the meaningful word '*purse*' would lead an examiner to conclude that a person is *'information-present'* from which an inference of guilt could be drawn by judicial authorities. On the other hand, an innocent person would be expected to react similarly to the word '*purse*' as well as the words *laptop, wallet, phone*, etc. since they are all meaningless to the examinee, and be deemed an *'information-absent'* person (Labkovsky & Rosenfeld, 2014). Finally, the probes and irrelevants stimuli can be presented auditorily (Deng, Rosenfeld, Ward, & Labkovsky, 2016), verbally (Meixner & Rosenfeld, 2011), or pictorially

(Rosenfeld, Ward, Thai, & Labkovsky, 2015). ERPs are also generated from other senses (i.e. smell, taste, pain) (Andreassi, 2007). Although conceivable, we are not aware of any CIT-based research on tactile, gustatory or olfactory sensory signals.

The seminal work of Craik and Lockhart (1972) and Craik and Tulving (1975) enlighten us in the way memory is processed, stored and retrieved, and thus it is helpful in our appreciation of the functioning of the P300 CIT. Through a series of experiments, they demonstrated that "the durability of the trace is a positive function of the "depth" of processing, where depth refers to greater degrees of semantic involvement" (Craik & Tulving, 1975, p. 268). Their investigation revealed two lines of evidence in relation to levels of processing (LOP): (1) stimuli that are sensorially attended only at a low level of analysis, or shallowly encoded, will result in evanescent memory traces, (2) stimuli that received complete attention, and are further enriched by images, or deeply encoded, leave a longer lasting trace. The conclusion of their findings was that deeply encoded information was associated with improved memorability on subsequent retrieval tests.

Supplementary evidence about LOP came from Zinchenko (see Craik & Lockhart, 2008; and Smirnov, 1973), Ferlazzo, Conte, and Gentilomo (1993), and Knott and Dewhurst (2009). As reported in Craik and Lockhart (2008), Zinchenko first introduced the idea in 1939 that the qualitative type of processing was crucial. He argued that intentionality to encode in addition to an orienting task focused on "meaning and comprehension" (p. 55), were necessary components of LOP to support the notion that deeper processing, through learning and integration of novel items into existing schemas, would favor enhanced memorability much like the ability of young children over adults to grasp new material. Smirnov (1973), for his part, stressed that active action on the part of the memorizer was fundamental. In other words, carrying out an actual motor action rather than reading about it would likely enhance memory performance (Craik & Lockhart, 2008). The second group of researchers later confirmed, through an 'old' versus 'new' paradigm, that recall and recognition performance of word pairs was better when their meaning was more deeply processed (tested via semantic relatedness), relative to a shallow processing task (tested via rhymical similarity). Finally, Knott and Dewhurst (2009) manipulated LOP at study and levels of attention (i.e. full versus divided) at test of words and anagrams in a Know and Remember paradigm. They concluded that remembering relies more on automatic retrieval processes while knowing relies more on controlled retrieval processes. As it applies to our

experiment, they contend that shallowly encoded items rely predominantly on familiarity-based responding, and deep encoding conditions increase recollective retrieval.

In terms of processing depth of pictorial stimuli, Marzi and Viggiano (2010) found that deeply encoded faces were recognized more accurately than shallowly encoded faces where an orientation judgement had to be made in the shallow condition and an actor/politician categorization was required in the deep encoding condition. They opined that, generally speaking, deep versus shallow encoding improved performance across a variety of encoding strategies.

We set out to verify the implications of processing levels with the CTP. Rosenfeld and colleagues have published over one dozen articles to date in line with the CTP (see Rosenfeld, Hu, Labkovsky, Meixner, & Winograd, 2013 for a review). While the CTP's performance has been tested through several research paradigms involving a variety of verbal stimuli (e.g. participant's hometown, family name, mother's first name, or investigator's first or last name) or pictorial stimuli (i.e. the image of the stolen item), none has investigated the role played by different levels of cognitive processing during the course of a mock theft scenario.

Verbal stimuli

Winograd and Rosenfeld (2011) tested their participants in a mock theft of a ring. They instructed their guilty group to attend the psychology department, to look for a certain envelope, to steal from it an item (i.e. a ring), and to return to the laboratory with the stolen item. The word "ring" was the probe and irrelevants were the words "wallet, earring, watch, locket, necklace, and bracelet". In LOP terms, we believe this paradigm would implicate a shallow level of processing since the task of taking physical possession of a ring from an envelope and submitting immediately thereafter to a P300 CIT involves little extra cognitive effort at encoding the probe stimulus, the ring. Yet, they were able to detect 10/12 (83% - sensitivity) of their guilty subjects and 11/12 (92% - specificity) of their innocent participants. An outstanding question remained, however, about the possible impact that a task involving deeper encoding of the stimulus might have had on detection rates. In a similar paradigm, Winograd and Rosenfeld (2014) verified whether prior exposure to crime details influenced detection rates, as prior knowledge of criminal details by potential suspects or innocent persons, through media leakage for instance, represents a major threat to CITs, whether they are ANS- or CNS-based. The manipulation of level of knowledge in this investigation provided us with some insight on LOP influence. Indeed,

subjects in the group Guilty-Informed (instructed to steal a ‘ring’ and did steal a ‘ring’) were all correctly identified (13/13) at a .80 bootstrap confidence level, while participants in the Guilty-Naïve group (instructed to steal an ‘item’ and did steal a ‘ring’) were detected at a rate of 79% (11/14). Applied to this study, we surmised that the Guilty-Informed group was exposed to a deeper LOP, through the prior exposure of the stimulus word ‘ring’ contained in the pre-test instructions, and that the Guilty-Naïve group constituted the shallow processing condition. The outstanding question then is whether the experimental conditions of prior-exposure and naivety really coincide with the respective LOP we attached to them.

The visual modality of a verbal stimulus was also looked at in comparison to an auditory modality of the same stimulus either presented alternatively or simultaneously (Deng, Rosenfeld, Ward, & Labkovsky, 2016). Participants were instructed to read silently the names of cities appearing on a computer monitor while the acoustic equivalent was heard through a computerized voice. The probe in this case was the respective hometown of each participant and the irrelevants were other cities (i.e. Atlanta, Buffalo, Orlando, Pittsburgh, Stockton, and Wichita). Their findings indicated that simultaneous presentation of visual and acoustic stimuli produced greater detection results in the visual only modality. In LOP terms, the shallow condition could be associated to the single alternating modality (i.e. audio or visual), and the deep processing condition could be thought of as the equivalent to the simultaneous presentation in both modalities (i.e. audio and visual). Since none of the previous studies manipulated LOP, a theoretical gap exists as to the role of processing depth of verbal stimuli on P300 CIT performance, and most importantly as it relates to the CTP.

Pictorial stimuli

A number of investigations have looked into the memorability of images at encoding and at test, and a pictorial-superiority effect has often been reported (Israel & Schacter, 1997). In fact, this particular effect dates back well over 100 years (Kirkpatrick, 1894). Galli (2014) summarized the state of the literature like this in relation to LOP: “In general, theorists agree that deep encoding results in more elaborate memory traces, and that this in turn affects later memorability.” (p. 1). On the other hand, others have found that exposure to real objects (3D) results in greater memorability than pictures (2D) (Snow, Skiba, Coleman, & Berryhill, 2014). Very few, however, have compared recognition memory of words and pictures in ERP studies. Schloerscheidt and Rugg (1997) presented their participants with 130 digitized color

photographs of common objects and 130 words made up of object names in an old/new paradigm while cerebral activity was measured at several sites including Fz, Cz, and Pz. During the study phase of pictures and words, volunteers were asked to imagine what the object would look like in real life and to state whether the object would be bigger or smaller than the computer monitor. The test took place about 5 minutes after the last study trial. Investigators found hit and false alarm (in brackets) rates of 90% (9%) for pictures and 88% (9%) for words. Again, LOP was not manipulated here and it is difficult to assess the processing depth involved in the encoding strategy employed in this study or whether detection rates might have improved under deeper processing.

With respect to the P300-based CIT and the CTP, Rosenfeld and colleagues have tested the CTP's performance with pictorial stimuli either in isolation or in contrast with words of object or family names. Labkovsky and Rosenfeld (2014) modified the original single probe CTP into a dual probe CTP and tested 47 students in an envelope-containing-item mock theft scenario. The name 'Meixner' (verbal probe) was inscribed on the envelope which contained a USB drive (pictorial probe). The encoding of the verbal probe became incidental to the commission of the mock crime. Pictorial Irrelevants were a pen, notebook, iPad, cell phone, watch, CD, computer mouse, and DVD player, and verbal equivalents were Alden, Bridges, Carswell, Gigler, Hechtman, Kamat, Martin, and Proctor. The newly configured CTP meant that Part 1 was therefore akin to the conventional four-stimuli CTP and Part 2 became a three-stimuli protocol. They tested their participants in a counterbalanced fashion such that all participants underwent both parts of the reconfigured CTP. They correctly identified 14/14 (1.0 specificity) of their innocent control participants, 11/15 (.73 sensitivity) of their simply guilty subjects in Part 1, and 13/14 (.93 specificity) of the innocent controls, 14/15 (.93 sensitivity) of the simply guilty individuals in Part 2. In Rosenfeld, Ward, Thai, and Labkovsky (2015), 12 participants were tested with a variety of probe stimuli (i.e. iPod, ring, keys, USB key, pen, and coin) on a computer screen in two counterbalanced blocks. For example, in one block the pictorial probe 'ring' was shown pictorially, and in the other block the same probe was displayed verbally (i.e. the word 'ring'). They exposed the probe on a computer screen for 30 s in the study phase and instructed the participants to carefully examine it. They removed the probe and further directed the participants to visualize it while mentally recalling its details for another 30 s. Participants did not undergo a mock theft. They were instead advised that they were suspected of having

stolen the probe in question. Their findings indicated increased values over three indexes (P300 probe minus irrelevant amplitude difference, and two bootstrap related dependent variables) favoring pictorial over verbal presentation (individual hit scores were not reported). Finally, Lu et al. (2017) explored the role of collaborative versus individual implication in the mock theft of a ring. Irrelevants were a watch, earrings, necklace, bangle, brooch and bracelet. All stimuli were displayed pictorially. From their knowledgeable individuals (guilty suspects), this group accurately detected 10/16 (63% - specificity) of those acting collaboratively with another participant and 3/16 (19% - specificity) of those acting alone using the bootstrapped probe-irrelevant amplitude difference variable, and 4/16 (25% - specificity) persons in the collaborative group, and 12/16 (75% - specificity) persons in the individual group using the variable showing the number, out of 1,000, of bootstrapped iterations greater than the accepted .9 criterion (Rosenfeld, 2011). False alarm rates were not reported in this study. Overall, depth of processing was not addressed in any of the studies outlined above.

With this in mind, we sought to verify the operability of different levels of processing on the memory encoding of participants in a mock theft scenario and the impact varying LOP may have on the CTP's performance. In keeping with Smirnov (1973), we added the feature of an action performed by our participants in the scenario to enhance memorability. Hence, regardless of their exposure to verbal or pictorial stimuli, we hypothesized that participants engaged in a deep processing task would generate higher P300 amplitudes than those involved in the shallow processing of the stimulus item (H1), and that the same relationship would occur between the shallow processing group and the innocent controls (H2).

Experiment 1 - Word

Method

Participants.

Following a power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), an estimated sample group of about 42 (approx. 14 per group) was necessary to enable the detection of an effect size of f(V) = 0.7, or approximately 0.3 Cohen's *d*, at an alpha of .05. A total of 46 (7 males) participants were recruited for this study. The mean age was 21.8 (*SD* = 4.2), ranging from 18 to 41 years old. All were undergraduate students from Concordia University's psychology department and were offered a course credit for their participation. All had normal or corrected-to-normal vision and expressed fluency in English. None reported being color blind,

nor suffering or diagnosed with a major psychological disorder (i.e. schizophrenia). The data from four participants was excluded for making too many errors. These are described further below in the procedure section. They were excluded for exceeding a threshold of 20% behavioral errors (including one that made too many cognitive miscues as well as having a handicapped right hand). A fifth participant was removed from analysis for technical reasons (electrode at A2 disconnected at approx. trial 190). This left 41 datasets for analysis.

This research was authorized by Concordia University's ethics committee (certificate #30006647). All participants signed a written consent form prior to commencing the experiment. This document clearly explained the purpose of the research, the general procedure, the risks and benefits, and the conditions of participation, which included a confidentiality commitment from the experimenters.

Research design.

We used a 3 (groups: innocent control, shallow processing, deep processing) x 3 (electrode sites: Fz, Cz, Pz) x 2 (stimuli type: probe or irrelevant) mixed-between-within-subjects factorial design for this study, where groups was the between-group variable, and sites and stimuli served as the within-subject variables.

Procedure.

Volunteer participants were first asked to read and sign a written consent and to complete a demographic data sheet. They were then randomly assigned to one of three groups, innocent control (IC) (n = 14), guilty immediate shallow processing (GISP) (n = 13), or guilty immediate deep processing (GIDP) (n = 14).

Once greeted by a research assistant, individuals in the GISP group were handed out written instructions on how to perform a mock theft. The mock theft briefing sheet read as follows. "You are to walk over to room PY-051.00. This room is located straight down the hall from the lab. Once in the room locate a Green & Beige North Face backpack. Inside the backpack is an object. Steal the object from the backpack. Leave the backpack there. Hide the item on your person and return to the lab for further instructions." The item they were to 'steal' was a silver watch. They were invited immediately thereafter to enter the laboratory for testing.

Persons assigned to the GIDP condition were given the same instructions on how to commit the mock theft, and to return to the main laboratory room for testing. However, prior to testing they were asked to read a short text and write-in the missing words (e.g. In the past few

minutes I took part of a laboratory experiment during which I stole a _____. The _____ was located in a room near the main laboratory). The expected word 'watch' was missing in 10 slots. (see Appendix D).

Candidates in the control group were not subjected to the mock theft scenario and were directly tested upon completion of the required initial documentation. The test lasted approximately 15 minutes irrespective of the condition.

Trial structure and testing procedure.

Typically, the CTP involves the presentation of four types of stimuli on a computer monitor: a *probe* (the concealed item known only to the author of a crime and the authorities), *irrelevant items* (an assortment of similar stimuli acting as fillers to the probe item), a *target item* (a string of numbers, usually 11111), and a series of four *non-target items* (a string of numbers, ordinarily from 22222, … to 55555). Following a baseline of 100ms of recorded pre-stimulus brain activity, the stimuli, regardless of their type, are always presented for 300ms at the center of the computer screen.

In the CTP, a trial consists in the presentation of two stimuli. The pair is always made up of a probe or an irrelevant item, followed by a target or non-target item, and separated by a fixation cross. The button press response from one mouse to the first stimulus is intended to confirm that the participant has implicitly seen the stimulus in question, while the conditional button presses from a second mouse in response to the second stimulus is meant to confirm the participant's explicit attention to the stimuli presentation (Rosenfeld et al., 2008).

In this experiment, investigators instructed participants to press the right button from a mouse on the left as fast as they could each time they saw a word, the first stimulus. They were also told to immediately press either the right or left button from another mouse on the right when they saw the second stimulus. If it was the target item 11111, they had to press the right button, and if it was one of any non-target items (i.e. 22222, 33333, 44444, or 55555), they had to press the left button on that same mouse. Participants who committed more than 20% of button press errors on either stimulus were excluded.[13] These miscues are called behavioral errors.

[13] This threshold is in keeping with the Rosenfeld laboratory.

Finally, the CTP investigator usually pauses the experiment periodically to seek the identity of the last stimulus seen by the participant. According to Rosenfeld et al., (2008), this step further ensures that the participant is fully attentive, cooperative and not employing any countermeasure. However, the literature is quite unclear in terms of the number of interruptions. A review indicates that the number of pop quizzes varies from a little as 7-12 trials (Rosenfeld, Ward, Frigo, Drapekin, & Labkovsky, 2015) to as high as 50-70 trials (Deng, Rosenfeld, Ward, & Labkovsky, 2016). One CTP experiment did not report its number of inquisitive pauses (Meixner & Rosenfeld, 2010), while some others remained vague about it, such as stating that the number of interrogative stops were made every few trials (Dietrich, Hu, & Rosenfeld, 2014), that participants' attention was monitored but not maintained (Labkovsky & Rosenfeld, 2014), that surprise tests were conducted at unpredictable times (Labkovsky & Rosenfeld, 2012; Rosenfeld, Ward, Drapekin, Labkovsky, & Tullman, 2017), or that the procedure was occasionally paused (Lu et al., 2017).

No finding has ever been offered to date as to the optimal number of interruptions. We took a middle of the road approach and quizzed our participants about every 20 trials, ranging from 19 to 25 trials (M = 21.5), for a total of eight pauses over 177 trials. As this step could be interpreted as a reinforcement procedure, care was taken to ensure that a pause coincided with all stimuli at least once. Since there were seven items, six irrelevants and one probe, and nine pauses, that allowed for only one item to be reinforced twice. As a precaution against a possible reinforcement of the probe stimulus, investigators chose an irrelevant item for the extra interruption. Participants were informed prior to testing that they would be questioned periodically on the last stimulus seen, and that more than two slip-ups, or a 25% error rate, would lead to their data being set aside. This type of mistake was called a cognitive error. Participants were not informed of a practice session, but investigators edited out the first 10 trials. To summarize, a total of 187 trials were presented but only 177 were kept for analysis. No real practice run exists in preparation for the application of the CTP. However, investigators normally use the first 10 trials to act as a built-in rehearsal.[14] These trials were edited out from the final analysis. The probe was presented 29 times (p = .078), irrelevants 158 times (p = .422), target 39 times (p = .104), and non-targets (p = .493).

[14] This preparatory stage is in keeping with the Rosenfeld laboratory.

Stimuli.

As in Winograd and Rosenfeld (2011) and Lukacs et al., (2016), we used one probe and six irrelevant items in our experiment. The probe stimulus was the word "WATCH", and the irrelevants were the words: "CREDIT CARDS", "IPHONE", "SUNGLASSES", "USB KEY", "CAMERA", and "MONEY". In keeping with Lykken's (1998) plausibility criterion, the irrelevant items were "equally plausible alternatives" (p. 39).

Stimuli presentation was done through PsyTask and displayed on a 55cm HP Compact (LA2206x) flat monitor with a 1280 x 1024 resolution in a dimly lit room. The average stimuli size was 6.77cm x 6.93cm. At a viewing distance of 63.5cm (measured from the participant's right eye to the center of the screen) the average stimulus subtended 5.5° x 6.1° of visual angle. The viewing distance from the participant's nasion to the fixation cross at the center of the monitor was 61cm. All items were presented in Times New Roman, 96 font, Black on a White background surrounded by a wide Black edge. The inter-stimulus interval varied from 1600ms to 2700ms. The first interval refers to the time between the presentations of the probe or irrelevant item to the next. The second interval consists of the time from the exposure of a target to a non-target item.

EEG data acquisition.

EEG data was recorded with a Mitsar amplifier, model 201 (Mitsar company, St-Petersburg, Russia) sampling at 500 Hz, and seven conductive gel filled Ag/AgCl electrodes. The ground electrode was placed on the forehead above the corrugator muscles. The electrooculogram (EOG) electrode was placed approximately one cm above the center of the left eyebrow. Three electrodes were attached to the scalp midline at sites Fz, Cz, Pz and referenced to linked mastoids. In accordance with the International 10-20 system, and prior to being tested, the distance between the inion and the nasion for each participant was measured such that the Cz electrode was consistently placed at the 50% mark on the scalp. Participants were asked to refrain from making head and upper torso movements, speaking or fidgeting in their seat, and to keep their feet flat on the floor during the test. Impedance between the scalp and electrodes was kept at below 5 KΩ. Signals were passed through the amplifier with a 30 Hz low cut filter, a 0.16 Hz high pass filter, a notch of 55-65 Hz, and a gain of 70 μV.

Offline analysis was conducted with WinEEG software (version 2.103.70, 2014). Eyeblink artifacts were corrected according to Semlitsch, Anderer, Schuster, and Presslich (1986), and all EEG and EOG segments with an amplitude over +/- 70 μV were removed from analysis.

Analysis Methods

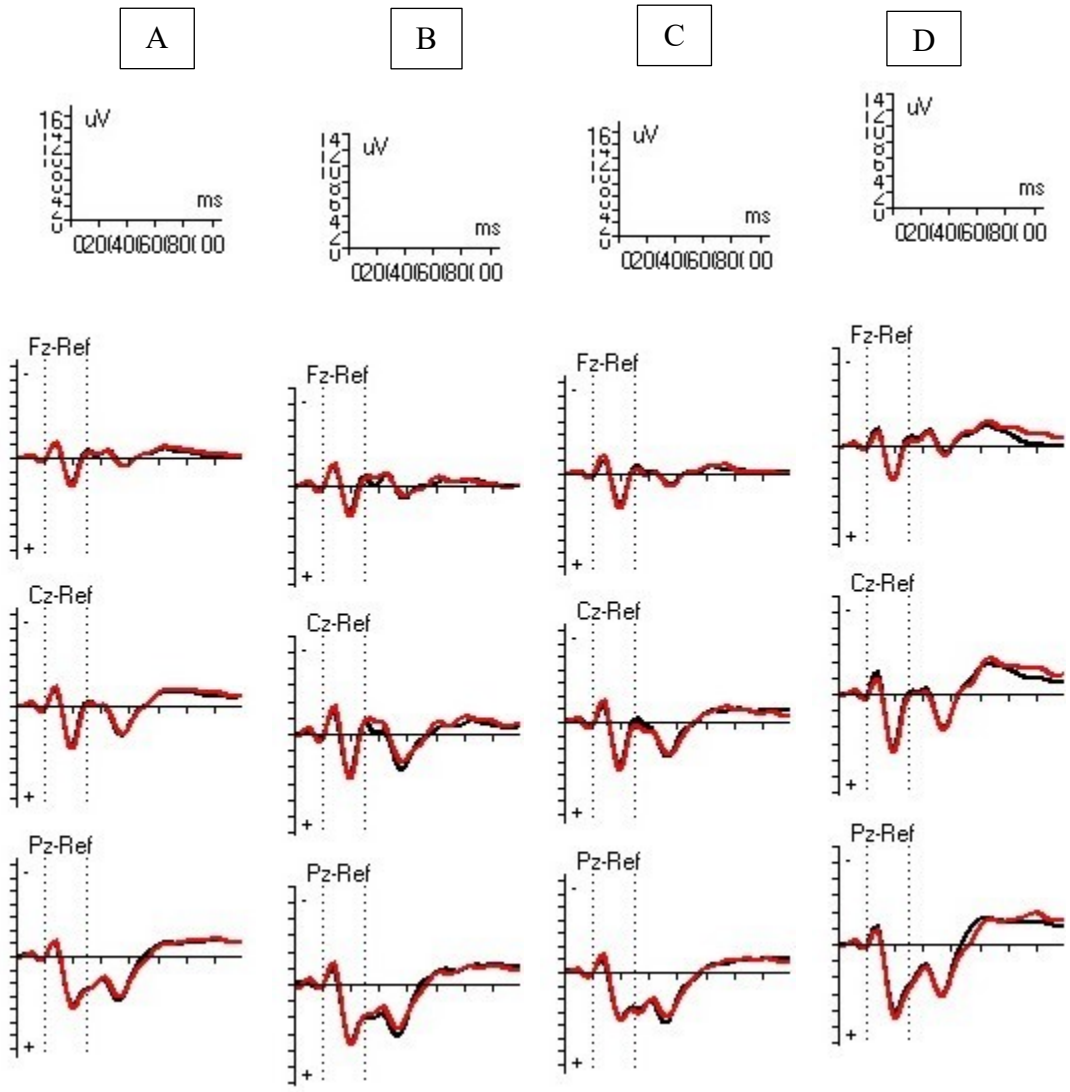
P300 amplitude and latency.

