Exploring the Validity, Applications, and Accuracy of the P300-Based Concealed Information Test (CIT)

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Psychology

By
Anne C. Ward

EVANSTON, ILLINOIS

December 2018
ABSTRACT

Exploring the Validity, Applications, and Accuracy of the P300-Based Concealed Information Test (CIT)

Anne C. Ward

The concealed information test (CIT) has garnered more empirical support than other methods of recognition detection and has a firm theoretical foundation. Because it occurs involuntarily, even when recognition is denied, P300 amplitude is a robust indicator of concealed information. Although the P300-based CIT shows great promise for field use, there are outstanding questions regarding its validity, applications, and diagnostic accuracy. Thus, with the goal of refining the P300-based CIT for