Investigators used the peak-to-peak method of analyzing P300 amplitude has it has been found to be superior to the base-to-peak method (Soskins, Rosenfeld, & Niendam, 2001), and more sensitive in concealed information detection (Rosenfeld, 2011). All three sites were analyzed, but our final analysis was based on data from the Pz site as it has been found to produce the largest amplitudes (Rosenfeld, 2011) and it is the site most often reported in the literature.

Grand averages (see Fig. 7) were calculated with all groups and conditions according to Keil et al. (2014), and search windows were established to be 408ms, 734ms, and 1300ms from the probe curve. The search parameters were established using an objective method such that T1 (408ms) was established as the point where the curve begins its downward trajectory[15], T2 (734ms) was determined to be the point where the curve intersected the X axis past the most downward point, and T3 (1300ms) was selected arbitrarily as the point where the algorithm stopped searching for any waveform. We used a non-commercially available Matlab compatible software, supplied by Rosenfeld (personal communication, May 2015), to identify the most positive and negative peaks. The algorithm then searched for the mid-point of the most positive 100ms average potential segment (also called the P300 latency) in the 408-734ms look window, and then subtracted the average of the mid-point of the most negative 100ms segment amplitude found between the 734-1300ms window. The subsequent value after subtraction was defined as the P300 peak-to-peak (p-p) amplitude.

[15] The polarity for this experiment was purposely inverted with positive amplitudes below the Y axis.

Figure 7. Grand Averages – Word - All Groups Combined (A), Innocent Control (B), Guilty Immediate Shallow Processing (C), and Guilty Immediate Deep Processing (D)



Three dependent variables were used here as in Lu et al. (2017), and Rosenfeld, Ward, Frigo, Drapekin, and Labovsky (2015), but for each electrode site. First was the p-p P300 amplitude difference in microvolts (labelled as Dx) between probe and irrelevant items of participants in each condition. The second dependent measure was the difference in means of the iterated bootstrapped average p-p P300s for probe and irrelevant items. The third was the greater number, out of 100, of p-p bootstrapped iterations between probe and irrelevant items where an iteration achieved a confidence value over the .9 criterion. Although this criterion level is considered traditional and somewhat arbitrary (Rosenfeld, 2011), it has been used in many P300 studies from the Rosenfeld laboratory, and it has been found to be discriminating effectively at the individual level (Meixner and Rosenfeld, 2009). In other words, for a participant to be classified as *knowledgeable*, at least 90 out of 100 bootstrapped p-p P300 iterations for the probe had to be greater than 0.9 than that of the p-p P300 iterations for irrelevants.

Group statistical analyses.

A series of repeated measures analysis of variance (ANOVA) were conducted for group analysis. Where necessary, post-tests were conducted using a Bonferroni correction.

Individual diagnostics.

The preferred statistical measure for individual diagnostics in EEG concealed information test (CIT) research is bootstrapping (Rosenfeld, 2011; Rosenfeld & Donchin, 2015). This technique permits the random resampling with replacement (n-1) of an EEG single sweep data distribution instead of repeating the test many more times within an individual. An average amplitude can then be calculated by bootstrapping a set of P300 probe waveforms for each participant. The same procedure is then applied to a corresponding set of waveforms for irrelevant items. The irrelevant P300 amplitude mean is then subtracted from the probe mean and subjected to multitude iterations. In our case, we elected on 100 iterations as recommended by Rosenfeld, Ward, Meijer, and Yukhnenko (2017).

In order to reliably identify knowledgeable participants, one must answer the question, proposed by Rosenfeld and Donchin (2015), whether "the finding that 90% or more of these bootstrapped] Probe-Irrelevant [P300 differences were greater than zero" (p. 970). Posed differently, the question that the bootstrap method answers is the following: "Is the probability more than 90 in 100 that the true difference between the average probe P300 and the average irrelevant P300 is greater than zero?" (Rosenfeld et al., 2008, p. 909). A non-commercial

computer program (private communication with Rosenfeld, 2015) draws at random with replacement, a set of n1 probe waveforms and a set of n2 with replacement irrelevant waveforms. It then averages these and calculates P300 amplitudes from this single average. The calculated irrelevant mean P300 is then subtracted from the comparable probe value to produce a difference value. This process is repeated 100 times. One thus obtains 100 values to place in a distribution. To state with 90% confidence, the accepted criterion used in earlier studies by Rosenfeld (2011) and others (Lu et al., 2017; and Lukacs et al., 2016), that probe and irrelevant evoked ERPs are significantly different, we require that the value of zero difference or less (a negative difference) not be > -1.29 *SD* below the mean distribution of differences. In a one-tail distribution, a 1.29 criterion yields a $p < .1$ confidence level. The null hypothesis that the probe evoked P300 is greater than the irrelevant evoked P300 is then rejected if the two are not significantly different or if the irrelevant P300 is found to be larger.

Receiver Operating Characteristic (ROC) Analysis.

We used a ROC analysis to further substantiate the CTP's effectiveness in determining knowledgeable participants from non-knowledgeable ones. The resulting values of sensitivity (the true-positive rate) and specificity (the true-negative rate) from a ROC analysis is the AUC. The AUC permits the assessment that a CIT investigator "will correctly identify the positive case when presented with a randomly chosen pair of cases in which one case is positive and one case is negative." (Eng, 2005, p. 910).

Results

Between-Groups Comparisons.

P300 p-p amplitudes.

The data was verified for normality, skewness, and kurtosis. With the exception of one outlier, all other participants were within +3 / -3 Z score. The original value for the amplitude level at FzIallDx for participant 121 was 11.93 μV. The solution for dealing with the outlier was to seek the next highest valid value, in this case 8.18 μV, add 1.00 to it, and replace it with the new value of 9.18 μV.

The first step of our data analysis was to conduct a mixed repeated measure ANOVA with SPSS (version 25) to determine the site that presented the best results. We found a main effect of sites (F (2, 76) = 83.19, p = .000, ηp^2 = .686). As expected, the Pz site produced mean amplitude values (M = 10.36 μV, SE = 0.592) significantly higher than Fz (M = 4.61 μV, SE = 0.297) and

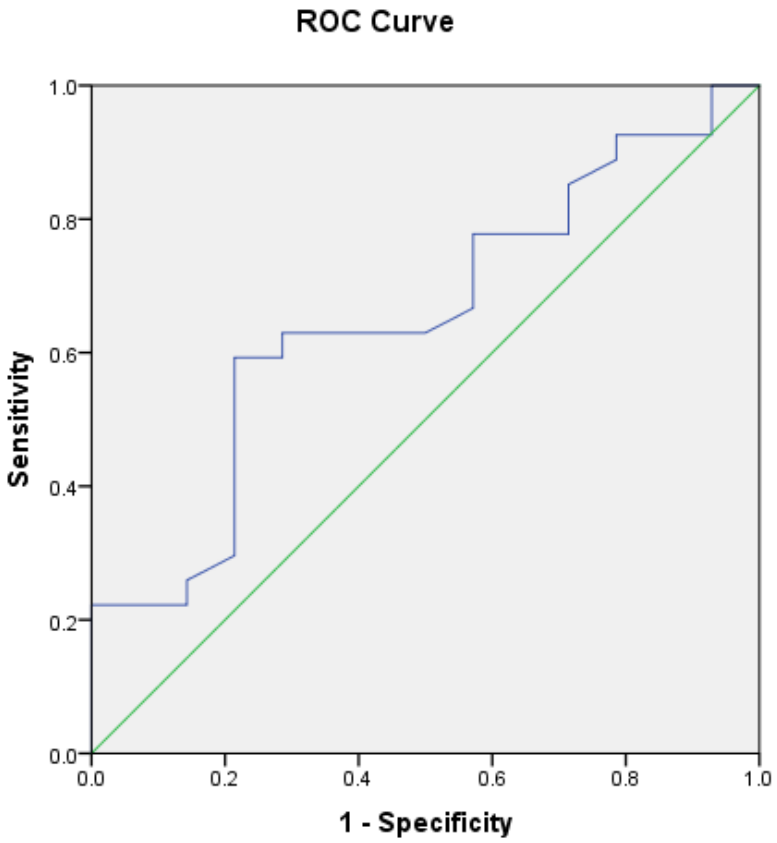
Cz (*M* = 8.32 µV, *SE* = 0.491). Accordingly, we continued our data analysis based on the Pz site for this experiment (study 1) as well as for study 2. However, a further analysis at the Pz site did not produce any significant findings in study 1.

Individual Classification.

ROC Curves.

None of the results produced significant findings in detection rates of IC versus GISP groups (*AUC* = .624, *SE* = .112, *p* = .275, 95% CI: .404-.843) or the IC versus GIDP groups (*AUC* = .679, *SE* = .103, *p* = .108, 95% CI: .476-.881). Figure 8 represents the ROC findings of the IC group compared to both GISP and GIDP groups combined (*AUC* = .652, *SE* = .089, *p* = .114, 95% CI: .477-.827).

Figure 8. ROC Curve Word – GISP & GIDP Combined



Bootstrapping.

The bootstrapping analysis described above produced the following outcomes. We identified all of our participants in the IC group (14/14, 100%). However, we only selected accurately 2/13 (15%) of our GISP subjects and only 1/14 (7%) of our GIDP individuals. (Table 3). The individual bootstrap scores for each participant can be found at Table 4.

Table 3: Individual Classification Rates – Verbal and Pictorial Stimuli

	Modality	IC	GISP	GIDP
Correct	Verbal	(14/14) 100%	(2/13) 15%	(1/14) 7%
	Pictorial	(14/14) 100%	(6/14) 43%	(9/15) 60%
Incorrect	Verbal	(0/14) 0%	(11/13) 85%	(13/14) 93%
	Pictorial	(0/14) 0%	(8/14) 57%	(6/15) 40%

Table 4: Probe vs. Irrelevants Bootstrap Results at Pz for each Participants – Verbal vs Pictorial

No of Participant	Verbal P vs. Iall			Pictorial P vs. Iall		
	IC	GISP	GIDP	IC	GISP	GIDP
1	33	77	35	73	100	100
2	77	49	24	55	100	100
3	17	88	72	62	72	100
4	53	27	98	63	54	100
5	17	100	75	65	95	98
6	42	8	73	37	94	100
7	39	76	39	29	100	88
8	4	29	38	72	100	85
9	84	37	60	49	8	87
10	33	92	87	63	80	90
11	43	71	65	16	71	99
12	43	15	85	47	41	84
13	24	81	56	50	83	100
14	83		18	52	29	75
15						35
Mean	**42**	**58**	**59**	**52**	**73**	**89**

Discussion

We had expected to see the P300 amplitude levels of the deep processing group to be significantly higher than those from the shallow and the latter meaningfully higher than the innocent group. Neither hypothesis was confirmed. In fact, our detection rates were unacceptably low, and this outcome was not anticipated. On the other hand, all innocent controls were correctly identified as true negatives (14/14), indicating a high level of specificity (100%). But our poor sensitivity detection rates demand that we turn our attention to possible explanations.

Our review began with our scenario and its realistic qualities. The experiment of reference for our investigation was Winograd and Rosenfeld (2011). Their mock theft scenario consisted of asking their participants (in the guilty conditions) to bring a manila envelope to the office of the Psychology department and enquire with the secretaries as to the location of Dr. Rosenfeld's mailbox. Having located the mailbox, they were further instructed to look for a matching manila envelope labeled in Dr. Rosenfeld's name, to surreptitiously steal an item from inside that envelope, and to return to the lab with the stolen item. The participants were also informed of the secretaries' (also lab confederates) naïveté about the study and to do their best not to get caught. They were to have the secretaries contact the lab in the event they were discovered. In contrast, we asked our guilty participants to simply walk over to a nearby room outside the laboratory, locate a backpack once inside the room, find the only object inside the backpack, steal it, leave the backpack there, hide the object on their person, and to return to the lab for further instructions. The experimental condition described above in Winograd and Rosenfeld (2011) resembled our shallow processing condition. Our deep processing condition included a missing word text exercise where participants were expected to fill 10 blanks spots spread across 12 sentences with the word 'WATCH'. The word 'stole' in one of the sentences was the only direct association of criminality to the experimental task. The other sentences were either descriptive in nature (i.e. The _____ is a man's _____. The make of the _____ is Seiko. The back face of the _____ is Blue. The overall colour of the _____ and bracelet is Silver.) or meant to enhance the semantic significance of the item (i.e. A _____ is a mechanical instrument designed to tell time. This particular _____ indicates the time as well as the date and day of the week). There was no indication of proprietorship in relation to the backpack nor to the room in question.

Compared to Winograd and Rosenfeld (2011), our scenario lacked realism and may not have elicited much arousal among our guilty participants. Greater amygdala activation has been found to correlate positively with memory performance (Canli, Zhao, Brewer, Gabrieli, & Cahill, 2000). The minimal level of arousal experienced by our guilty subjects during the mock crime could have had the cascaded effect of poor encoding. As in a domino effect, insufficient encoding, especially for our shallow processing group, may account for the lackluster detection efficiency. Peth, Vossel, and Gamer (2012) manipulated the level of stress during mock-crime execution and concluded that "emotional arousal might facilitate the detection of concealed information sometime after the crime" (p. 381). Furthermore, Klein Selle, Verschuere, Kindt, Meijer, and Ben-Shakhar (2017) suggested that emotional arousal may enhance detection efficiency with the SCR measure. It bears reminding that probe P300 amplitudes steadily increased from 9.39 μV (IC group), to 10.71 μV (GISP group), to 11.48 μV (GIDP group), while irrelevants amplitudes remained relatively steady at 9.86 μV (IC group), to 10.04 μV (GISP group), to 10.69 μV (GIDP group). The direction of these curves is consistent with our expectations. Mertens, Allen, Culp, and Crawford (2003) also experienced surprisingly low rates of classification accuracy for guilty subjects (47% using a bayesian method and 27% using a bootstrap method), leading them to conclude that "Even the recent and salient knowledge of facts from a realistic mock crime does not ensure guilty individuals will be correctly classified." (p. 40).

Additionally, we did not pre-test the stimuli, as it seemed counterintuitive to a real-life theft, despite the suggestion by Honts, Raskin, and Kircher (2002) in the context of a polygraphic-based CIT. Carmel, Dayan, Naveh, Raveh, and Ben-Shakhar (2003) manipulated the type of mock crime, by comparing a standard condition with a more realistic mock theft while measuring skin conductance. In the standard condition, participants were informed of the crime details in advance and if they had trouble remembering any of the details, they were reminded of them. In the realistic condition, participants were instructed to steal a CD-ROM from an office but none of the other details were raised. Surprisingly, the realistic mock crime resulted in weaker detection efficiency than the standard condition. These discrepancies in the pre-test confirmatory level of stimuli memorability in our mock crime may have impacted the encoding quality of the probe watch, especially for the shallow group. Be that as it may, in real life a CIT examiner would not know for certain if a crime suspect would have paid attention to the probe

stimulus, perceived it, and stored it into memory for later retrieval, and even less likely would the examiner prime a crime suspect with the stimuli. Hence, the kind of incidental exposure to the probe stimulus our participants in the shallow condition went through may represent a limitation of the CTP. Its performance in identifying significant P300 probe-irrelevants differences in mock crime scenarios may be restricted to those situations where pre-test memory confirmation (priming), optimal arousal, and realistic conditions are met. While this may amount to a fixable methodological problem for researchers in laboratories, ecological tests may have to contend with crime situations that do not necessarily come along with all these pristine testing conditions. For instance, crimes charged with high emotionality (i.e. robbery, violent assault, homicide) are likely to produce the necessary arousal if a P300-based CIT was conducted on a suspect, victim or witness. But this may not be the case with a host of other less arousing crimes (i.e. fraud, theft, possession of stolen goods) or more trivial offences (i.e. mischief to property, disturbing the peace). In other words, the CTP may not be suitable for real life testing of all types of criminal infractions.

Another less explored limitation of the CTP would be its detection sensibility with individuals whose personality make up is characterized with low anxiety levels. Visu-Petra, G., Miclea, and Visu-Petra, L. (2012) failed to find any significant associations with state-trait anxiety levels and deception in a reaction time-based CIT involving a realistic mock theft scenario. The combination of a minimally anxious person involved in a low arousing crime might represent a challenge for the CTP.

The final issue of concern was the significance our probe stimulus embodied for our guilty participants. What happens when the selected probe stimulus used by a CTP examiner is less meaningful for the examinee? Put differently, the question to be answered is whether incidental encoding of a minimally meaningful stimulus is sufficiently detectable by the CTP. Meixner and Rosenfeld (2014) addressed this issue by displaying to their knowledgeable group subjects (equivalent to our guilty groups) verbal items that corresponded to events captured by a body camera worn by participants while they went about their daily activities. Their control group consisted of a nonknowledgeable group of persons who were shown verbal stimuli unrelated to events from their video footage (analogous to our innocent control group). They successfully discriminated all subjects from both groups. But memorability was not expressly captured or manipulated in their study. It is well documented that an OR occurs upon the rare presentation of

a meaningful piece of information among a random series of frequently presented meaningless pieces of information of the same category (Rosenfeld, 2011; klein Selle, Verschuere, & Ben-Shakhar, 2018). Skinner and Price (2019) report that "increased meaningfulness of to-be-remembered information can directly impact memory for that information." (p. 1). Prior research by van Hooff and Golden (2002) demonstrated that ERP-based memory assessment may "not be sensitive enough to detect memories for words that may have formed weak memory traces." (p. 20). Meaningfulness as a cognitive construct has a considerable history beginning with Noble's (1952) attempt at quantifying its elements into a meaning (*m*) index, a linear function of the number of S_x (verbal stimuli) multiple $R_1, R_2, R_3 \ldots R_n$ (conditioned responses) connections which are formed, to Craik and Lockhart's (1972) more recent LOP approach. Nevertheless, repetition may not be a sufficient condition to guarantee acquired meaningfulness (Noble, 1952) and subsequent recognition (van Hooff & Golden, 2002). This is highlighted by responses to a post-test question indicating that 12% (3/26) of our guilty participants reported not having recognized the probe at all during the CTP. It is then likely that the memory trace for the probe was weak as well in the remainder of the subjects.

Experiment 2 - Image

Method

Experiment 2 with pictorial stimuli was conducted in the same manner as experiment 1. We addressed below only the methodological sections that differed from those applicable to experiment 1.

Participants.

A power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) produced an estimated sample group of about 42 (approx. 14 per group) to enable the detection of an effect size of f(V) = 0.7, or approximately 0.3 Cohen's *d*, at an alpha of .05. A total of 51 (4 males) participants were recruited for this study. The mean age was 22.0 (*SD* = 5.7), ranging from 18 to 52 years old. All were undergraduate students from Concordia University's psychology department and were offered a course credit for their participation. All had normal or corrected-to-normal vision and expressed fluency in English. None reported being color blind, nor suffering or diagnosed with a major psychological disorder (i.e. schizophrenia). The data from six participants was excluded for making too many errors either behavioral or cognitive, another

was removed on suspicion of drug use, and a participant was not tested because of a hair extension that made it impossible to apply electrodes. This left 43 datasets for analysis.

Procedure.

Volunteer participants were first asked to read and sign a written consent and to complete a demographic data sheet. They were then randomly assigned to one of three groups, innocent control (IC) (n = 14), guilty immediate shallow processing (GISP) (n = 14), or guilty immediate deep processing (GIDP) (n = 15).

Participants were provided with the same briefing for the mock crime as the subjects in Experiment 1. The only difference is in the experimental manipulation of the deep encoding strategy for the GIDP condition. Prior to testing, participants were asked to complete a short questionnaire made up of five questions: 1) Is this a man's watch or a woman's watch? 2) What is the make of the watch? 3) What is the color of the watch and bracelet? 4) What time is displayed on the watch? and 5) What date is displayed on the watch? The objective of these questions was to force the participant to examine the watch more closely, pay more attention to its details while simultaneously inducing a deeper level of memorability processing.

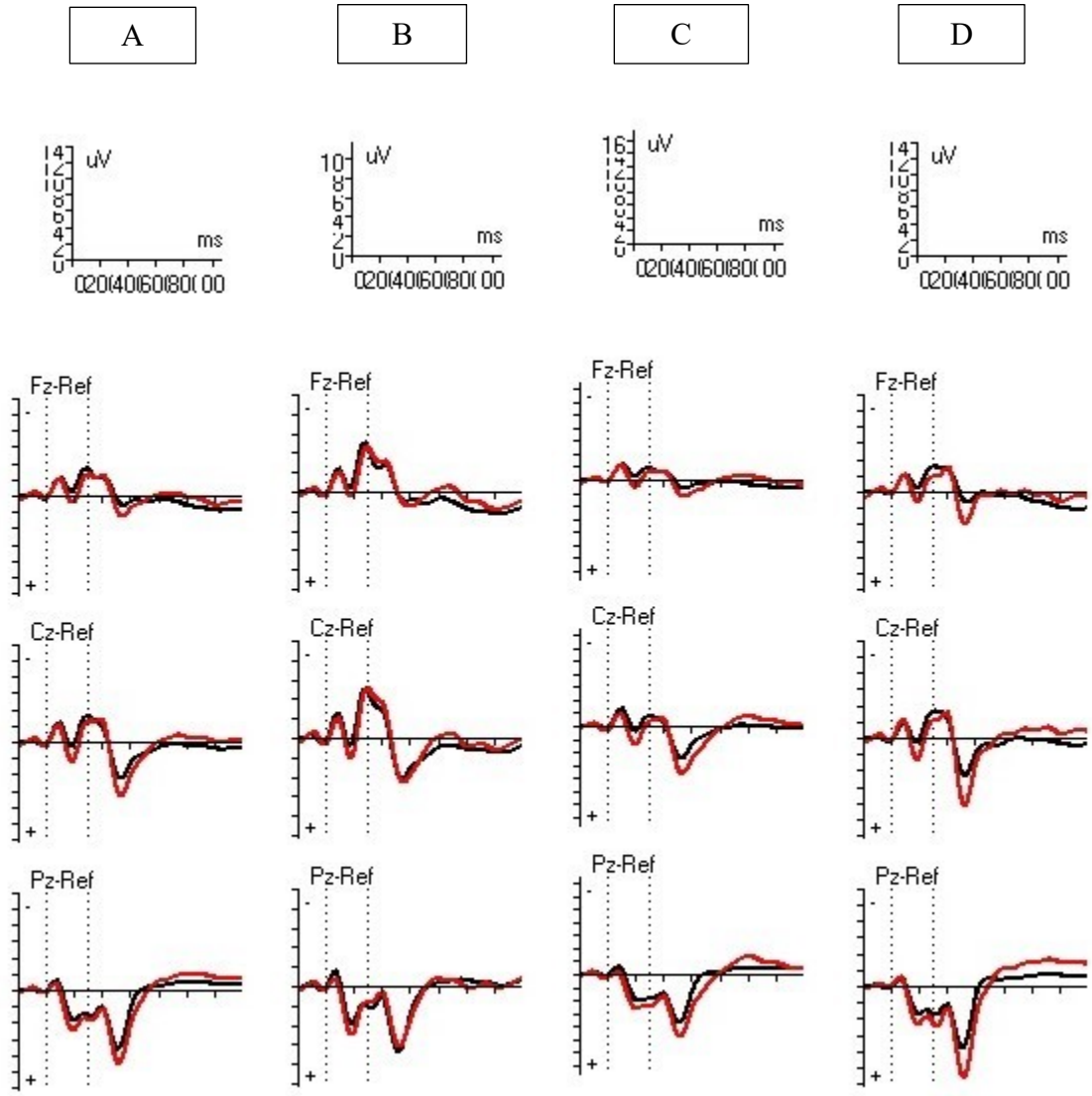
Stimuli.

The stimuli, probe and irrelevants alike, used in experiment 2 were the pictorial equivalents of the stimuli used in experiment 1 and presented in the same fashion.

Search windows.

The search windows for this study were established to be 394ms, 720ms, and 1300ms from the probe curve. The grand averages for each respective group are found in Figure 9.

Figure 9. Grand Averages – Picture - All Groups Combined (A), Innocent Control (B), Guilty Immediate Shallow Processing (C), and Guilty Immediate Deep Processing (D)



Results

Between-Groups Comparisons.

P300 p-p amplitudes.

The data was verified for normality, skewness, and kurtosis. With the exception of one outlier, all other participants were within +3 / -3 Z score. The original value for the amplitude level at FzPrDx for participant 10 was 14.35 μV. The solution for dealing with the outlier was to seek the next highest valid value, in this case 9.54 μV, add 1.00 to it, and replace it with the new value of 10.54 μV.

Given our initial finding in experiment 1 of a main effect of site, the first step of our data analysis was to conduct a mixed repeated measure ANOVA with SPSS (version 25) at site Pz to determine the presence of main effects and interactions. We found a significant main effect of stimuli (F (1, 40) = 30.34, p = .000, ηp^2 = .431), and one significant interaction; stimuli (probe & irrelevant) x group_new (IC, GISP, and GIDP) (F (2, 40) = 6.92, p = .003, ηp^2 = .257).

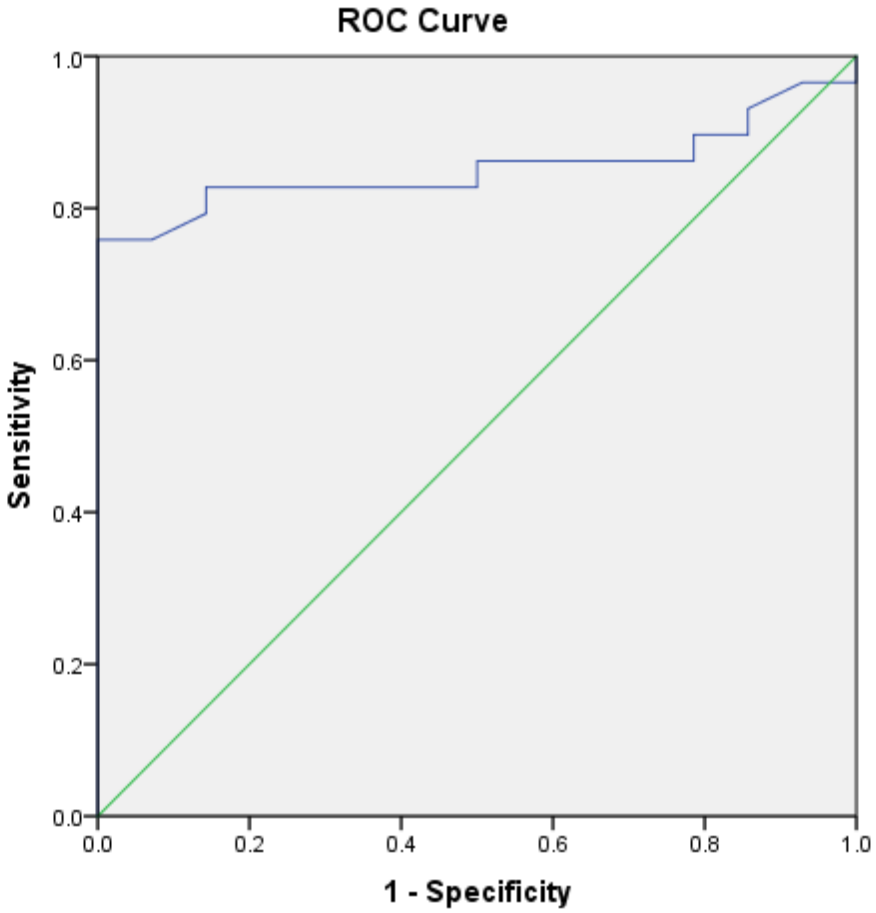
Post hoc analyses revealed a few significant relationships. The P300 probe values in the GISP (M = 12.58 μV, SE = 1.45) were significantly higher (p = .000) than irrelevants values (M = 9.01 μV, SE = 1.06) resulting in a significant mean amplitude difference (M_{diff} = 3.566 μV, SE = 0.899). Similarly, probe amplitude levels (M = 14.88 μV, SE = 1.40) in the GIDP were significantly greater than irrelevants amplitudes (M = 10.17 μV, SE = 1.02) with a significant mean amplitude difference (M_{diff} = 4.708 μV, SE = 0.869). Finally, probe amplitude values (M = 14.88 μV, SE = 1.40) in the GIDP group were meaningfully higher (p = .025) than those of the IC group (M = 9.29 μV, SE = 1.45) with a significant mean amplitude difference (M_{diff} = 5.594 μV, SE = 2.02). All other pairwise comparisons failed to reach significance levels.

Individual Classification.

ROC Curves.

Participants from the IC and GISP groups were marginally identified at 76% (AUC = .755, SE = .102, p = .022, 95% CI: .556-.955) while those from the IC and GIDP groups were accurately classified at 94% (AUC = .943, SE = .056, p = .000, 95% CI: .833-1.000). Figure 10 displays the AUC results of both groups, GISP and GIDP combined (AUC = .852, SE = .060, p = .000, 95% CI: .734-.970), relative to the IC group.

Figure 10. ROC Curve Picture – GISP & GIDP Combined



Bootstrapping.

We were able to accurately identify 100% (14/14) of IC participants, 43% (6/14) of the GISP subjects, and 60% (9/15) of the GIDP persons (see Table 3). The individual bootstrap scores are found at Table 4.

Discussion

We had made two predictions at the outset, that irrespective of the stimuli presentation modality, participants in the GIDP group would have elevated P300 amplitudes compared to the GISP group, and that the same relationship would be found between the GISP and IC groups. Dealing specifically with the pictorial modality in experiment 2, neither of our hypotheses were supported. Instead our findings pointed to a significant effect of deep LOP in relation to our innocent control group. Moreover, we must note that our specificity rate was 100%. Much of the concerns expressed in the discussion section of experiment 1 are also applicable to experiment 2. We address the overall issues in the next section.

General Discussion

The pictorial superiority effect over words has long been established. Our experiments are an added contribution to this large body of evidence across a variety of disciplines, but especially in psychology and pedagogy. The individual detection rates were much better in the pictorial modality [100% (14/14) of IC participants, 43% (6/14) of the GISP subjects, and 60% (9/15) of the GIDP persons] than in the word condition [100% (14/14) of IC participants, 15% (2/13) of our GISP subjects and only 7% (1/14) of our GIDP individuals]. In addition, our findings of perfect detection rate in both of our IC groups align with the well-established understanding that the CIT is better at detecting "innocent" individuals (specificity) than "guilty" persons (sensitivity) (Verschuere, Crombez, Koster, & De Clercq, 2007).

When the guilty groups in both image and word conditions were conflated into one group and compared with their respective IC group, a clearer outcome emerged favoring detection accuracy of participants exposed to pictorial stimuli. Indeed, the AUC results indicated that 85% of subjects in the image experiment were correctly selected (AUC = .852, SE = .060, p = .000, 95% CI: .734-.970) (Fig. 14) compared with a non-significant rate of 65% (AUC = .652, SE = .089, p = .114, 95% CI: .477-.827) for the word modality Figure 10.

However, the unacceptably low detection rates in our GI groups, both SP and DP, require a closer examination. The CIT is described as a memory recognition technique reliably capable of detecting the P300 ERP. This component is known to be elicited when a person is exposed to "a rare and meaningful stimulus within a series also including frequently presented, less meaningful stimuli" (Rosenfeld, 2019, p. 1). We do not suspect that rareness could have had a negative influence in our results. According to Johnson (1986) "probability is unrelated to the meaning or significance of an event" (p. 370). The 1:6 probe to irrelevants ratio has been used successfully in many CTP studies by Rosenfeld and colleagues and was independently replicated in Lukacs et al. (2016). The meaningful characteristic of the probe stimulus in relation to our mock crime scenario may be a source of possible answers and therefore merits further attention.

Several researchers have looked at the meaning of meaningfulness in the context of memory encoding. Johnson (1986) explained that the sensitive nature of P300 amplitudes to meaning is a function of three independent and additive variables. First, task complexity is understood as the cognitive efforts required to process a stimulus, where "P300 amplitude is directly related to the extent to which a stimulus must be processed." (p. 373). Second, stimulus complexity relates to the perceptual demand made on the organism, and where the intricate features of a stimulus are a direct function on P300 amplitudes. In essence, the more intricate the pattern the larger P300 amplitudes are likely to be found. Third, stimulus value means the significance of an event for the organism. Johnson (1986) reports several findings where higher monetary payoffs as well as other non-monetary rewards such as the sound of a car horn, flashing lights, noxious smell, intense heat all resulted in larger P300s. More recently, Bonin, Gelin, and Bugaiska (2014) demonstrated an animacy effect in words and pictures in recall and in 'remember/know paradigms. They showed that words and pictures representing animated items were recalled and remembered more often than inanimated items and provided evidence that the animacy effect was due to animated items being richer in terms of sensory features. Brady, Alvarez, and Störmer (2019) manipulated the meaningfulness of facial stimuli and measured the N170 component. They showed their participants Black and White unambiguous Mooney faces, ambiguous faces, and non-faces and found strong evidence supporting the involvement of meaning in visual memory. They argued that a meaningful image allowed for a more elaborative encoding enabling a stronger pathway to memory for later retrieval. "Thus,

meaning may act as a “hook” to allow the retrieval of even visual details by creating specific retrieval cues.” (Brady, Alvarez, & Störmer, 2019, p. 1107).

Taken together our mock crime scenario may not have generated enough of a meaningful event in our participants to leave durable memory traces. The perfunctory nature of our task (attend a room, find an item inside a backpack, remove it with the pretend intention to steal it, and return to the lab for testing), with the possible exception of the deep processing conditions, likely led to cursory encoding and the subsequent generation of P300 amplitudes at levels insufficiently high enough for the CTP to detect significant probe-irrelevants differences. Shallow encoding of a probe stimulus under laboratory conditions is a potential vulnerability for the CTP, and a limitation that future researchers should be mindful of as they attempt to conduct investigations under more ecologically valid conditions either in laboratory or in controlled field-like settings.

CHAPTER 4: P300-BASED MEMORY DETECTION APPLIED TO A MOCK TERRORISM SCENARIO USING THE COMPLEX TRIAL PROTOCOL AND MULTIPLE PICTORIAL STIMULI

Abstract

Terrorist events require the greatest of response from law enforcement and national security investigators. The Complex Trial Protocol is a type of Concealed Information Test capable of identifying culpable pieces of information stored in the memory of offenders involved in violent crime. Forty-one undergraduate participants were randomly assigned to one of three groups, Innocent Control (IC), Simply Guilty (SG), and Guilty Countermeasure (GCM). Individuals in the SG and GCM groups underwent a mock terrorism scenario and were exposed to three probes, the face of an accomplice, a crime scene, and a mock explosive device. The GCM group performed a memory suppression countermeasure. P300 amplitude response was measured and hit rates ranged from 54%-71% (bomb), 70%-93% (crime scene), and 77%-93% (male accomplice face). Sensitivity (true negatives) rates were marginal (64% - crime scene) to good (71% - male face, 79% - bomb). The countermeasure had the opposite effect of generating higher hit rates than simply guilty individuals. Stimuli quality and ecological issues are discussed.

Keywords: Terrorism, Extremism, Complex Trial Protocol, P300, Concealed Information Test, Memory Detection.

Introduction

On April 19, 1995, Gulf War veterans Timothy McVeigh and Terry Nichols, both Caucasian males, executed a plot that resulted in the death of 168 people and injured over 500 others with the bombing of the Alfred P Murrah Federal Building in Oklahoma City, Oklahoma (Spindlove & Simonsen, 2013). While Nichols plead guilty, McVeigh's defense team unsuccessfully employed the 'empty chair strategy', a defense scheme that focuses on placing the blame on an unknown individual (Ross, 1997). Nichols was sentenced to life imprisonment and McVeigh was executed in 2001. From a prosecutorial standpoint the outcome of this tragedy may be labeled a posteriori successful. However, many questions about the identity of a certain John Doe 2 and a possible Middle East connection, not only sidetracked the FBI's investigation

for months at the time, but still remain unanswered decades later (Ross, 1997; Rohrabacher & Dugan, 2009).

At the heart of the identification dispute over possible suspects were the testimonies of Tom Kessinger and Daina Bradley. Kessinger, an employee at Elliott Body Shop, a truck rental company, remembered seeing two males two days before the bombing who attended the business to rent a Ryder truck (Rohrabacher & Dugan, 2009). Kessinger was the only one of three employees able to provide a physical description of both men to FBI sketch artist Roy Rozycki (Rohrabacher & Dugan, 2009). The composite drawings were released to the public the day after the bombing and they led to thousands of calls to authorities. While McVeigh was clearly and rapidly identified as John Doe 1, the FBI then attempted to identify John Doe 2 who did not match Nichols' description. But after a two-month long investigation, investigators called off the search reasoning that Kessinger mistook two innocent customers, also US army servicemen, who had attended the store as well to rent a truck the day after McVeigh transacted his business (Rohrabacher & Dugan, 2009). For her part Bradley was inside the Murrah Federal building minutes before the bomb blast and saw a Ryder truck being parked close by while looking out through a window (Thomas, 1997). From her hospital bed, she first described to FBI agents that the passenger of the truck was someone other than McVeigh, but she ended up testifying that she saw a male resembling McVeigh exit the driver side of the truck and a male with an olive-complexion get out from the passenger side (Thomas, 1997). The passenger, John Doe 2, has never been identified.

Eyewitness testimony is notorious for being faulty at times with foibles of human memory at the center of many cases of wrongful convictions (Davis & Follette, 2001). "Notwithstanding the potential for error in memory, American courts rely extensively, and in some cases exclusively, on witnesses' recollections to provide the "*facts*" of the cases before them." (Davis & Follette, 2001, p. 1428). Given that forensic evidence was either unavailable or not important in 95% of cases in England (Baldwin & McConville, 1980) and that forensic clues were gathered in only 10% of offences investigated by American police agencies (Horvath & Meesig, 1996), it is difficult to imagine a tribunal reaching a verdict without hearing the testimony of any witness at any point during a criminal trial. In spite of its susceptibility to error, human memory remains an important source of information and evidence in civil and criminal courts.

Beginning in 1958, David Lykken, professor of psychiatry, developed the Guilty Knowledge Test (GKT) (David T. Lykken, Awards for Distinguished Scientific Applications of Psychology, 2001). The development of this test represented a turning point in deception detection research (Ambach & Gamer, 2018). Lykken, who thought polygraphic detection of lies through autonomic arousal was impossible, turned his attention instead to detecting the presence of guilty knowledge stored in someone's memory by examining the degree of physiological agitation when presented with a series of answers, including one correct and other incorrect but equally plausible responses (Lykken, 1998). He reasoned that a guilty person would demonstrate a stronger physiological disturbance to a correct alternative to a given crime than to non-relevant items. The GKT, known nowadays as the Concealed Information Test (CIT), operates on two assumptions. First, that a notable physiological difference would appear between knowledgeable subjects relative to non-knowledgeable persons when presented with crime relevant test items (Ambach & Gamer, 2018), much like what one would expect from a class of well-prepared students successfully recognizing the correct answer to a multiple-choice question of an exam (MacLaren, 2001; klein Selle, Verschuere, & Ben-Shakhar, 2018), versus unready peers failing to recall the correct response. Second, a pertinent piece of crime information constitutes a new and significant stimulus to an individual having experiential memory of the criminal event, triggering an orienting reflex, or a sort of marked bodily disturbance characterized by an elevation of electrodermal activity or neuronal magnitude (Rosenfeld, 2011; klein Selle, Verschuere, & Ben-Shakhar, 2018). However, with respect to the multiple-choice questionnaire example, a successful pupil in this instance would represent an individual having recognized a crime detail for its novelty and meaningfulness, and ignorant students exemplify innocent suspects, responding equally to all possible answers. "In sum, crime knowledge is inferred from systematic stronger responding to the correct alternatives." (Meijer & Verschuere, 2018, p. 214).

Whereas Lykken focused his research on the skin conductance response as the metric for signs of memorable traces, another technique based on the analysis of the P300 waveform has drawn noteworthy attention of late (Meijer & Verschuere, 2018). Rosenfeld et al. (2008) developed an EEG-based method to detect concealed information said to be accurate and countermeasure resistant. This technique is better known as the Complex Trial Protocol (CTP).

Typically, the CTP involves the presentation of four types of stimuli on a computer monitor: a *probe* (the item known only to the author of a crime and the authorities), *irrelevant*

items (an assortment of similar stimuli acting as equally plausible fillers to the probe item), a *target item* (a string of numbers, typically 11111), and a series of four *non-target items* (a series of strings of numbers, typically from 22222, … to 55555). For instance, in the terrorist attack cited above, a probe could be the bomb or any of its components, the face of a known accomplice, or the crime scene where the bomb was placed. Examples of irrelevant items could be, respectively, a variety of explosive devices, faces of comparable individuals, or matching scenes.

In the CTP, a trial consists in the presentation of two types of stimuli. The pair is always made up first of a probe or an irrelevant item, followed by the random presentation of a target or non-target item. Rosenfeld and colleagues (2008) decided to create two separate discernment tasks, a rare-frequent probe-irrelevant judgement rendered implicit with the operation of a single button press from the same computer mouse, and a target-non-target discriminatory exercise made explicit with the functioning of two button presses from a second mouse, one for infrequent targets and the other for frequent non-targets. The button press response from one mouse to the first stimulus is intended to confirm that the participant has implicitly seen the stimulus (probe or irrelevant) in question, while the conditional button presses from a second mouse in response to the second stimulus is meant to confirm the participant's explicit attention to the target or non-target stimuli presentation on the monitor (Rosenfeld et al., 2008).

Evidence of pictorial superiority over words has long been documented (Kirkpatrick, 1894; Stenberg, Radeborg, & Hedman, 1995; Stenberg, 2006;). This effect endures more than a century later in associative memory studies of images of common objects and words presented to children and adult participants in picture-picture, word-picture, and word-word conditions with the advantage to the pure picture condition (Baadte & Meinhardt-Injac, 2019). Despite this lasting effect, there is however, emerging evidence that three-dimensional stimuli supplanted two-dimensional test items. Findings from Snow, Skiba, Coleman, and Berryhill (2014), supported the idea that participants who were shown real household objects outperformed, in recall and recognition tests, those who had color photos or line drawing of the same object presented to them. In terms of facial pictures, additional evidence is provided by Huang et al. (2017) who examined earliest face familiarity effects (FFEs) between familiar and unfamiliar faces. They showed 20 college students photographs of faces from famous Chinese movie stars and faces of unknown persons and measured four ERP waves, P1, N170, N250, and P300 from

57 scalp sites (including Cz, Fz, and Pz). All faces were aged matched, had neutral expressions, and were frontal views. In addition, faces were cropped to remove their outer contours, converted to gray scale, and were similar in foreground size, luminance and contrast. After a pre-test selection faces were presented upright and inverted. Their findings suggested that long-term face familiarity may not start until the N250 FFE emerged, with largest amplitudes of P300 waves recorded at centro-parietal sites for famous faces relative to unknown faces. Their results are an indication of this later brainwave's ability to react to familiar facial features.

Few studies have looked at ERPs in a CIT. In Cutmore, Djakovic, Kebbel, and Shum (2009) participants took part of a mock crime (theft of a wallet) and were later shown pictures of faces, objects and words in a three-stimuli paradigm. The probe stimuli were the face of the wallet's owner, the photo of the wallet and the word 'wallet'. The target stimuli were the owner's face of a mobile phone (that participants were purposely and innocently exposed to during the mock crime), the photo of the phone and the word 'phone'. The irrelevants were pictures of common objects (bottle, orange, watch, and pencil) and their referent word, and the face of four novel females. Objects provided the best discrimination in terms of individual bootstrap scores (95%), followed by words (86%) and faces (85%), and accuracy when using all three cues was 94%. Ambach, Bursch, Stark, and Vaitl (2010) recorded EEG and four ANS signals (electrodermal activity-EDA, heart rate, respiration, and finger plethysmogram) from 31 participants who underwent a mock crime (handling household objects while reading instructions in which the object's referent word was mentioned four times and bolded). No response was required from the participants as they viewed either words or pictures of the probe items (objects previously handled). Additional irrelevants and target stimuli were also presented. They found significant probe-irrelevant differences for each signal recorded, with the EEG-EDA pair providing an incremental value in detection accuracy of 99.3% in each modality. These studies presented with key differences from our research here. First, both groups employed a three-stimuli paradigm, whereas we tested the four-stimuli CTP. Second, Ambach, Bursch, Stark, and Vaitl (2010) used the standard base-to-peak measurement method while we utilized the peak-to-peak (pp) method. Third, while Cutmore, Djakovic, Kebbel, and Shum (2009) did measure their probe-irrelevants amplitude difference with the pp method, the pictures were displayed for 800ms, the time exposure of words was not reported, and their interstimulus intervals (ISI) were

described as “short”. In the CTP, all stimuli are shown for 300ms and our ISI is clearly indicated as varying from 1600ms to 2700ms.

Our intention was to make the mock crime as realistic as possible given laboratory constraints. Carmel, Dayan, Naveh, Raveh, and Ben-Shakhar (2003) manipulated the type of crime (standard laboratory mock crime versus more realistic mock crime), and time delay upon testing (immediate versus one week) and measured skin conductivity. All their participants underwent the same mock scenario (enter the office of a teaching assistant and steal a CD-Rom containing the examination of a psychology course). The difference between both crime conditions is that all relevant details of the crime were specified in advance to participants assigned to the standard mock crime condition, whereas none were mentioned to those in the realistic condition. Sub-groups from each mock crime condition were tested at different time delay. Investigators registered participants’ EDA as they were posed questions related to various features of the crime and displayed on a computer monitor. All participants were instructed to respond ‘No’ to each question. Their study’s findings suggested that realistic settings (52% in both delay conditions) suffered from weaker detection accuracy relative to standard conditions (80% in delayed condition and 71% in immediate condition) but recall rates of central features (90%) were far higher than peripheral items (66%). Despite the artificial allure of their ‘more realistic’ mock crime, they reasoned that because recall rates in the realistic condition were low, once only correctly recalled items were taken into consideration the effect size of crime type dropped from 0.34 to 0.11 and was not significantly different. Moreover, they posited that perpetrators of crime and witnesses probably behave differently, with the former type paying more attention to central features than the latter. As a result, they advised that central features of a crime be used instead of peripheral items and recommended using less artificial scenarios and more realistic mock crimes.

Pictorially based experiments with the CTP have been conducted in the Rosenfeld laboratory as well as elsewhere but not always completely independent. Rosenfeld, Ward, Thai, and Labkovsky (2015) concluded that, relative to verbal items, the CTP performed better with the initial pre-test exposure of pictorial stimuli and the subsequent presentation of a stimulus in a congruent testing modality. But this group only relied on group differences and did not test for individual detection rates. The CTP may perform well with group differences but there is no evidence to date of its efficiency at the individual level in terms of a pre-test picture exposure to

testing with a picture presentation of non-verbal stimuli, nor with a pre-test 3D exposure to a picture presentation during the test. Lu et al. (2017) asked their participants to steal a ring by either acting alone or as a pair. They wanted to test the influence of collaborative versus individual participation in a crime. Although both groups showed elevated P300 to the pictorial probe ring, compared to pictures of other jewelry items, P300 amplitude differences between probe and irrelevants were significantly less in the collaborative group. Results of individual diagnostic rates were also significantly inferior for the collaborative group (3/16, 19%) than the individual group (10/16, 63%). Investigators concluded that acting collaboratively probably impaired encoding of crime relevant information for both crime actors as the responsibility of committing the theft is diffused rather than concentrated as in the case of the lone offender.

It is pure conjecture as to what may have happened if authorities had subjected McVeigh, Nichols, Kessinger, Bradley or others to memory detection techniques such as an EEG-based CIT. We sought to test the CTP in a laboratory-based environment against three different pictorial stimuli (a face, a crime scene, and a visually complex object) much like what investigators could have used to assess the memory of key actors involved in the bombing of the Murrah building. To our knowledge, the CTP's performance has never been assessed with any of these types of stimuli all at once in a single experiment. In previous work, Meixner and Rosenfeld (2011) used three stimuli in a mock terrorism scenario in which participants, upon being presented with a briefing document, role played a terrorist agent planning a mock attack against the USA. The participants had to choose several options, namely the names of cities where the attack would take place (i.e. Houston, Detroit, Atlanta), the methods of attack (i.e. bomb), and calendar months (i.e. July) when the attack would occur. These three options made up the probes and they were all presented in the written form. The investigators averaged all three blocks and accurately identified all guilty (12/12) and innocent (0/12) participants at a 0.9 confidence level. They did not report the relevant analyses to test the significance of each stimuli independently or whether they had done so or not. Labkovsky and Rosenfeld (2014) attempted to use multiple probe in a mock theft experiment. They incorporated two probes in a two-part CTP. They subjected their participants to a pictorial presentation of a probe (i.e. USB drive) in Part 1 followed by a verbal presentation of a second probe (i.e. the name 'Meixner') in Part 2 of the CTP. The first part was a typical CTP, but the second part became a three-stimulus protocol as a result of the reconfiguration of the CTP from a one-probe to a two-probe protocol. They also

used eight irrelevant items for each category, verbal and pictorial. They achieved good individual diagnostics by identifying 11/15 (.73 sensitivity) of their simply guilty participants (with no false positives, 1.0 specificity) from Part 1. They achieved better results in Part 2 by accurately detecting 14/15 (.93 sensitivity) participants in the simply guilty group but at the expense of having a lower specificity (.93) in the innocent group (13/14). Given the improved detection rate when the CTP is used with two probes, it is likely that incorporating more probes may improve the CTP's efficiency even more. A regular CTP test with one probe normally takes approximately 15-20 minutes, but it is desirable to have more than one probe in any CIT in order to protect against false positives diagnoses as Labkovsky and Rosenfeld (2014) noted. They suggested that 4-6 independent probe items would be more profitable, but Lykken (1998) demonstrated that the odds of innocence versus guilt increase from 18,203:1 in a 0-item CIT to 1:282,475 in a 10-item CIT.[16] It then becomes important to verify the extent to which a CTP test can perform well in circumstances where multiple probes are used. A gap remains to evaluate the CTP's sensitivity to detect independent multiple pictorial stimuli such as the face of an accomplice, a crime scene or an elaborate object like an explosive device in one sitting.

Countermeasures.

A critical weakness in lie or memory detection testing, whether assayed through autonomic nervous system (ANS) polygraphy or EEG-P300 instrumentation, is the ability of any test to resist to countermeasures (CM). CM's are physical (i.e. pressing toes on the floor), mental (i.e. counting backwards) or cognitive (i.e. imagining being slapped on the face by the examiner) actions taken by an examinee to influence their physiological or psychophysiological responses, so as to either suppress or augment bodily manifestations. For instance, a guilty individual may attempt to make an inhibitory reaction to a crime relevant item to appear more innocent looking or behave manifestly to irrelevant items as if they were criminally relevant to obscure an information-present finding.

One particular CM is the voluntary suppression of episodic memories where participants are instructed to make a mental effort to avoid thinking about the lab-based mock crime they have just committed (Hu, Bergström, Bodenhausen, & Rosenfeld, 2015). In this case, subjects performed a mock theft of a ring and the probe stimulus was the word 'ring'. Findings from Hu,

[16] Based on several assumptions, namely that there are five scorable alternatives per item and a probability of 70% that a knowledgeable person will 'hit' on any given item (Lykken, 1998).

Bergström, Bodenhausen, and Rosenfeld (2015) showed that suppression attenuated the P300 activity to the point where innocent and guilty participants were partially indistinguishable, with the exception of a distinct late-posterior-negative (LPN) slow wave, which permitted nonetheless to discriminate guilty subjects from innocent ones. However, the study by Hu et al. (2015) used a 50-50 target to non-target ratio whereas most CTP studies have used a lower ratio of 10-90 or 20-80 (Meixner & Rosenfeld, 2011, 2014). The implications of the 50-50 target to non-target ratio are an increased demand of cognitive resources due to a higher task demand in response switching (Ward & Rosenfeld, 2017). In a follow-up study, Ward and Rosenfeld (2017) replicated the investigation by Hu et al. (2015) with a 20-80 target to non-target ratio and found no significant differences between the suppression and guilty groups, but the experiment did not include an innocent control group. Consequently, no conclusion was possible on the discriminatory performance of the CTP under those conditions. We aimed to independently replicate the suppression experiments cited above with a 20-80 target to non-target ratio.

Based on this global evidence we first reasoned that asking participants to walkthrough a 3D crime scene, to manipulate an actual mock explosive device, and to focus on the face of a real life accomplice, would likely be as realistic as possible, given the artificial nature of a laboratory setting, and improve their encoding experience. Second, we posited that a graphic stimuli representation during the testing phase, instead of a verbal one, would probably ameliorate the CTP's performance. Consequently, we expected to accurately detect our guilty participants equally from anyone of the three types of stimuli alone (H1), or from the aggregate mean value of all three stimuli (H2), and that the memory inhibiting CM will be unsuccessful and result instead in greater P300 amplitudes than other non-countering participants (H3).

Method

Participants.

A power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) estimated a sample group of about 42 was necessary to enable the detection of an effect size of f(V) = 0.7, or approximately 0.3 Cohen's *d*, at an alpha of .05. A total of 52 participants were recruited for this experiment. The data from 11 participants was excluded for making too many errors, 10 for exceeding a threshold of 20% behavioral errors, and one for committing more than two cognitive mistakes. These miscues are described further below in the procedure section. This left 41 (6 males) datasets for analysis. The mean age was 23.7 (*SD* = 5.6), ranging from 18 to 44 years old.

All were undergraduate students from Concordia University's psychology department and were offered a 1.5 course credit for their participation. All had normal or corrected-to-normal vision and declared fluency in English. None reported being color blind, nor suffering or diagnosed with a major psychological disorder (i.e. schizophrenia).

This research was authorized by Concordia University's ethics committee (certificate #30004969). All participants signed a written consent form prior to commencing the experiment. This document clearly explained the purpose of the research, the general procedure, the risks and benefits, and the conditions of participation, which included a confidentiality commitment from the experimenters.

Research design.

A four-way, 3 x 3 x 3 x 2, between-within subject factorial design was used with groups (innocent, guilty, and countermeasure) as the between-subject variable, and sites (Fz, Cz, and Pz), blocks (pictures of exploding devices, indoor household scenes, and male faces), and stimuli (probes and irrelevant) consisting of the within-subject variables.

Procedure.

Volunteer participants were first asked to read and sign a written consent and to complete a demographic data sheet. They were then randomly assigned to one of three groups, Innocent Control (IC) (n = 14), Simply Guilty (SG) (n = 13), and Guilty Countermeasure (GCM) (n = 14).

Individuals in the SG and GCM groups were greeted by a research assistant in the main laboratory room and read out a scripted scenario. The fictional storyline required the participant to pretend being part of an international terrorist organization whose leader and accomplice (MF) summoned the participant to launch a terrorist attack by means of triggering an exploding device inside an office at Concordia University. The research assistant then informed the participant where to meet the accomplice, that he would only be known by his first name (Michel), and outlined the steps the participant would have to take to arm the exploding device (see Appendix E). The participant then walked over to a nearby room where the terrorist accomplice (MF) waited. MF greeted the participant and performed a three-stage scenario aimed at reinforcing the memorization of three probes, MF's face, the exploding device, and the overall crime scene where the device was to be set (see Appendix F). The probes are further described below in the stimuli section. The first stage of the script was meant to emphasize recognition of the mock device by drawing the participant's attention to each component of the device as MF instructed

the participant on how to arm it. Once armed, the participant was asked to place the device inside a cardboard box and slide it underneath an office desk. The second step of the scenario was to stress the participant's encoding of MF's face. For instance, MF purposely asked the participant to have a good look at MF's face at one point of the scenario. The last phase involved the mental recording of the crime scene. MF asked the participant to take a few steps back to a designated spot in the room and asked him or her to describe the scene out loud. At that point MF would leave the room and wait outside until the participant executed the steps to arm the mock device. Once the mission completed, the participant stepped out of the room where MF asked him or her to go over each step of the arming sequence. This aspect of the scenario was designed to provide an additional opportunity for the participant to encode MF's facial features into memory and to memorize the device's components. Both encounters with MF, inside and outside the room, lasted an average of approx. 2-3 minutes each, for a total encoding time of about 5-6 minutes. MF then invited participants to enter the main laboratory room for testing immediately thereafter. MF made sure to have no further contact with the participant from the moment the scenario ended until the post-test questionnaires were completed. A research assistant took over the entire testing session including the installation of electrodes, providing all the necessary instructions, conducting the pop quizzes (explained further below), and administering the post-test questionnaires.

Persons assigned to the GCM condition were given the same instructions as those from the SG group and to return to the main laboratory room for testing. However, prior to testing they were instructed on how to perform the same mental countermeasure as employed in Hu et al. (2015) (see Appendix G). Once they indicated that they understood the countermeasure they were allowed to begin the experiment. In addition, the countermeasure was explained again to participants before commencing the next block of stimuli to ensure the instructions were refreshed in their mind.

Candidates in the IC group were not subjected to the mock terrorism scenario and were directly tested upon completion of the required initial documentation. They were presented with the same stimuli as individuals from the other two groups.

The overall test lasted approximately 90 minutes irrespective of the experimental condition. A pause of 3-5 minutes was allowed in between each block to provide participants with a mental rest. No participant claimed or complained to be overly fatigued to the point where

the testing session had to be interrupted or outright cancelled. However, the level of fatigue or alertness were not measured.

Trial structure and testing procedure.

In this experiment, investigators instructed participants to press the right button from a mouse on the left as fast as they could each time they saw the first stimulus, an image of a probe or an irrelevant item. They were also told to immediately press either the right or left button from another mouse on the right when they saw the second stimulus. If it was the target item 11111, they had to press the right button, and if it was one of any non-target items (i.e. 22222, 33333, 44444, or 55555), they had to press the left button on that same mouse. The purpose of the second button press, to discern between target and non-target items, was to maintain the person's attention focused on the computer screen. Participants who committed more than 20% of button press errors were excluded.[17]

Finally, the CTP investigator usually pauses the experiment periodically to quiz the participant on the identity of the last stimulus seen. Rosenfeld et al., (2008) claim that this step further ensures that the participant is fully attentive, cooperative and not employing any countermeasure. However, the literature is quite unclear on the number of interruptions to use. A review indicates that the number of pop quizzes varies from once each 7-12 trials (Rosenfeld, Ward, Frigo, Drapekin, & Labkovsky, 2015) to every 50-70 trials (Deng, Rosenfeld, Ward, & Labkovsky, 2016). One CTP experiment did not report its number of inquisitive pauses (Meixner & Rosenfeld, 2010), while some others remained vague about it, such as stating that the number of interrogative stops were made every few trials (Dietrich, Hu, & Rosenfeld, 2014), that participants' attention was monitored but not maintained (Labkovsky & Rosenfeld, 2014), that surprise tests were conducted at unpredictable times (Labkovsky & Rosenfeld, 2012; Rosenfeld, Ward, Drapekin, Labkovsky, & Tullman, 2017), or that the procedure was occasionally paused (Lu et al., 2017).

There is no available conclusion in the CTP's literature in relation to the optimal number of interruptions. We took a middle of the road approach and quizzed our participants about every 40 stimuli presentation, ranging from 38 to 50 (M = 43), for a total of eight pauses over 354 stimuli presentations. This excluded one pause, the very first one, which was used as a practice run

[17] This threshold is in keeping with the Rosenfeld laboratory.

during the first 10 trials. As these interruptions could be interpreted as a reinforcement procedure, care was taken to ensure that a pause coincided with all stimuli at least once. Since there were seven items, six irrelevant and one probe, and nine pauses (including the practice pause), this allowed for only two items to be reinforced twice. To prevent a possible reinforcement of the probe stimulus, investigators chose an irrelevant item for the extra interruptions. Participants were informed prior to testing that they would be questioned periodically on the last stimulus seen, and that more than two miscues, or a 25% error rate, would lead to their data being set aside. These are considered behavioral errors. Participants were not informed of a practice session, but investigators edited out the first 10 trials. To summarize, a total of 374 stimuli presentations were shown but only 354 were kept for analysis.

Stimuli.

As in Winograd and Rosenfeld (2011) and Lukacs and colleagues (2016), we used one probe and six irrelevant items in our experiment. The stimuli consisted of three blocks of images. The probe for the block of exploding devices was a homemade mock bomb, the size of a shoe box, 45cm×15cm×8cm (L×W×H). It was made out of wood and contained a Black alarm clock, a rat trap, a Red squeezable toy ball, a stuffed mouse, and a removable safety pin which protruded from one of the sides. The wooden box had a lid which could be opened from the top. The box was painted in Blue on the outside and White on the inside. In keeping with Lykken's (1998) plausibility criterion, the irrelevant items were "equally plausible alternatives" (p. 39). In other words, these items were not disproportionate in size, shape, colour, meaning or value such as to extraordinarily stand out in the participant's mind. The non-pertinent items were photos from the Internet of a land mine, a grenade, a Molotov cocktail, a flame thrower, sticks of dynamite, and a pipe bomb.

The probe for the second block of stimuli was an office like setting inside one of the laboratory's room. The décor was made to look like a typical office desk with a chair, lamp, coffee maker, radio, box of tissues, stapler, coffee mug, and markers. The non-criminal photos, obtained from the Internet, were of a bedroom, a living room, a dining room, a bathroom, kitchen and the inside of a garage.

The probe for the accomplice sets of stimuli was a photograph of the experimenter's face (MF) taken outdoor with a suburban residential background. The irrelevant stimuli were a selection of Caucasian males, bearded and clean shaven, with and without prescription glasses

with an aged appearance between 25 and 50 years old. These photos were taken from the Internet and were a mixed of indoor and outdoor settings.

Stimuli presentation was done through PsyTask and displayed on a 55cm HP Compact (LA2206x) flat monitor with a 1280 x 1024 resolution in a dimly lit room. The stimuli sizes were 19cm x 11cm (explosive device & crime scene), and 34cm x 19cm (male faces). At a viewing distance of 63.5cm (measured from the participant's right eye to the center of the screen) the average stimulus subtended 9.9° and 30° of visual angle respectively. The viewing distance from the participant's nasion to the fixation cross at the center of the monitor was 61cm. All items were presented in colour on a White background surrounded by a wide Black edge. The inter-stimulus interval varied from 1600ms to 2700ms. The first interval refers to the time between the presentations of the probe or irrelevant item to the next. The second interval consists of the time from the exposure of a target to a non-target item.

EEG data acquisition.

EEG data was recorded with a Mitsar amplifier, model 201 (Mitsar company, St-Petersburg, Russia) sampling at 500 Hz, and seven conductive gel filled Ag/AgCl electrodes. The ground electrode was placed on the forehead above the corrugator muscles. The electrooculogram (EOG) electrode was placed approximately one cm above the center of the left eyebrow. Three electrodes were attached to the scalp midline at sites Fz, Cz, Pz and referenced to linked mastoids. In accordance with the International 10-20 system, and prior to being tested, the distance between the inion and the nasion for each participant was measured such that the Cz electrode was consistently placed at the 50% mark on the scalp. Participants were asked to refrain from making head and upper torso movements, speaking or fidgeting in their seat, and to keep their feet flat on the floor during the test. Impedance between the scalp and electrodes was kept at below 5 KΩ. Signals were passed through the amplifier with a 30 Hz low cut filter, a 0.16 Hz high pass filter, a notch of 55-65 Hz, and a gain of 70 μV.

Offline analysis was conducted with WinEEG software (version 2.103.70, 2014). Eyeblink artifacts were corrected according to Semlitsch, Anderer, Schuster, and Presslich (1986), and all EEG and EOG segments with an amplitude over +/- 70 μV were removed from analysis.

Analysis Methods.

P300 amplitude and latency.

Investigators employed the peak-to-peak method of analyzing P300 amplitude as it has been found to be superior to the base-to-peak method (Soskins, Rosenfeld, & Niendam, 2001), and more sensitive in concealed information detection (Rosenfeld, 2011). All three sites were analyzed but our final analysis was based on the Pz site since it has been found to produce the largest amplitudes (Rosenfeld, 2011) and it is the most often site reported in the literature.

Grand averages (see Fig. 16, 17, 17, 19) were calculated with all groups and conditions according to Keil et al. (2014), and search windows were established to be 394ms, 862ms, and 1300ms for the probe. The search parameters were established using an objective method. Based on the Grand mean of the probe curve, T1 (394ms) was established as the point where the probe curve began its post-stimulus downward trajectory[18], T2 (862ms) was determined to be the point where the probe curve re-intersected the X axis past the most downward point, and T3 (1300ms) was selected arbitrarily as the point where the algorithm stopped searching for any waveform. We used a non-commercially available Matlab compatible software, supplied by Rosenfeld (personal communication, May 2015), to identify the most positive and negative peaks. The algorithm searched for the mid-point of the most positive 100ms average potential segment (also called the P300 latency) in the 394-862ms look window, and then subtracted the average of the mid-point of the most negative 100ms segment amplitude found between the 862-1300ms window. The subsequent value after subtraction was defined as the P300 peak-to-peak (p-p) amplitude.

[18] The polarity for this experiment was purposely inverted with positive amplitudes below the Y axis.

Figure 11. Grand Averages all Groups Combined

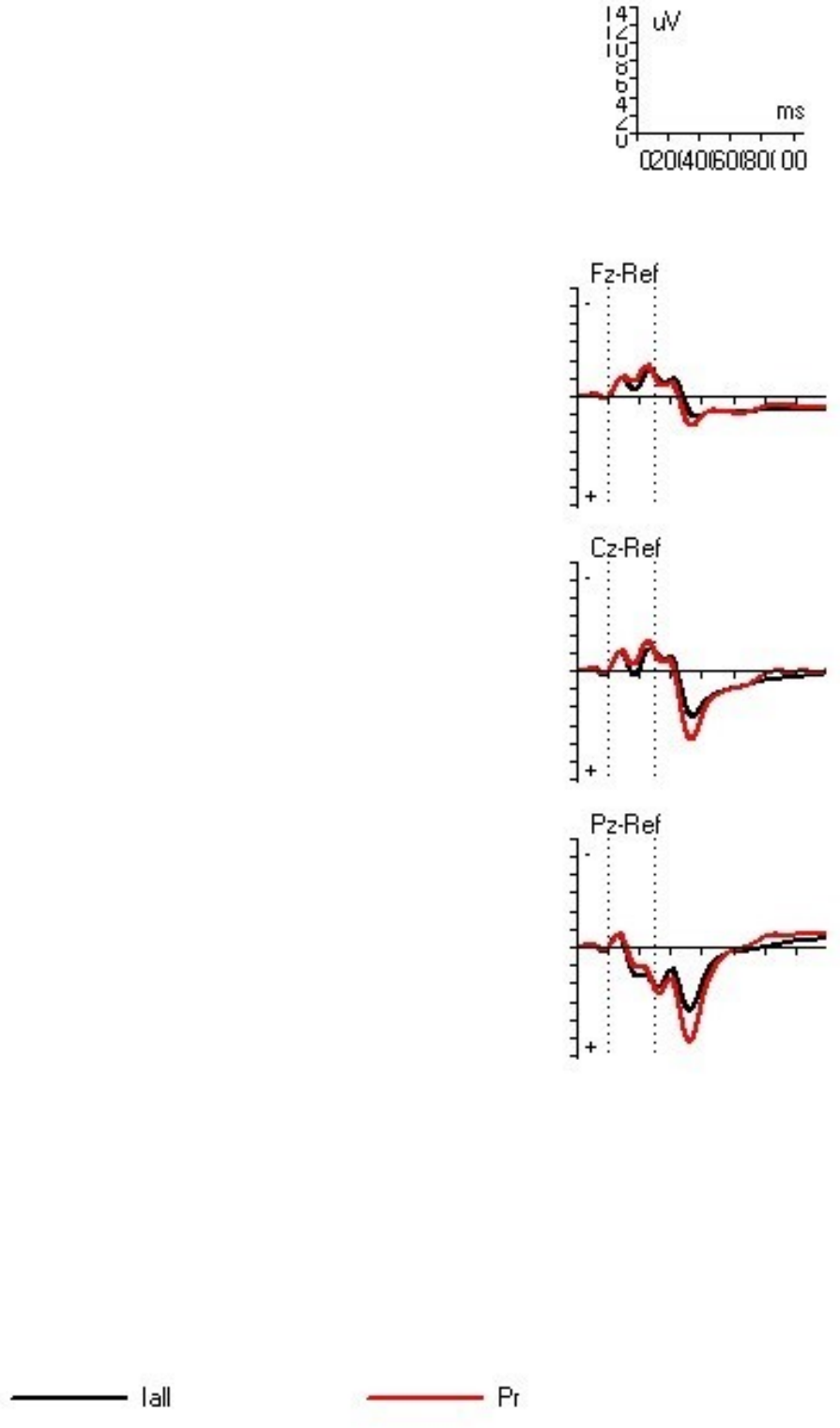


Figure 12. Grand Averages Bomb

a) Innocent Control Group b) Simply Guilty Group c) Guilty Countermeasure Group

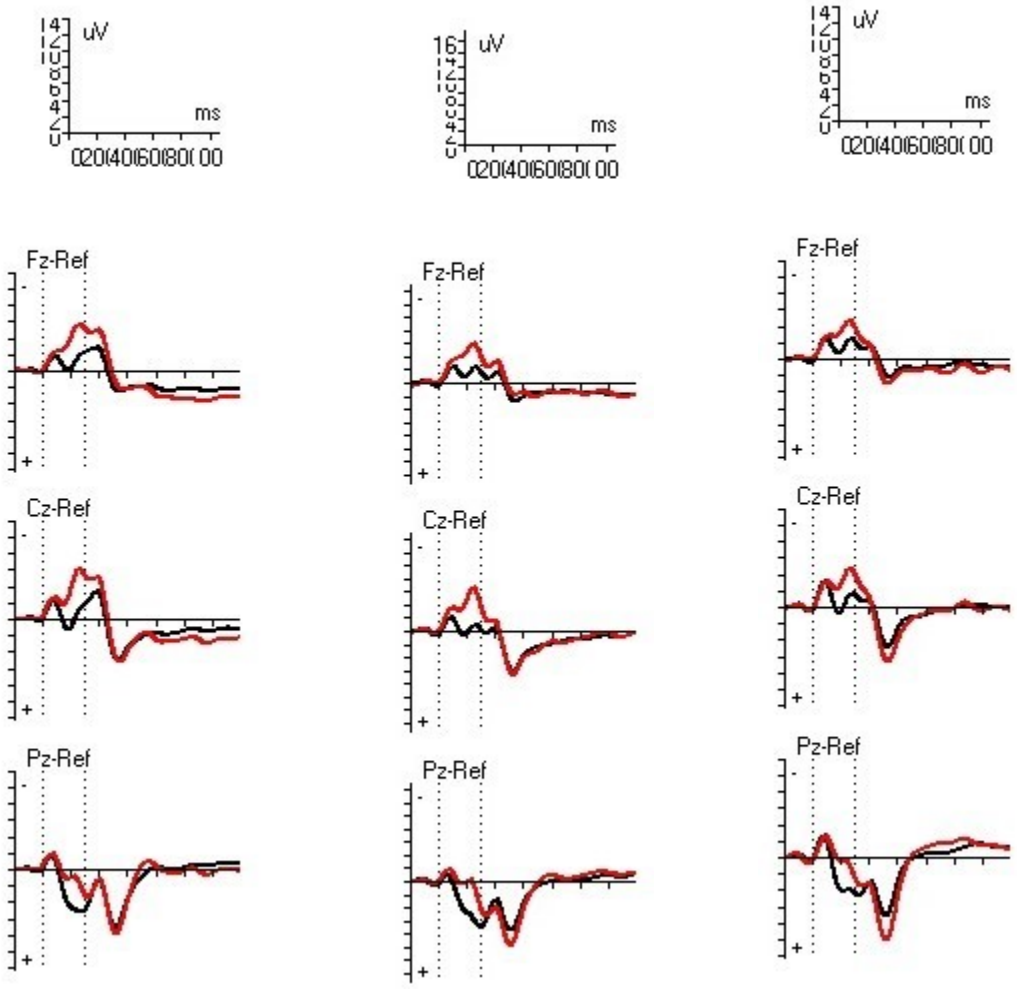
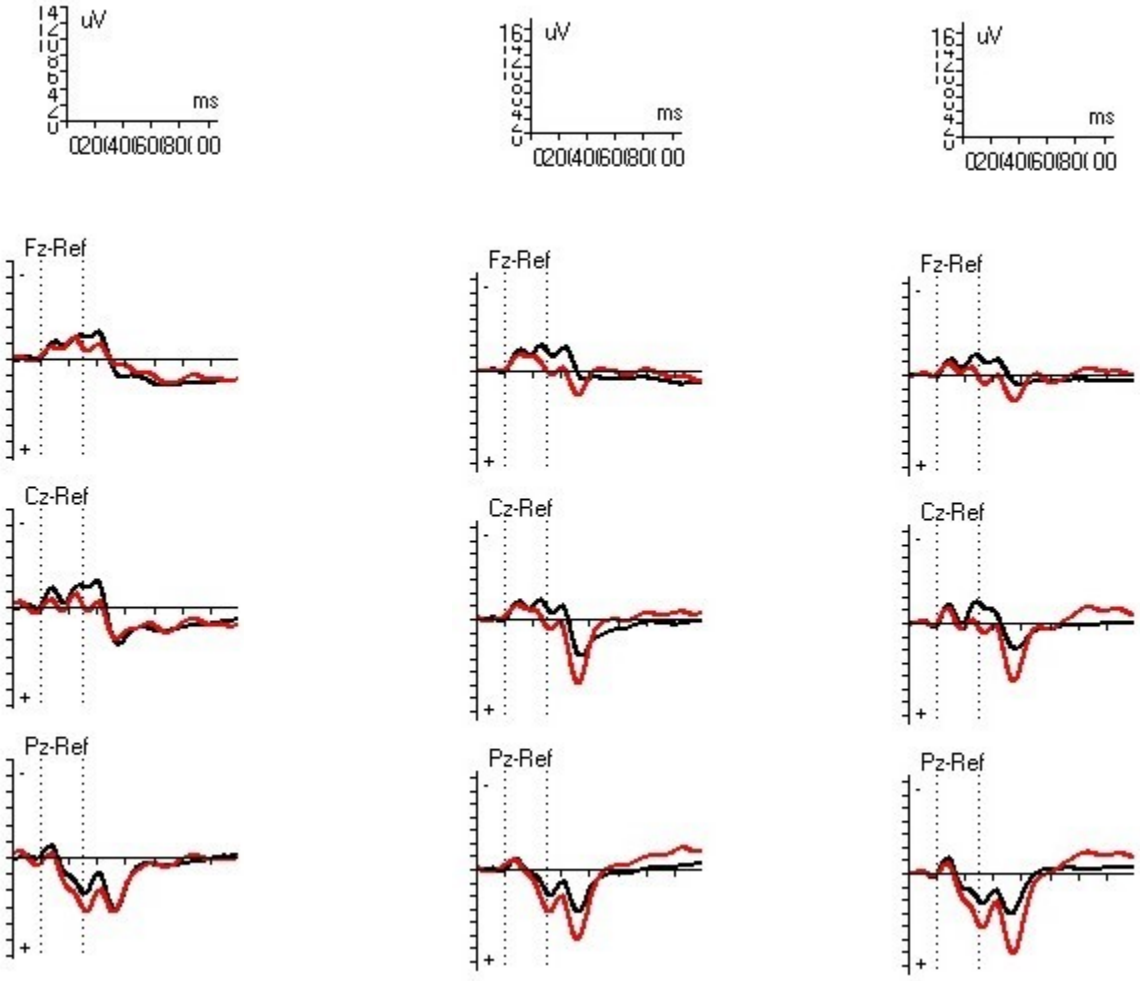


Figure 13. Grand Averages Crime Scene

a) Innocent Control Group b) Simply Guilty Group c) Guilty Countermeasure Group

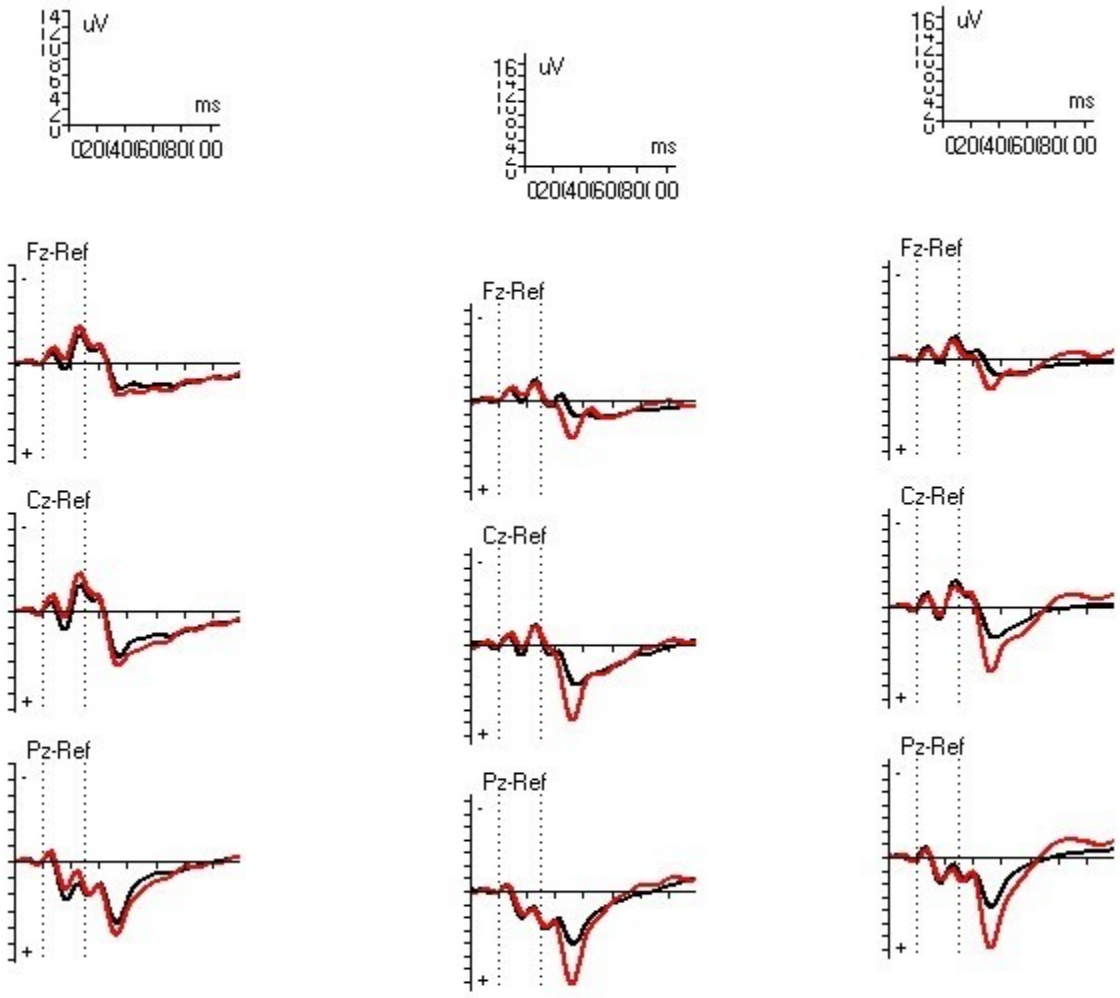


Probe

Irrelevant

Figure 14. Grand Averages Male Accomplice

a) Innocent Control Group b) Simply Guilty Group c) Guilty Countermeasure Group



Group statistical analyses.

A series of repeated measures analysis of variance (ANOVA) were conducted for group analysis. Where necessary, post-tests were conducted using a Bonferroni correction.

Individual diagnostics.

The statistical measure of choice for individual diagnostics in P300-based CIT research is bootstrapping (Rosenfeld, 2011; Rosenfeld & Donchin, 2015). This technique permits the random resampling with replacement (n-1) of an EEG single sweep data distribution instead of repeating the test many more times with the same individual. An average amplitude can then be calculated by bootstrapping a set of P300 probe waveforms for each participant. The same procedure is then applied to a corresponding set of waveforms for irrelevant items. The irrelevant P300 amplitude mean is then subtracted from the probe mean and subjected to multitude iterations. We opted here on 100 iterations as recommended by Rosenfeld, Ward, Meijer, and Yukhnenko (2017).

We used three dependent variables here as in Lu et al. (2017), and Rosenfeld, Ward, Frigo, Drapekin, and Labovsky (2015), but for each electrode site: first, the p-p P300 amplitude difference in microvolts (labelled as Dx); second, the difference between the means of the iterated bootstrapped average p-p P300s for probe and irrelevant items; and third, the greater number, out of 100, of p-p bootstrapped iterations between probe and irrelevant items where an iteration achieved a confidence value over the .9 criterion. Although this criterion level is considered traditional and somewhat arbitrary (Rosenfeld, 2011), it has been used in many P300 studies from the Rosenfeld laboratory, and it has been found to be discriminating effectively at the individual level (Meixner & Rosenfeld, 2009). In other words, for a participant to be classified as *knowledgeable*, at least 90 out 100 bootstrapped p-p P300 iterations for the probe had to be greater than 0.9 than that of the p-p P300 iterations for irrelevants. Finally, the number of bootstrapped iterations for probe also served to calculate the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) analysis.

In order to reliably identify knowledgeable participants, one must answer the question, proposed by Rosenfeld and Donchin (2015), whether "the finding that 90% or more of these bootstrapped] Probe-Irrelevant [P300 differences were greater than zero" (p. 970). Posed differently, the question that the bootstrap method answers is the following: "Is the probability more than 90 in 100 that the true difference between the average probe P300 and the average

irrelevant P300 is greater than zero?" (Rosenfeld et al., 2008, p. 909). A non-commercial computer program[19] draws at random with replacement, a set of n1 probe waveforms and a set of n2 with replacement irrelevant waveforms. It then averages these and calculates P300 amplitudes from this single average. The calculated irrelevant mean P300 is then subtracted from the comparable probe value to produce a difference value. This process is repeated 100 times. One thus obtains 100 values to place in a distribution. To state with 90% confidence, the accepted criterion used in earlier studies by Rosenfeld (2011) and others (Lu et al., 2016; Lukacs et al., 2016), that probe and irrelevant evoked ERPs are significantly different, we require that the value of zero difference or less (a negative difference) not be > -1.29 *SD* below the mean distribution of differences. In a one-tail distribution, a 1.29 criterion yields a $p < .1$ confidence level. The null hypothesis that the probe evoked P300 is greater than the irrelevant evoked P300 is then rejected if the two are not significantly different or if the irrelevant P300 is found to be larger. Lastly, as in Meixner and Rosenfeld (2011), we intend to calculate the average of the bootstrapped results of all three blocks to assist us in relation to our third hypothesis.

Receiver Operating Characteristic Analysis (ROC).

We used a ROC analysis to further substantiate the CTP's effectiveness in determining knowledgeable participants from non-knowledgeable ones. The resulting values of sensitivity (the true-positive rate) and 1-specificity (the true-negative rate) from a ROC analysis is the AUC. The AUC permits the assessment that a CIT investigator "will correctly identify the positive case when presented with a randomly chosen pair of cases in which one case is positive and one case is negative." (Eng, 2005, p. 910).

Grier's A'.

We also used Grier's *A*' (Grier, 1971) to further substantiate our individual diagnostics since *A*' takes into account false alarm rates and the ROC AUC does not (Stanislaw & Todorov, 1999). *A'* establishes the overall discriminative ability of an instrument, in this case the CTP, by computing the hit rates (true positives) and false alarms (false positives). A value of 1.0 represents perfect discernment, and 0.5 corresponds to chance discrimination.

[19] Supplied by Dr. Peter Rosenfeld.

Post-test questionnaires.

Once each test was finished, participants were asked to complete a post-test questionnaire designed for their respective condition. These questionnaires served to validate certain information such as general instructions, CM compliance, and salience of stimuli.

Results

Between-Groups Comparisons.

P300 p-p amplitudes.

The data was verified for normality, skewness, and kurtosis and found to be normally distributed. Outliers were analyzed by individual groups. One participant was found to be marginal and was retained for analysis. Otherwise all other participant's data was within +3 / -3 Z score.

Data analysis was first conducted with a series of mixed repeated measures ANOVA in SPSS (version 25) to determine main effects and interactions. Of particular interest to us was to ascertain which site produced the best results, as Pz has been reported repeatedly in the literature to be the site from which all subsequent analyses are performed. Our main dependent variable was the P300 p-p amplitude. First, a sites (Fz, Cz, Pz) x stimuli (probe, irrelevants) x blocks (bomb, crime scene, accomplice) x groups (IC, SG, GCM) mixed ANOVA revealed main effects of sites (F (2, 76) = 118.44, p = .000, ηp^2 = .757) and stimuli (F (1, 38) = 82.99, p = .000, ηp^2 = .686). Mean amplitude values were larger at Pz (M = 11.06 µV, SE = .660), relative to Fz (M = 3.94 µV, SE = .283) and Cz (M = 7.31 µV, SE = .539). Similarly, mean amplitude values were higher for probes (M = 9.24 µV, SE = .568) than irrelevants (M = 5.63 µV, SE = .388).

Several interactions also reached significance levels, such as stimuli x group (F (2, 38) = 9.59, p = .000, ηp^2 = .335), sites x stimuli (F (2, 76) = 45.30, p = .000, ηp^2 = .544), sites x blocks (F (4, 152) = 2.44, p = .049, ηp^2 = .060), stimuli x blocks (F (2, 76) = 14.72, p = .000, ηp^2 = .279), and sites x stimuli x groups (F (4, 76) = 5.36, p = .001, ηp^2 = .220). Post-hoc tests for the triple interaction indicated that the largest differences between probes and irrelevant were generated at the Pz site. As a result, we then resumed our analyses from the data collected at Pz [2 (stimuli) x 3 (blocks) x 3 (groups) repeated measure Anova].

We found a main effect of stimuli (F (1, 38) = 106.45, p = .000, ηp^2 = .737), and significant interactions of stimuli x groups (F (2, 38) = 12.06, p = .000, ηp^2 = .388) and stimuli x blocks (F (2, 76) = 8.00, p = .001, ηp^2 = .174). Neither the groups x blocks (F (4, 76) = .931, p =

.451, ηp^2 = .047) nor the blocks x stimuli x groups interaction (F (4, 76) = 1.924, p = .115, ηp^2 = .092) reached significance.

A closer examination of the pairwise comparisons from the stimuli x groups interaction indicated that P300 amplitude probe values for the IC group (M = 10.02 μV, SE = 1.35) were significantly lower than those of the GCM (M = 15.41 μV, SE = 1.35) group (p = .022). The probe amplitude levels between the IC and SG groups were not significantly different (p = .069) as were the levels between the SG and GCM groups (p = 1.000). But, as expected, the irrelevants' amplitudes were all non-significant for the three pairs of groups. However, the P300 levels of probes and irrelevants were significantly different in the IC group (p = .028), as were those of the SG (p = .000) and GCM (p = .000) groups. With respect to the stimuli x blocks interaction, the mean probe amplitude values within each block were all significantly higher than those of irrelevants; Bomb (M_{pr} = 12.34 μV, SE = .825, M_{Iall} = 9.44 μV, SE = .606, p = .000), CS (M_{pr} = 13.81 μV, SE = .960, M_{Iall} = 8.21 μV, SE = .648, p = .000), and MaleAcc (M_{pr} = 13.89 μV, SE = .892, M_{Iall} = 8.68 μV, SE = .646, p = .000).

When difference values between probes and irrelevants were compared by function of each group and blocks, only those in the SG and GCM groups reached significance. The IC group recorded probe-irrelevants differences that were marginally nonsignificant in the Bomb (p = .051) and CS (p = .103) blocks, but more convincingly nonmeaningful in the MaleAcc (p = .167) group. With one exception, all probe-irrelevants differences in all three blocks of the SG and GCM groups were significant at p = .000. The pairwise comparison in the SG-Bomb block was significant at p = .010. Taken together, it appears that the bomb stimuli may have triggered an unanticipated reaction in the IC group and pushed up the overall P300 probe-irrelevant difference average of all three blocks. This issue is addressed in greater details in the discussion portion.

When blocks were examined individually, the Bomb failed to generate significant probe amplitude differences between any of the IC, SG, GCM group pairs. In contrast, the CS produced a significant probe amplitude difference between the IC (M_{pr} = 9.76, SE = 1.64) and GCM (M_{pr} = 16.34, SE = 1.64) (p = .022), while the IC-SG (M_{pr} = 15.34, SE = 1.70) difference was not significant (p = .070). The MaleAcc stimulus lead to significant probe differences between the pairs of IC (M_{pr} = 9.82, SE = 1.53) versus SG (M_{pr} = 15.84, SE = 1.58) (p = .028)

and IC versus GCM (M_{pr} = 16.00, *SE* = 1.53) (p = .020). The SG-GCM pairwise comparison was not significant.

Individual Classification.

ROC Curves.

The AUC results from the ROC analyses are displayed at Table 6. With respect to the bomb, participants from the IC and SG groups were identified in 63% (*AUC* = .629, *SE* = .114, p = .254, 95% CI: .407-.852) while those from the IC & GCM groups were accurately classified in 76% (*AUC* = .763, *SE* = .093, p = .018, 95% CI: .580-.946). In relation to the crime scene, the IC and SG groups participants were classified in 74% (*AUC* = .739, *SE* = .098, p = .035, 95% CI: .547-.931) and individuals from the IC & GCM groups were properly sorted in 87% (*AUC* = .865, *SE* = .079, p = .001, 95% CI: .711-1.000). Finally, the face of the male accomplice resulted in an indexing rate of 81% (*AUC* = .813, *SE* = .087, p = .006, 95% CI: .643-.984) for persons in the IC and SG groups, and 94% (*AUC* = .939, *SE* = .049, p = .000, 95% CI: .844-1.000) for those in the IC & GCM groups. We also performed a ROC analyses using the mean bootstrap iterations from all three blocks. The findings indicate that 79% (*AUC* = .794, *SE* = .086, p = .009, 95% CI: .626-.962) of participants in the IC and SG groups were correctly discriminated while the rate improved to 91% (*AUC* = .908, *SE* = .062, p = .000, 95% CI: .787-1.000) for the pair of IC and GCM groups. There is no general consensus in the field on the evaluation of discriminating ability of ROC curve (AUC) (Meyers, Gamst, & Guarino, 2006), but these authors' suggest the following guidelines, .50-.60, no discrimination; .60-.70, poor; .70-.80, acceptable/fair/good; .80-.90, very good; and .90 and higher, excellent.

Table 5: Probe vs. Irrelevants Mean Bootstrap Results for each Participants for all Three Blocks

Participant	P vs. Iall		
	IC	SG	GCM
1	56	57	92
2	79	79	98
3	72	78	98
4	56	98	100
5	86	85	100
6	66	59	88
7	93	97	100
8	29	80	99
9	83	97	100
10	52	99	100
11	87	100	97
12	40	86	99
13	72	87	55
14	64		75
	13/14	5/13	11/14
Mean	**67**	**85**	**93**

Table 6: Probe vs. Irrelevants Bootstrap, Hit Rates, A', and AUC Results for each Participants for Each Block

No of Participant		Bomb			Crime Scene			Maleacc	
	IC	SG	GCM	IC	SG	GCM	IC	SG	GCM
1	67	8	77	15	71	100	86	91	100
2	97	95	95	85	82	98	54	61	100
3	88	46	100	40	97	99	88	90	96
4	28	94	100	99	100	100	40	100	100
5	100	92	100	71	100	100	86	63	100
6	41	76	65	87	69	99	71	32	100
7	84	95	100	100	97	99	95	100	100
8	26	41	97	2	98	100	60	100	99
9	59	92	100	94	100	100	97	100	100
10	43	96	100	74	100	99	38	100	100
11	79	100	100	91	100	99	90	100	91
12	33	89	97	71	68	100	17	100	100
13	49	75	28	83	92	38	85	95	100
14	100		69	66		98	25		58
Hit rate	11/14 78.6%	7/13 53.8%	10/14 71.4%	9/14 64.3%	9/13 69.2%	13/14 92.9%	10/14 71.4%	10/13 76.9%	13/14 92.9%
False rate	.214			.357			.286		
Grier A'		.742	.834		.751	.876		.826	.898
AUC		.629	.763		.739	.865		.813	.939

Bootstrapping.

In addition to ROC curves for individualized diagnostic, we utilized the bootstrapping technique as described above to predict participants' group membership (Tables 5-6). For the bomb stimulus, we were able to identify 79% of the participants in the IC group (11/14) as true negatives, 7/13 (54%) subjects in the SG group, and 10/14 (71%) individuals in the GCM group as true positives at a .9 confidence level. For the crime scene stimulus, we correctly sorted 64% of the participants in the IC group (9/14) as true negatives, 9/13 (69%) subjects in the SG group, and 13/14 (93%) individuals in the GCM group as true positives at a .9 confidence level. For the male accomplice stimulus, we accurately classified 71% of the participants in the IC group (10/14) as true negatives, 10/13 (77%) subjects in the SG group, and 13/14 (93%) individuals in the GCM group as true positives at a .9 confidence level. When all three blocks are combined, the average accuracy rates for the IC group is 71%, the SG group is 67%, and the GCM group is 86%.

We further calculated Grier's *A*' (Grier, 1971) for both SG and GCM groups (Table 6). For the SG group, the values obtained were .697 (bomb), .751 (crime scene), and .826 (male accomplice) respectively. For the GCM group the values were .800 (bomb), .876 (crime scene), and .898 (male accomplice). Grier's parameter determines the overall discriminative ability between hit rates (true positives) and false alarms (false positives) of an instrument, in this case the CTP. A value of 1.0 represents perfect discrimination, and 0.5 corresponds to chance discrimination.

Post-test questionnaire.

Nearly one third (32%) of GCM participants stated they often/always followed the instructions to suppress the mock crime. A majority of GCM (69%) participants found the memory suppression attempt moderately difficult, and a further 69% of GCM subjects were motivated to prove their innocence. 92% of SG participants rated that crime relevant memories came to mind very/extremely automatically upon seeing the relevant items.

Discussion

The findings reported here illustrate, just as Meixner and Rosenfeld (2011) found with verbal stimuli, that the CTP can detect, through multiple series of pictorial stimuli and with varying degrees of effectiveness, individuals with possession of information relative to a terrorist attack. This is the first time, to our knowledge, where the CTP was used with three successive blocks of pictorial probes during one sitting and included an item containing multiple features (e.g. explosive device), a crime scene consisting of an office-like setting, and the face of an alleged accomplice. Considering ethical constraints normally encountered when conducting laboratory experiments, we believe that our walkthrough scenario was more realistic, compared to video, photographic or written scenarios. Six out 10 of our SG subjects rated the terrorism scenario as either 'a little/somewhat/moderately' realistic.

We initially set out three hypotheses. First, we expected that the CTP would be able to detect our individuals based on any of our block of pictorial stimuli. From a group perspective, the probe-irrelevant amplitude differences were all significant within each group (IC, SG, GCM) or block (Bomb, CS, MaleAcc). This finding provides mixed evidence. On one hand, it adds to the growing body of literature that ERP-based CIT is a reliable index of memory recognition. But on the other, the IC group reacted to the probe differently than irrelevants when this is not expected. Clearly, something about the probes we used attracted too much attention from our IC

participants. On an individual basis, our hypothesis was partially supported. The bomb stimulus discriminated our SG subjects at chance level (54%). The quality of our mock bomb probe stimulus may have been an issue here. The probe lacked the realism of an improvised exploding device. It was made up of a rat trap, an alarm clock, a bright pink colored and squeezable pet toy, and a stuffed mouse. In contrast, the irrelevant items consisted of pictures of realistic-looking explosive devices drawn from the Internet, such as a grenade, dynamite sticks, etc. The non-natural look of the bomb probe could have had a reverse effect, in that its artificiality compared to the more realistic-looking irrelevant stimuli could have increased the probe's salience even for the IC. The crime scene probe, it appears, may have stood out more for innocent controls, as indicated by its high false hit rate (.357), and not constitute enough of a memorable event for simply guilty subjects, based on the marginally fair hit rate of 69%. In contrast, individuals in the GCM group were diagnosed at a much higher rate (93%). More about this below.

Second, we predicted that the mean bootstrap scores of all three types of stimuli would provide a diagnostic accuracy enough to identify our participants. This hypothesis was supported based on AUC and *A*' computations ranging respectively from .794 and .773 for SG individuals to .908 and .869 for GCM participants. In addition, bootstrap averages from all three blocks showed detection rates of 71% for the IC group, 67% for the SG group, and 86% for the GCM group.

Third, we prognosticated that the countermeasure would not affect the CTP's discriminatory efficiency of guilty participants. Our findings clearly support the CTP's ability to resist a memory suppression countermeasure. Not only did the countermeasure in question failed to influence guilty participants' P300 amplitudes, but it elevated them considerably and significantly. Participants were asked to make efforts at not remembering an event in which they actively participated in. Instead of ignoring the memories associated with the event of undergoing a mock terrorism scenario, it seems that the countermeasure forced participants to bring them back to consciousness and to further consolidate the memory traces related to the scenario. Marzi and Viggiano (2010) suggest that "deeper levels of processing involve the extraction of meaning and the deeper the level, i.e. the greater the degree of semantic or abstract processing, the more robust the memory trace." (p. 239). In sum, the countermeasure may have acted as a mechanism enabling such a deeper level of processing. An interesting avenue of investigation to verify this possibility would be to test participants after a longer delay. If the

assertion that deeper processing leads to a greater memory trace, then testing in longer time delay conditions would be valuable.

Perhaps the most interesting finding of this experiment was the CTP's performance with the face of the male accomplice. Meijer, Smulders, Merckelbach, Harald, and Wolf (2007) explored the sensitivity of the P300 component in relation to facial recognition, and found support for the use of pictures of faces in a P300-based CIT. Researchers in this study showed facial photographs of participants' siblings and good friends and asked them to deny recognition of the highly familiar probe by classifying it as unfamiliar through conditional button responses. But in a follow-up experiment, when participants (undergraduate students) were not instructed to conceal recognition of personally less familiar faces (e.g. university professors), detection was unsuccessful. The authors concluded that mere recognition could not account for the successful detection of faces, and that an affective component may underly facial recognition. Our experiment stands in sharp contrast with those findings. Notwithstanding that neither concealment nor the valence of our facial stimuli were manipulated in our study, our results point towards mere recognition as the trigger to successful detection of facial stimuli, and more so when individuals were instructed not to remember their experience.

Our findings in relation to facial recognition could add value to eyewitness accuracy in criminal matters. The body of evidence in respect to mistaken eyewitness identification as a major contributing factor to wrongful convictions is considerable (Saks & Koehler, 2005; Smith & Dufraimont, 2014; see Lefebvre, Marchand, Smith, & Connolly, 2009). What is largely insufficient is research data relative to P300-based CIT in its application to eyewitnesses of crime. Lefebvre, Marchand, Smith, and Connolly (2007) investigated the use of ERPs as a neurophysiological measure of eyewitness identification accuracy. They exposed their participants to four 60-s simulated nonviolent crime scenarios involving a male suspect entering a room and stealing a woman's purse. In each video the offender's face was visible for about 15 seconds including brief opportunities lasting 2-3 seconds where it was possible to see the culprit's face from a frontal view. The four videos were used in four delay conditions (no-delay, 1-h delay, 1-week delay, and no-delay culprit-absent condition). The photograph of the victim was shown as a target to maintain attention and responsiveness to the photographs. They found that P300 elicitation upon the presentation of the culprit, relative to fillers, resulted in individual correct identification rates varying from 79% (no-delay culprit-present), 83% (1-h delay culprit-

present), 58% (1-week culprit-present), and 46% correct rejection (no-delay culprit-absent). The results from button presses were far better (83%) with respect to the last condition. The researchers used ERP mean z scores as their dependent variables and the base-to-peak method. In a follow-up investigation, the same group of authors (2009) utilized a similar paradigm but manipulated the deception level and tested two different statistical methods to analyze their data, the bootstrap and an independent *t*-test comparing the bootstrapped averages of the culprit with the filler having produced the highest P300 amplitude. Each participant of the group exposed to the no-deception condition were accurately identified (20/20) based on both statistical methods. Under the deception condition, the bootstrap results revealed a hit rate of 90% (18/20) and the independent *t*-test method generated a hit rate of 70% (14/20). However, a Grier *A*' computation identified the independent *t*-test method as a superior detection procedure (.93) compared to the bootstrap method (.88). These two experiments highlight the reliability and effectiveness of P300-based CIT in memory recognition technique with eyewitnesses. Our study with the face of a crime accomplice adds credible evidence to the idea that EEG-based CIT can be used interchangeably with victims of crime just as offenders.

We mentioned earlier in the method section that 11 participants were discarded for committing too many cognitive or behavioral errors in any of the three blocks. This subset constitutes a potential of 33 data cells (Table 7). This is a limitation in our study. We note here that four of those breached the error threshold in all three blocks (Bomb, CS and MaleAcc), four violated the error limit in one block, and three erred too often in two blocks. The seven individuals who made too many miscues in one or two blocks also means that they produced valid results in a total of 11 out of 21 cells. According to the bootstrap data of this subset, these persons were accurately classified in 8 of the 11 data cells (73%) spread throughout the three blocks. The accepted rationale for setting aside the data of delinquent individuals is that the commission of an excess number of errors is interpreted as a lack of cooperation from the person being tested. The possible underlying behaviors of non-cooperation could be defined as inattention, lack of concentration and focus, or the application of a countermeasure of some kind. None of the uncooperative behaviors and their respective motivations have ever been explored in the context of the CTP. Such simple explanations may be easily attributable to the four participants who failed all three blocks. But they clearly lack the sophistication to explain the latent instigations behind participants who were errorless or committed a number of errors within

acceptable, albeit arbitrary, limits in one or two blocks, but exceeded the error threshold in another block. This methodological and theoretical gap poses a problem for the CTP when it comes to field use. A sizable number of participants are normally recruited in regular CTP experiments in accordance with power analysis, whereas the number of participants in field usage would be limited to one individual. The obvious question of how to treat the data of one person who exceeds the error threshold in one or more block but remains within acceptable margins in other blocks of stimuli remains unanswered. This issue is bound to be raised in a legal arena if such CTP findings are ever tendered as evidence.

Table 7: Probe vs. Irrelevants Bootstrap Scores for Rejected Participants for Each Block

No of Participant		Bomb			Crime Scene			Maleacc	
	IC	SG	GCM	IC	SG	GCM	IC	SG	GCM
1-S59F	77-tmm			70-tmm			43-tmm		
2-S60F	91-tmm			96			92		
3-S61F		79-tmm			98-tmm			99	
4-S84F		94			92-tmm			36-tmm	
5-S85F		83-tmm			96			60-tmm	
6-S89F		70-tmm			85-tmm			92-tmm	
7-S97F		94-tmm			87-tmm			100-tmm	
8-S98F		100			100-tmm			100	
9-S67F			93-tmm			100			100
10-S73F			33-tmm			87			93
11-S90F			97-tmm			100-tmm			100-tmm
Hit rate on valid cells		2/2 100%		0/1 0%	1/1 100%	1/2 50%	0/1 0%	2/2 100%	2/2 100%

Legend: tmm = Too many mistakes (cognitive and/or behavioral), X = incorrectly classified, X = correctly classified

An additional limitation of our experiment is time delay at testing, one that is frequently raised in CIT literature. Research paradigms into memory recognition often involve testing in no-delay or very short delay (e.g. 1-h, 1-d) conditions. These lengths of delays are not representative of real-life criminal investigations where an offender may be identified several weeks, months or even years after the commission of the crime, a factor *a priori* dependent on when the offence is reported to authorities in the first place. For example, sexual offences are notoriously reported to police after lengthy periods of time. Researchers should turn their attention to the impact natural memory decay has on the effectiveness of EEG-based CIT.

A final limitation is that we did not control for potential cross-racial/ethnic identification issues. The male accomplice (MF) is Caucasian and so were the neutral fillers that made up the irrelevants. Potential issues arising out of cross-racial eyewitness identifications have been documented such as the confidence-accuracy relationship (Sauer, Palmer & Brewer, 2019) and "own-race" bias has been found in field and laboratory studies (Platz & Hosch, 1988). We cannot make any claim with respect to either issues principally because we did not collect any ethnic or race related data, or confidence ratings from our participants. Investigators in future P300-based CIT studies involving facial stimuli should be mindful of these important concerns.

CHAPTER 5: GENERAL DISCUSSION

This dissertation began with the successful independent validation of the CTP using autobiographical data in chapter 2 and ended in chapter 4 with the fruitful demonstration that this protocol could be used with three stimuli in a mock terrorism scenario and with very good detection rates. However, in chapter 3, the CTP proved to be ineffective in the mock theft of a watch with the verbal probe 'watch' and its pictorial referent.

The foundation for subsequent investigations into the CTP was laid out in chapter 2. With the exception of three P300-based CIT experiments using the CTP, a fully independent one in Europe (Lukacs et al., 2016) and two semi-independent others in China (Deng, Rosenfeld, Ward, & Labkovsky, 2016; Lu et al., 2017), all extant investigations about the CTP have been the product of researchers from the Rosenfeld laboratory at Northwestern University in Chicago. Our achievement reported in chapter 2 represents the second known completely independent replication of the CTP with autobiographical data. The detection accuracy results obtained in our experiment with participant's family name were very high ranging from 93% and 94% for true positives (sensitivity) in the SG and GCM group respectively to 100% for true negatives (specificity) in the IC group. Additionally, we exposed the CTP to a novel CM and showed that the CTP was resistant to it. These findings are an important contribution to the reliability of the CTP as a potential forensic tool.

Our work in chapter 3 was clearly a disappointment. Our classification rates were unacceptably low in both modality of presentation. We should note, however, that the pictorial probe provided better results than its verbal equivalent. Our findings that LOP had little influence on the outcome also bears further investigation. From a positive point of view, we identify in the section below some weaknesses of the CTP and point to possible solutions.

Our last study highlighted some positive developments in CTP literature. First, the CTP had never been tested with three consecutive blocks of pictorial items. One of the reasons Labkovsky and Rosenfeld (2014) created the dual-probe version of the CTP was to avoid presenting multiple blocks of stimuli and prevent potential fatigue in participants. Neither Labkovsky and Rosenfeld (2014) nor us measured fatigue as a dependent variable. Notwithstanding this factor, our findings were more than satisfactory with mean detection rates across all blocks ranging from 67% to 86%. It should be noted that no participant reported being fatigued in our experiment, but this factor would definitively be required to be tested in future

research before any claim is made. Lastly, we extended the work of Hu, Bergström, Bodenhausen, and Rosenfeld (2015) and Ward and Rosenfeld (2017) in terms of memory inhibition as a possible CM. Our work demonstrated that attempting to omit from memory an episodic event had the opposite effect of forgetting the encoded stimuli. Indeed, our detection rates increased to 71%, 93% and 93% for the bomb, crime scene and accomplice's face respectively in the GCM condition. Bringing back the relevant stimuli into conscious memory in order to avoid remembering them probably consolidated their memorability. This memory consolidation may have a bearing from an applied perspective. Our participants were only exposed a few minutes to any of the three stimuli giving them minimal time to consolidate the stimuli. The CM condition would have given them the opportunity for greater consolidation. In real life, the culprit of a serious crime would likely be brought into contact with the pertinent criminal pieces of information for a longer period of time. However, serious crimes have been known to take place in a very short time span. In any event, field investigators would have to take this variable into consideration in their assessment of which probe item to select prior to conducting a CIT.

Limitations

Several limitations bear mentioning in each of our experiment. First, the autobiographical study in chapter 2 hardly reflected a realistic scenario. A person's last name is memorially rich. It is difficult, although not impossible, to imagine a criminal incident where investigators would be faced with a situation where a crucial piece of evidence would hinge on a crime suspect's family name. One possibility would be where a suspect used an alias in the execution of a crime and the alias constituted a probative piece of evidence. The outstanding question remain whether the alias would contain features as rich as the suspect's real family name. Once again, this aspect would have to be prodded further by field investigators before conducting a CIT.

Second, and in relation to our word versus pictorial study in chapter 3, before LOP is dismissed as a potential key variable in P300-based CIT future research perhaps the question of how to quantify LOP should be addressed. Despite that signals from ERPs, ocular movements, heart rate and reaction time were believed to be promising leads as a decent measurement scale (Craik, 2002), an objective index of depth of processing remains elusive (Craik, 2016). We created two deep processing tasks by asking our participants to fill incomplete sentences with the missing word watch in the verbal experiment and then asked our subjects to describe the watch

they allegedly stole in the pictorial experiment. These deep processing tasks are probably not representative of the cognitive processes a thief would undergo in real life. Furthermore, the LOPs we attempted to implement probably did not generate sufficient meaningfulness. As Craik (2002) reported "…to be effective for later memory, further processing must enrich the representation "meaningfully" in the broadest sense"…] since [… "Further processing at shallow levels of analysis does not lead to better memory." (p. 311). We laid out possible explanations in support of the larger arguments that the lack of realism of our mock theft scenario and the resulting insufficient level of meaningfulness or arousal the probe held with our guilty participants may have been more at issue than the efficiency of the CTP itself. A replication of this study in a more realistic environment where probe stimuli play a greater material role is therefore necessary before this issue is settled.

Third, our mock terrorism study demonstrated that the selection of stimuli is very important for P300-based CIT. Despite all the care in the selection of our probe items, especially the mock bomb, the latter appeared to have generated an unwanted attention from our IC group and an insufficient one from our SG and GCM groups. Clearly, the appropriate selection of probes is crucial for field investigators intent on using a P300-based CIT to further their criminal probe, for they literally have to transpose themselves in the 'shoes' of an offender and attempt to determine what stimuli the culprit would have encoded.

Future Directions

Notwithstanding the set back with the mock theft, the CTP shows good promise as an investigative tool. The CTP performed best with facial stimuli in the mock terrorism scenario. This outcome opens the door to many investigative avenues for P300-based CIT researchers, one no longer limited at developing evidence of guilt against suspects of criminal infractions.

False identifications by eyewitnesses is an international problem, and one that can have disastrous consequences. Repercussions of wrongful convictions are broad and serious. From the accused perspective, a misidentification can lead to psychological trauma and financial ruin. From a community standpoint, damaged credibility to justice related organizations (i.e. police, courts, legal profession, etc.) are likely outcomes. Dioso-Villa, Julian, Kebbell, Weathered, and Westera (2016) reported that more than 1,500 exonerees were listed on the US National Registry of Exonerations, 350 cases of corrected miscarriages of justice have been chronicled in England and 18 in Canada, and 71 individuals have been wrongly convicted in Australia between 1922

and 2015. The most common cause of wrongful convictions is eyewitness misidentification (Scheck, Neufeld, & Dwyer, 2003). Saks and Koehler (2005) examined the factors associated with wrongful conviction in 86 DNA exoneration cases in the USA. They found that eyewitness error was a contributing factor in 71% of them, ahead of forensic science testing errors (63%), police (44%) or prosecutorial (28%) misconduct, false/misleading testimony by forensic scientists (27%), dishonest informants (19%), incompetent defense representation (19%), false testimony by lay witnesses (17%), and false confessions (17%).

A great deal of research has been done in relation to eyewitness memory over the past decades. Wells et al. (1998) recommended four rules that authorities should adopt about photospreads to minimize the likelihood of a miscarriage of justice caused by misidentification. One of them is to record the confidence in the eyewitness' identification at the time of identification. This is based on the notion that post-identification events that have nothing to do with the witness' memory can dramatically affect the confidence statements of that witness (Wells et al., 1998). "By recording the eyewitness's confidence at the time of the identification, post-identification factors (which have little to do with the witness's memory) will not yet have influenced the confidence judgment" (p. 635).

Police organizations oftentimes develop their operational policies (i.e. search, seizure, arrest, etc.) in accordance with the legal framework operating within their respective jurisdiction rather than the prevailing scientific literature. But what happens when legal policy makers provide little or no guidance to police? "Currently there is no universal legislation in Australia regulating how identifications are conducted" (Dioso-Villa, Julian, Kebbell, Weathered, & Westera, 2016, p. 162). Neither the *Evidence Act 1995* (NSW) nor the *Police Powers and Responsibilities Act 2000* (Qld) meet this criterion. In Canada, the operational policy manual of the Royal Canadian Mounted Police in terms of presenting photograph packs instructs the investigator to ask the witness, upon the selection of a particular subject, "*Tell me why you recognize this person*". (RCMP, Operational Manual, Ch. 25.4), and the *Canada Evidence Act* (R.S., c. E-10, s. 1, 1985) is silent on the matter. In the USA, safeguards against eyewitness misidentifications are mostly legal in nature and post facto (i.e. motion to suppress evidence, witness cross examination, expert testimony, etc.) (Wells et al., 1998). The manner in which photospreads are presented to witnesses, in Australia as in Canada and the USA, is therefore left

to each individual police agency as they develop their respective operational policies. These policies, unfortunately, are not always based on scientific evidence, and this situation often leads to a patchwork application of procedures within the same jurisdiction (Wells et al., 1998). The use of polygraphy in lie detection and the Reid interview technique by police agencies are notorious examples of other investigative tools based on questionable evidence.

However, eyewitness confidence to diagnose the accuracy of an individual decision has some pitfalls of its own (Sauer, Palmer, & Brewer, 2019). While there is a growing body of theoretical and empirical literature in support of the conclusion that "confidence can be informative about likely accuracy in the eyewitness identification domain" (p. 162), the limitations to measuring boundary conditions with conventional methods (i.e. Confidence Accuracy Characteristics) demonstrate the difficulty in drawing "decisive conclusions about the accuracy of an individual identification made with a particular level of confidence." (Sauer, Palmer, & Brewer, 2019, p. 162).

In contrast, EEG-based CITs enable scientists and trained professionals to quantitatively assess the memory of eyewitnesses with statistically sound techniques (i.e. bootstrapping, Grier *A'*, and Receiver Operating Characteristic's Area Under the Curve). Furthermore, whether or not police investigators question eyewitnesses of crime on their confidence in identifying the culprit in a photo parade, there is currently no way for an eyewitness to place a reliable and objective metric value to their recognition of an individual. At best, eyewitnesses can subjectively assess their level of confidence by attaching to it a verbal comment (e.g. "I am sure it's him", "I am almost positive it's him") or place a quantitative value to their assessment (e.g. "I am 95% sure", "I am confident at 80-90%"). Clearly, those kinds of assessments of memory recognition are fraught with danger and vulnerable to human error. Powerful statistical techniques can fill the gap and provide an unbiased numerical value of an eyewitness' memory. It would then be left for the trier of fact to interpret the probative value of the eyewitness' confidence level in their identification of an offender.

Another direction worthy of pursuit to increase the efficiency of the CIT is by coupling EEG-based technology with eyetracking equipment. Recent evidence indicates that ocular movements (i.e. fewer fixations and longer fixation durations) are reliable markers of facial

recognition and countermeasure resistant (Millen & Hancock, 2019). Hu and Rosenfeld (2012) combined the P300-based CIT with a reaction time-based autobiographical Implicit Association Test (aIAT) and further increased the efficiency of memory detection. The combination of ERP with either behavioral or ocular measures might be a propitious approach towards the development of a multi-outcome diagnostic CIT.

REFERENCES

Ambach, W., & Gamer, M. (2018). Physiological measures in the detection and concealed information. In Peter Rosenfeld (Ed.), *Detecting concealed information and deception* (pp. 3-33). London: Academic Press.

Ambach, W., Bursch, S., Stark, R., & Vaitl, D. (2010). A concealed information test with multimodal measurement. *International Journal of Psychophysiology, 75*(3), 258-267. doi:10.1016/j.ijpsycho.2009.12.007

Andreassi, J. L. (2007). *Psychophysiology: Human behavior and physiological response* (Fifth edition ed.). Psychology Press.

Baadte, C., & Meinhardt-Injac, B. (2019). The picture superiority effect in associative memory: A developmental study. *British Journal of Developmental Psychology, 37*(3), 382-395. doi:10.1111/bjdp.12280

Baldwin, J., & McConville, M. (1980). *Confessions in crown court trials.* (No. 5). London: HMSO.

Benning, S. D., Kovac, M., Campbell, A., Miller, S., Hanna, E. K., Damiano, C. R., . . . Aaron, R. V. (2016). Late positive potential ERP responses to social and nonsocial stimuli in youth with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 46*(9), 3068-3077.

Ben-Shakhar, G. (2011). 11 countermeasures. *Memory Detection: Theory and Application of the Concealed Information Test,* 200.

Birch, C. D., Kelln, B. R. C., & Aquino, E. P. B. (2006). A review and case report of pseudologia fantastica. *Journal of Forensic Psychiatry & Psychology, 17*(2), 299-320. doi:10.1080/14789940500485128

Bonin, P., Gelin, M., & Bugaiska, A. (2014). Animates are better remembered than inanimates: Further evidence from word and picture stimuli. *Memory & Cognition, 42*(3), 370-382.

Bradley, M. M. (2009). Natural selective attention: Orienting and emotion. *Psychophysiology, 46*(1), 1-11. doi:10.1111/j.1469-8986.2008.00702.x

Brady, T. F., Alvarez, G. A., & Störmer, V. S. (2019). The role of meaning in visual memory: Face-selective brain activity predicts memory for ambiguous face stimuli. *Journal of Neuroscience, 39*(6), 1100-1108.

Canli, T., Zhao, Z., Brewer, J., Gabrieli, J. D., & Cahill, L. (2000). Event-related activation in the human amygdala associates with later memory for individual emotional experience. *Journal of Neuroscience, 20*(19), 1-5.

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the validity of the guilty knowledge test from simulated experiments: The external validity of mock crime studies. *Journal of Experimental Psychology: Applied, 9*(4), 261-269. doi:10.1037/1076-898X.9.4.261

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

Craik, F. I. M., & Lockhart, R. S. (2008). Levels of processing and Zinchenko's approach to memory research. *Journal of Russian & East European Psychology, 46*(6), 52-60. doi:10.2753/RPO1061-0405460605

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294. doi:10.1037/0096-3445.104.3.268

Craik, F. I. M. (2002). Levels of processing: Past, present... and future? *Memory, 10*(5-6), 305-318. doi:10.1080/09658210244000135

Craik, F. I. M. (2016). *Memory, attention, and aging: Selected works of Fergus I. M. Craik* Routledge Ltd. doi:10.4324/9781315440446

Cutmore, T., Djakovic, T., Kebbel, M., & Shum, D. (2009). An object cue is more effective than a word in ERP-based detection of deception. *International Journal of Psychophysiology; Int. J. Psychophysiol., 71*(3), 185-192. doi:10.1016/j.ijpsycho.2008.08.003

David T. lykken - award for distinguished scientific applications of psychology. (2001). *American Psychologist; Am. Psychol., 56*(11), 883-885. doi:10.1037//0003-066X.56.11.883

Davis, D., & Follette, W.C. (2001). Foibles of witness memory in high profile/traumatic cases. *Journal of Air Law and Commerce, 66*, 1421-1549.

Deng, X., Rosenfeld, J. P., Ward, A., & Labkovsky, E. (2016). Superiority of visual (verbal) vs. auditory test presentation modality in a P300-based CIT: The complex trial protocol for concealed autobiographical memory detection. *International Journal of Psychophysiology, 105*, 26-34. doi:10.1016/j.ijpsycho.2016.04.010

DePaulo, B. M., Ansfield, M. E., Kirkendol, S. E., & Boden, J. M. (2004). Serious lies. *Basic and Applied Social Psychology, 26*(2-3), 147-167. doi:10.1207/s15324834basp2602&3_4

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70*(5), 979-995. doi:10.1037/0022-3514.70.5.979

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118. doi:10.1037/0033-2909.129.1.74

Dietrich, A. B., Hu, X., & Rosenfeld, J. P. (2014). The effects of sweep numbers per average and protocol type on the accuracy of the P300-based concealed information test. *Applied Psychophysiology and Biofeedback, 39*(1), 67-73. doi:10.1007/s10484-014-9244-y

Dike, C. C., Baranoski, M., & Griffith, E. E. H. (2005). Pathological lying revisited. (diagnosis of pseudologia phantastica). *Journal of the American Academy of Psychiatry and the Law, 33*(3), 342-349. doi:10.1084/jem.20051128

Dioso-Villa, R., Julian, R., Kebbell, M., Weathered, L., & Westera, N. (2016). *Investigation to exoneration: A systemic review of wrongful conviction in Australia* Taylor & Francis. doi:10.1080/10345329.2016.12036066

Eng, J. (2005). Receiver operating characteristic analysis: A primer. *Academic Radiology, 12*(7), 909-916. doi:10.1016/j.acra.2005.04.005

Farwell, L. A. (2012). Brain fingerprinting: A comprehensive tutorial review of detection of concealed information with event-related brain potentials. *Cognitive Neurodynamics, 6*(2), 115-154. doi:10.1007/s11571-012-9192-2

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Ferlazzo, F., Conte, S., & Gentilomo, A. (1993). Event-related potentials and recognition memory within the 'levels of processing' framework. *Neuroreport: An International Journal for the Rapid Communication of Research in Neuroscience, 4*(6), 667-670. doi:10.1097/00001756-199306000-00016

Furedy, J. J., Davis, C., & Gurevich, M. (1988). Differentiation of deception as a psychological process: A psychophysiological approach. *Psychophysiology, 25*(6), 683-688.

Galli, G. (2014). What makes deeply encoded items memorable? insights into the levels of processing framework from neuroimaging and neuromodulation. *Frontiers in Psychiatry, 5*, 1-8. doi:10.3389/fpsyt.2014.00061

Gallo, D. A., Meadow, N. G., Johnson, E. L., & Foster, K. T. (2008). Deep levels of processing elicit a distinctiveness heuristic: Evidence from the criterial recollection task. *Journal of Memory and Language, 58*(4), 1095-1111. doi:10.1016/j.jml.2007.12.001

Gamer, M. (2011). Detecting concealed information using autonomic measures. In B. Verschuere, G. Ben-Shakhar & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 27-45)

Gamer, M., & Berti, S. (2012). P300 amplitudes in the concealed information test are less affected by depth of processing than electrodermal responses. *Frontiers in Human Neuroscience, 6*, 1-10. doi:10.3389/fnhum.2012.00308

Ganis, G. (2015). Deception detection using neuroimaging. In P. A. Granhag, A. Vrij & B. Verschuere (Eds.), *Detecting deception: Current challenges and cognitive approaches* (pp. 105-121) Wiley Online Library.

Ganis, G. (2018). Detecting deception and concealed information with neuroimaging. In P. Rosenfeld (Ed.), *Detecting concealed information and deception* (pp. 145-166) Elsevier.

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: An fMRI investigation. *Cerebral Cortex, 13*(8), 830-836.

Ganis, G., Rosenfeld, J. P., Meixner, J., Kievit, R. A., & Schendan, H. E. (2011). Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *NeuroImage, 55*(1), 312-319.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*(6), 424-429. doi:10.1037/h0031246

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science, 23*(1), 3-10.

Grubin, D., & Madsen, L. (2005). Lie detection and the polygraph: A historical review. *Journal of Forensic Psychiatry & Psychology, 16*(2), 357-369. doi:10.1080/14789940412331337353

Gudjonsson, G. (1999). The making of a serial false confessor: The confessions of Henry Lee Lucas. *The Journal of Forensic Psychiatry, 10*(2), 416-426. doi:10.1080/09585189908403693

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of polygraph tests. *Journal of Applied Psychology, 79*(2), 252.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (2002). The scientific status of research on polygraph techniques: The case for polygraph tests., 446-483.

Horvath, F., & Meesig, M. S. (1996). The criminal investigation process and the role of forensic evidence: A review of empirical findings. *Journal of Forensic Sciences, 41*(6), 963-969.

Hu, X., Bergström, Z. M., Bodenhausen, G. V., & Rosenfeld, J. P. (2015). Suppressing unwanted autobiographical memories reduces their automatic influences: Evidence from electrophysiology and an implicit autobiographical memory test. *Psychological Science, 26*(7), 1098-1106. doi:10.1177/0956797615575734

Hu, X., Hegeman, D., Landry, E., & Rosenfeld, J. P. (2012). Increasing the number of irrelevant stimuli increases ability to detect countermeasures to the P300-based complex trial protocol for concealed information detection. *Psychophysiology, 49*(1), 85-95. doi:10.1111/j.1469-8986.2011.01286.x

Huang, W., Wu, X., Hu, L., Wang, L., Ding, Y., & Qu, Z. (2017). Revisiting the earliest electrophysiological correlate of familiar face recognition. *International Journal of Psychophysiology, 120*, 42-53. doi:10.1016/j.ijpsycho.2017.07.001

Iacono, W. G., & Ben-Shakhar, G. (2019). Current status of forensic lie detection with the comparison question technique: An update of the 2003 national academy of sciences report on polygraph testing. *Law and Human Behavior, 43*(1), 1-13.

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review, 4*(4), 577-581. doi:10.3758/BF03214352

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving, 7*(1), 2-9.

Johnson Jr, R. (1986). For distinguished early career contribution to psychophysiology: Award address, 1985: A triarchic model of P300 amplitude. *Psychophysiology, 23*(4), 367-384.

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., . . . Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology, 51*(1), 1-21. doi:10.1111/psyp.12147

Kirkpatrick, E. A. (1894). An experimental study of memory. *Psychological Review, 1*, 602-609.

klein Selle, N., Verschuere, B., & Ben-Shakhar, B. (2018). Concealed information test: Theoretical background. In P. Rosenfeld (Ed.), *Detecting concealed information and deception* (pp. 35-57). London: Academic Press.

klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., & Ben-Shakhar, G. (2017). Unraveling the roles of orienting and inhibition in the concealed information test. *Psychophysiology, 54*(4), 628-639.

Knott, L. M., & Dewhurst, S. A. (2009). Investigating the attentional demands of recognition memory: Manipulating depth of encoding at study and level of attention at test. *European Journal of Cognitive Psychology, 21*(7), 1045-1071. doi:10.1080/09541440802539515

Kozel, F. A., Johnson, K. A., Grenesko, E. L., Laken, S. J., Kose, S., Lu, X., . . . George, M. S. (2009). Functional MRI detection of deception after committing a mock sabotage crime. *Journal of Forensic Sciences, 54*(1), 200-231.

Labkovsky, E., & Rosenfeld, J. P. (2012). The P300-based, complex trial protocol for concealed information detection resists any number of sequential countermeasures against up to five irrelevant stimuli. *Applied Psychophysiology and Biofeedback, 37*(1), 1-10. doi:10.1007/s10484-011-9171-0

Labkovsky, E., & Rosenfeld, J. P. (2014). A novel dual probe complex trial protocol for detection of concealed information. *Psychophysiology, 51*(11), 1122-1130. doi:10.1111/psyp.12258

Lee, T. M. C., Liu, H., Tan, L., Chan, C. C. H., Mahankali, S., Feng, C., . . . Gao, J. (2002). Lie detection by functional magnetic resonance imaging. *Human Brain Mapping, 15*(3), 157-164. doi:10.1002/hbm.10020

Lefebvre, C. D., Marchand, Y., Smith, S. M., & Connolly, J. F. (2007). Determining eyewitness identification accuracy using event-related brain potentials (ERPs). *Psychophysiology, 44*(6), 894-904. doi:10.1111/j.1469-8986.2007.00566.x

Lefebvre, C. D., 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*(3), 218-225. doi:10.1016/j.ijpsycho.2009.03.003

Loaiza, V. M., McCabe, D. P., Youngblood, J. L., Rose, N. S., & Myerson, J. (2011). The influence of levels of processing on recall from working memory and delayed recall

tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1258-1263. doi:10.1037/a0023923

Lu, Y., Rosenfeld, J. P., Deng, X., Zhang, E., Zheng, H., Yan, G., . . . Hayat, S. Z. (2017). Inferior detection of information from collaborative versus individual crimes based on a P300 concealed information test. *Psychophysiology,* 1-13. doi:10.1111/psyp.13021

Luck, S. J. (2014). *An introduction to the event-related potential technique* MIT press.

Lukács, G., Weiss, B., Dalos, V. D., Kilencz, T., Tudja, S., & Csifcsák, G. (2016). The first independent study on the complex trial protocol version of the P300-based concealed information test: Corroboration of previous findings and highlights on vulnerabilities. *International Journal of Psychophysiology, 110*, 56-65.

Lykken, D. (1998). *A tremor in the blood: Uses and abuses of the lie detector*. New York: Plenum Press.

Lynn, R. (1966). In LYNN R. (Ed.), *Attention, arousal and the orienting reaction* Pergamon. doi:https://doi.org/10.1016/B978-0-08-011524-5.50006-0 Retrieved from http://www.sciencedirect.com/science/article/pii/B9780080115245500060

Maclaren, V. V. (2001). A quantitative review of the guilty knowledge test. *Journal of Applied Psychology, 86*(4), 674-683. doi:10.1037/0021-9010.86.4.674

Mann, S., Vrij, A., Nasholm, E., Warmelink, L., Leal, S., & Forrester, D. (2012). The direction of deception: Neuro-linguistic programming as a lie detection tool. *Journal of Police and Criminal Psychology, 27*(2), 160-166. doi:10.1007/s11896-011-9097-8

Marzi, T., & Viggiano, M. P. (2010). Deep and shallow encoding effects on face recognition: An ERP study. *International Journal of Psychophysiology, 78*(3), 239-250. doi:10.1016/j.ijpsycho.2010.08.005

Maschke, G. W., & Scalabrini, G. J. (2005). The lie behind the lie detector. *Antipolygraph.Org,* 219.

Masip, J., Garrido, E., & Herrero, C. (2004a). Defining deception., 147-171. Retrieved from http://hdl.handle.net/10201/8026

Masip, J., Garrido, E., & Herrero, C. (2004b). The nonverbal approach to the detection of deception: Judgemental accuracy. *Psychology in Spain, 8*, 48-59.

Meijer, E. H., & Verschuere, B. (2018). Detection deception using psychophysiological and neural measures. In H. Otgaar, & M. Howe (Eds.), *Finding the truth in the courtroom: Dealing with deception, lies, and memories* (pp. 209-224). New York: Oxford University Press. doi:10.1093/oso/9780190612016.003.0010

Meijer, E. H., Smulders, F. T. Y., Merckelbach, Harald, L. G. J., & Wolf, A. G. (2007). The P300 is sensitive to concealed face recognition. *International Journal of Psychophysiology, 66*(3), 231-237. doi:10.1016/j.ijpsycho.2007.08.001

Meijer, E. H., Verschuere, B., Gamer, M., Merckelbach, H., & Ben-Shakhar, G. (2016). Deception detection with behavioral, autonomic, and neural measures: Conceptual and

methodological considerations that warrant modesty. *Psychophysiology, 53*(5), 593-604. doi:10.1111/psyp.12609

Meixner, J. B., & Rosenfeld, J. P. (2014). Detecting knowledge of incidentally acquired, real-world memories using a P300-based concealed-information test. *Psychological Science, 25*(11), 1994-2005. doi:10.1177/0956797614547278

Meixner, J. B., Haynes, A., Winograd, M. R., Brown, J., & Rosenfeld, J. P. (2009). Assigned versus random, countermeasure-like responses in the P300 based complex trial protocol for detection of deception: Task demand effects. *Applied Psychophysiology and Biofeedback, 34*(3), 209-220. doi:10.1007/s10484-009-9091-4

Meixner, J. B., & Rosenfeld, J. P. (2010). Countermeasure mechanisms in a P300-based concealed information test. *Psychophysiology, 47*(1), 57-65. doi:10.1111/j.1469-8986.2009.00883.x

Meixner, J. B., & Rosenfeld, J. P. (2011). A mock terrorism application of the P300-based concealed information test. *Psychophysiology, 48*(2), 149-154. doi:10.1111/j.1469-8986.2010.01050.x

Mertens, R., Allen, J., Culp, N., & Crawford, L. (2003). The detection of deception using event-related potentials in a highly realistic mock-crime scenario. Paper presented at the *Psychophysiology, 40* S60.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation*. Thousand Oaks: SAGE Publications.

Millen, A. E., & Hancock, P. J. B. (2019). Eye see through you! eye tracking unmasks concealed face recognition despite countermeasures. *Cognitive Research, 4*(1), 1-14. doi:10.1186/s41235-019-0169-0

National Research Council. (2003). *The polygraph and lie detection*. Washington: National Research Council of the National Academies.

Noble, C. E. (1952). An analysis of meaning. *Psychological Review, 59*(6), 421.

Palmer, R. C. (1989). Trial by ordeal. *Michigan Law Review, 87*(6), 1547-1556.

Peth, J., Sommer, T., Hebart, M. N., Vossel, G., Büchel, C., & Gamer, M. (2015). Memory detection using fMRI—Does the encoding context matter? *NeuroImage, 113*, 164-174.

Peth, J., Vossel, G., & Gamer, M. (2012). Emotional arousal modulates the encoding of crime-related details and corresponding physiological responses in the concealed information test. *Psychophysiology, 49*(3), 381-390.

Platz, S. P., & Hosch, H. M. (1988). Cross-racial/ethnic eyewitness identification: A field study. *Journal of Applied Social Psychology, 18*(11), 972-984.

Podlesny, J. A. (1993). Is the guilty knowledge polygraph technique applicable in criminal investigations-a review of FBI case records. *Crime Laboratory Digest, 20*(3), 57-61.

Podlesny, J. A. (2003). A paucity of operable case facts restricts applicability of the guilty knowledge technique in FBI criminal polygraph examinations. *Forensic Science Communications, 5*(3)

Porter, S., & ten Brinke, L. (2010). The truth about lies: What works in detecting high-stakes deception? *Legal and Criminological Psychology, 15*(1), 57-75. doi:10.1348/135532509X433151

Rohrabacher, D., & Dugan, P. (2009). The Oklahoma City bombing: Was there A foreign connection? *PDF). Oversight and Investigations Subcommittee of the House International Relations Committee. Archived from the Original (PDF) on March, 24*

Rosenfeld, J. P. (2005). 'Brain fingerprinting': A critical analysis. *The Scientific Review of Mental Health Practice: Objective Investigations of Controversial and Unorthodox Claims in Clinical Psychology, Psychiatry, and Social Work, 4*(1), 20-37. Retrieved from http://0-search.ebscohost.com.mercury.concordia.ca/login.aspx?direct=true&db=psyh&AN=2006-04004-004&site=ehost-live&scope=site

Rosenfeld, J. P., & Donchin, E. (2015). Resampling (bootstrapping) the mean: A definite do. *Psychophysiology, 52*(7), 969-972. doi:10.1111/psyp.12421

Rosenfeld, J. P., Hu, X., Labkovsky, E., Meixner, J., & Winograd, M. R. (2013). Review of recent studies and issues regarding the p300-based complex trial protocol for detection of concealed information. *International Journal of Psychophysiology, 90*(2), 118-134. doi:10.1016/j.ijpsycho.2013.08.012

Rosenfeld, J. P., & Labkovsky, E. (2010). New P300-based protocol to detect concealed information: Resistance to mental countermeasures against only half the irrelevant stimuli and a possible ERP indicator of countermeasures. *Psychophysiology, 47*(6), 1002-1010.

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., & Chedid, E. (2008). The complex trial protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology, 45*(6), 906-919. doi:10.1111/j.1469-8986.2008.00708.x

Rosenfeld, J. P., Sitar, E., Wasserman, J., & Ward, A. (2018). Moderate financial incentive does not appear to influence the P300 concealed information test (CIT) effect in the complex trial protocol (CTP) version of the CIT in a forensic scenario, while affecting P300 peak latencies and behavior. *International Journal of Psychophysiology, 125*(3), 42-49. doi:10.1016/j.ijpsycho.2018.02.006

Rosenfeld, J. P., Ward, A., Drapekin, J., Labkovsky, E., & Tullman, S. (2017). Instructions to suppress semantic memory enhances or has no effect on P300 in a concealed information test (CIT). *International Journal of Psychophysiology, 113*, 29-39. doi:10.1016/j.ijpsycho.2017.01.001

Rosenfeld, J. P. (2011). P300 in detecting concealed information. In B. Verschuere, G. Ben-Shakhar & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 63-89) Cambridge University Press Cambridge, UK.

Rosenfeld, J. P. (2019). P300 in detecting concealed information and deception: A review. *Psychophysiology,* 1-12. doi:10.1111/psyp.13362

Rosenfeld, J. P., Nasman, V. T., Whalen, R., Cantwell, B., & Mazzeri, L. (1987). Late vertex positivity in event-related potentials as a guilty knowledge indicator: A new method of lie detection. *International Journal of Neuroscience, 34*(1-2), 125-129. doi:10.3109/00207458708985947

Rosenfeld, J. P., Ozsan, I., & Ward, A. C. (2017). P300 amplitude at pz and N200/N300 latency at F3 differ between participants simulating suspect versus witness roles in a mock crime. *Psychophysiology, 54*(4), 640-648. doi:10.1111/psyp.12823

Rosenfeld, J. P., Ward, A., Frigo, V., Drapekin, J., & Labkovsky, E. (2015). Evidence suggesting superiority of visual (verbal) vs. auditory test presentation modality in the P300-based, complex trial protocol for concealed autobiographical memory detection. *International Journal of Psychophysiology, 96*(1), 16-22. doi:10.1016/j.ijpsycho.2015.02.026

Rosenfeld, J. P., Ward, A., Meijer, E. H., & Yukhnenko, D. (2017). Bootstrapping the P300 in diagnostic psychophysiology: How many iterations are needed? *Psychophysiology, 54*(3), 366-373. doi:10.1111/psyp.12789

Rosenfeld, J. P., Ward, A., Thai, M., & Labkovsky, E. (2015). Superiority of pictorial versus verbal presentation and initial exposure in the P300-based, complex trial protocol for concealed memory detection. *Applied Psychophysiology and Biofeedback, 40*(2), 61-73. doi:10.1007/s10484-015-9275-z

Ross, R. (1997). Impact in absentia. *American Bar Association Journal, 83*(3), 20-21.

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science, 309*(5736), 892-895.

Sauer, J. D., Palmer, M. A., & Brewer, N. (2019). Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision. *Psychology, Public Policy, and Law, 25*(3), 147-165.

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers* HMH.

Scheck, B., Neufeld, P., & Dwyer, J. (2003). Actual innocence: When justice goes wrong and how to make it right. *New American Library,* 157-172.

Schloerscheidt, A. M., & Rugg, M. D. (1997). Recognition memory for words and pictures: An event-related potential study. *Neuroreport: An International Journal for the Rapid Communication of Research in Neuroscience, 8*(15), 3281-3285. doi:10.1097/00001756-199710200-00018

Semlitsch, H. V., Anderer, P., Schuster, P., & Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology, 23*(6), 695-703. doi:10.1111/j.1469-8986.1986.tb00696.x

Seymour, T. L., & Fraynt, B. R. (2009). Time and encoding effects in the concealed knowledge test. *Applied Psychophysiology and Biofeedback, 34*(3), 177-187. doi:10.1007/s10484-009-9092-3

Skinner, D. J., & Price, J. (2019). The roles of meaningfulness and prior knowledge in younger and older adults' memory performance. *Applied Cognitive Psychology,* 1-10. doi:10.1002/acp.3552

Smirnov, A. A. (1973). *Problems in the psychology of memory*. New York: Prosveshchenie. Retrieved from http://0-search.ebscohost.com.mercury.concordia.ca/login.aspx?direct=true&db=psyh&AN=1967-16229-000&site=ehost-live&scope=site

Smith, A., & Dufraimont, L. (2014). Safeguards against wrongful conviction in eyewitness identification cases: Insights from empirical research. *Canadian Criminal Law Review, 18*(2), 199-217.

Snow, J. C., Skiba, R. M., Coleman, T. L., & Berryhill, M. E. (2014). Real-world objects are more memorable than photographs of objects. *Frontiers in Human Neuroscience, 8*, 1-11. Retrieved from http://0-search.ebscohost.com.mercury.concordia.ca/login.aspx?direct=true&db=psyh&AN=2015-06646-001&site=ehost-live&scope=site

Sokolov, E. N. (1963). Higher nervous functions: The orienting reflex. *Annual Review of Physiology,* (25), 545-580.

Sokolovsky, A., Rothenberg, J., Labkovsky, E., Meixner, J., & Rosenfeld, J. P. (2011). A novel countermeasure against the reaction time index of countermeasure use in the P300-based complex trial protocol for detection of concealed information. *International Journal of Psychophysiology, 81*(1), 60-63. doi:10.1016/j.ijpsycho.2011.03.008

Soskins, M., Rosenfeld, J. P., & Niendam, T. (2001). Peak-to-peak measurement of P300 recorded at 0.3 hz high pass filter settings in intraindividual diagnosis: Complex vs. simple paradigms. *International Journal of Psychophysiology, 40*(2), 173-180. doi:10.1016/S0167-8760(00)00154-9

Spindlove, J., & Simonsen, C. (2013). *Terrorism today: The past, the players, the future* (Fifth ed.). Boston: Pearson.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149. doi:10.3758/BF03207704

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology, 18*(6), 813-847. doi:10.1080/09541440500412361

Stenberg, G., Radeborg, K., & Hedman, L. R. (1995). The picture superiority effect in a cross-modality recognition task. *Memory & Amp; Cognition, 23*(4), 425-441. doi:10.3758/BF03197244

Suchotzki, K., Berlijn, A., Donath, M., & Gamer, M. (2018). Testing the applied potential of the Sheffield lie test. *Acta Psychologica, 191*, 281-288.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143*(4), 428-453. doi:10.1037/bul0000087

Synnott, J., Dietzel, D., & Ioannou, M. (2015). A review of the polygraph: History, methodology and current status. *Crime Psychology Review, 1*(1), 59-83. doi:10.1080/23744006.2015.1060080

Thomas, J. (1997, 24 May). Testimony backfires on defense at Oklahoma City bombing trial. *The New York Times*

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2009). Diagnostic methods 2: Receiver operating characteristic (ROC) curves. *Kidney International, 76*(3), 252-256.

Turchet, P. (2004). *La synergologie - comprendre son interlocuteur à travers sa gestuelle*. Montréal: Les éditions de l'homme.

Van Hooff, J. C., & Golden, S. (2002). Validation of an event-related potential memory assessment procedure: Intentional learning as opposed to simple repetition. *Journal of Psychophysiology, 16*(1), 12-22.

Verschuere, B., Crombez, G., Koster, E. H., & De Clercq, A. (2007). Antisociality, underarousal and the validity of the concealed information polygraph test. *Biological Psychology, 74*(3), 309-318.

Visu-Petra, G., Miclea, M., & Visu-Petra, L. (2012). Reaction time-based detection of concealed information in relation to individual differences in executive functioning. *Applied Cognitive Psychology, 26*(3), 342-351.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities (2nd ed.)*. New York, NY US: John Wiley & Sons Ltd. Retrieved from http://0-search.ebscohost.com.mercury.concordia.ca/login.aspx?direct=true&db=psyh&AN=2008-01237-000&site=ehost-live

Vrij, A. (2015). Verbal lie detection tools: Statement validity analysis, reality monitoring and scientific content analysis. *Detecting Deception: Current Challenges and Cognitive Approaches,* 1-35.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling, 5*(1), 39-43. doi:10.1002/jip.82

Vrij, A., Mann, S., & Fisher, R. P. (2006). Information-gathering vs accusatory interview style: Individual differences in respondents' experiences. *Personality and Individual Differences, 41*(4), 589-599. doi:10.1016/j.paid.2006.02.014

Vrij, A., & Verschuere, B. (2013). *Lie detection in a forensic context* Oxford University Press. doi:10.1093/obo/9780199828340-0122 Retrieved

from http://oxfordbibliographiesonline.com/view/document/obo-9780199828340/obo-9780199828340-0122.xml

Ward, A. C., & Rosenfeld, J. P. (2017). Attempts to suppress episodic memories fail but do produce demand: Evidence from the P300-based complex trial protocol and an implicit memory test. *Applied Psychophysiology and Biofeedback, 42*(1), 13-26.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads.*Law and Human Behavior, 22*(6), 603-647. doi:1025750605807

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

Winograd, M. R., & Rosenfeld, J. P. (2014). The impact of prior knowledge from participant instructions in a mock crime P300 concealed information test. *International Journal of Psychophysiology, 94*(3), 473-481. doi:10.1016/j.ijpsycho.2014.08.002

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*(5), 800.

Zernicki, B. (1987). Pavlovian orienting reflex. *Acta Neurobiologiae Experimentalis, 47*(5-6), 239-247. Retrieved from http://0-search.ebscohost.com.mercury.concordia.ca/login.aspx?direct=true&db=psyh&AN=1989-14558-001&site=ehost-live&scope=site

APPENDICES

Appendix A – (Chapter 2) Mock Burglary Scenario Briefing Sheet

Someone broke into room PY-037 at Concordia University by forcing the main door open with a screwdriver. The suspect then stole a computer, a monitor and some cash. Authorities were called to the crime scene and lifted a set of fingerprints. The police has obtained a list of students' family names from the university. The focus of the police investigation has now turned towards you.

Appendix B – (Chapter 2) Testing instructions

A theft occurred at Concordia University. We will try to use our brainwave test to find out who did it.

In this experiment, sensors will be placed on your scalp (3), forehead (2), and behind your ears (2) so that we can record brainwaves. Harmless conductive paste will be applied to the electrodes. It is easily cleaned off of the skin with water. A sink is available in the washroom outside the lab if you wish to also clean yourself immediately afterwards. We can supply you with a towel.

During the brain wave test, a series of family names and numbers are presented on a blank screen in front of you. If you need glasses to read numbers or family names, you **MUST** be sure to **WEAR YOUR GLASSES**. **Do NOT wear contacts.**

This experiment lasts about 15 minutes.

PAY ATTENTION TO ALL STIMULI! A high number of incorrect identifications will result in unusable data and early termination of your participation. All button responses are actively monitored by the experimenter.

Finally, remember to remain as relaxed as possible throughout testing. During the course of a trial, keep your eyes focused on the center of the screen where the stimulus appears. A fixation cross will appear between each stimulus. It is important that you do not blink, move your eyes, or activate any facial/head muscles, particularly right before, during, and right after the first stimulus appear; it is okay to blink quickly when you see the fixation cross. If you accidentally blink or twitch in any way during a trial, do not worry about it – just prepare yourself for the next trial.

In each "trial," you will be presented with two items, one after the other. First, a family name will be very briefly presented, and then it will disappear. Then a string of numbers will be presented. There are two responses to be made, one to each stimulus:

1) When a family name is presented, **press the 'Yes' button of the LEFT mouse**. This indicates simply that you saw the item.

2) When a string of numbers is presented (i.e. 11111, 22222, 33333, 44444, 55555), your task here is to determine whether or not the number you see is your target, which is 11111.

3) If the string is **11111, your "target" number,** you **press the RIGHT ("Yes") button of the RIGHT mouse**.

4) For any of the **other non-target numbers**, **press the LEFT ("No") button of the RIGHT mouse**.

5) The presentation of the number is also very brief, a fraction of a second, so you have to be paying attention or you will miss the number and fail to press the target or non-target buttons correctly. It is important that you pay attention to all the stimuli!

   Here is a quick recap of how each trial will occur:
   1. A family name or a string of numbers is presented
   2. If a family name appears, press the **"Yes"** button of the mouse on the left
   3. If the **"target" (i.e. 11111)** appears, press the **"Yes"** button of the mouse to the right
   4. If any other string of numbers appears **(i.e. 22222, 33333, 44444, 55555)** press the **"No"** button of the mouse to the right

Do you have any questions? Please repeat back to me what you are supposed to do.

Appendix C – (Chapter 2) List of irrelevant names

The following is a list of family names. Please have a look at them and indicate with a check mark if any of these names are particularly meaningful to you (e.g. name of a close relative or friend) or otherwise appear markedly unique compared to the other names on the list.

Scherrer
Rollison
Quill
Orlandi
Mauch
Dierks
Layden
Kembel
Holgate
Gilland
Fedler
Crosslin
Tillson
Pelzer
Yokum
Burket
Adger
Voris
Nordberg
Zank

Appendix D – (Chapter 3) Deep processing theft narrative

Read the following text and complete the blanks spaces with the appropriate word:

In the past few minutes I took part of a laboratory experiment during which I stole a _____. The _____ was located in a room near the main laboratory. I entered the room where I located a White & Grey North Face backpack. Inside the backpack was a _____. I took it from the backpack, and hid it on my person. I left the backpack there. The _____ is a man's _____. The make of the _____ is Seiko. The back face of the _____ is Blue. The overall colour of the _____ and bracelet is Silver.

A _____ is a mechanical instrument designed to tell time. This particular _____ indicates the time as well as the date and day of the week.

Appendix E – (Chapter 4) Script read by research assistant

You are a member of an international terrorist organisation. Your leaders have summoned you to launch a terrorist attack by means of exploding a device. You are to walk over to room PY-051 (down the hallway from this laboratory) and meet an accomplice by the name of [_____]. You will only know him/her by his/her first name.

[_____] will hand you over a wooden box. The box contains a mouse trap, an alarm clock, a stuffed mouse, a pet toy, and a safety pin.

[_____] will instruct you what to do with the box and where to place it. These instructions are described below:

- Open the lid,
- Squeeze the pet toy,
- Switch the pet toy with the stuffed mouse, already on the mouse trap,
- Close the lid,
- Remove the safety pin to arm the device,
- Place the wooden box inside a cardboard box,
- Slide the cardboard box underneath the desk table in the corner as pointed out by your accomplice, and
- Return to the laboratory.

If you have any question about the scenario, please ask the experimenter now.

PLEASE PAY ATTENTION TO ALL OBJECTS, FACES AND OFFICE SCENES AS YOU GO THROUGH THE SCENARIO

Appendix F – (Chapter 4) Script read by terrorist male accomplice

Experimenter waits for participant to knock on the door. The mock explosive device is exposed on the desk with the lid open. When the experimenter hears the knock he/she opens the door and says:

Experimenter: Hi, come in.

Experimenter goes to stand beside the explosive device and closes the door.

Experimenter: Do you know who I am?

If yes: That's right I am _____ and I am your terrorist leader.

If no: I am _____ and I am your terrorist leader.

Experimenter: Look at this device, I created it myself for our cause.

Experimenter: I don't want to touch it because I don't want to leave my fingerprints on it. Tell me what you see inside the box? (Wait for answer).

Experimenter: Exactly, it looks harmless…. A pink ball…..a mouse trap…. nothing in here that looks dangerous. That's the way it needs to look.

Experimenter: Now look at me closely. This may be the last time we see each other. I created this device and you are here to arm it. You bear the responsibility of arming this device and for what happens next.

Experimenter: When I leave this room, you will pick up the pink ball and squeeze it three times to activate the nuclear agent inside…. You will then swap it places with the stuffed mouse…. Then, close the lid and pull the pin very slowly until it is completely removed.

Now come over here…

Experimenter moves away from the desk and device so that the participant gets a more global view of the desk area.

Experimenter: Have a good look. Tell me what you see in general. (Wait for description answer).

Experimenter: Once you are done you will take the device and gently place it in the cardboard coffee maker box that you see under the desk and slide the box under the desk. Then leave this room and bring me the pin…I will be waiting outside this door.

Experimenter: Do you understand?

If yes: Instructor leaves the room

If no: Instructor repeats instructions and leaves the room.

When participant comes out:

Experimenter: Is it done?

Experimenter: Tell me what you did exactly. (Wait for answer).

Experimenter: Now give me the safety pin so I can destroy this evidence. Have one last look at my face. I don't forget faces and if you rat me out to the police I'll track you down. Now go back to the office you came from.

Appendix G – (Chapter 4) Memory suppression instructions

As a suspect, you really want to conceal your crime to avoid being detected. Since the brainwave based test relies on your recognition of the crime-relevant information, the best way to beat the test is to always avoid thinking of anything at all from the lab crime. Remember, to beat the test, it is extremely important that you should try your best to stop any memories about the lab crime from coming to mind during the test.

If you can stop yourself from remembering, our test may not be able to tell from your brain activity that you have the knowledge of the crime detail.

Because this task is very important to our experiment, we will take a moment to describe exactly what we want you to do when a photo appears. We always want you to view each photo, and to press the button to each photo as fast and accurately as you can, even if the photo refers to the lab crime. Do NOT simply look away from the screen, or not view such a reminder photograph, that will be a sign of un-cooperation, so you do not want to do that.

Critically, you should never think back, and you should never evaluate whether the stimulus comes from the lab crime. In other words, try as hard as you can to stop your mind from wandering to the lab crime, and process each stimulus as if it is completely unrelated to the crime. Repeating, to avoid detection, it is IMPERATIVE that you PREVENT the lab crime memories from coming to mind at all during the whole test, while looking at and paying full attention to the photos on the screen. If the lab crime does come to mind, you should push it out straight away. Although it may be challenging at first, please try your best to learn to not think about the lab crime at all, not even for a second, and not even after the photos have gone off the screen. Your goal should be never to think of the lab crime. You should accomplish this by trying to block thinking of the lab crime, but NOT by replacing it with other thoughts, any other thoughts. To repeat: DO NOT THINK of anything else than the photos on the screen while you are blocking the lab crime memories from coming to mind, just keep paying attention to, and looking at, the photos the entire time.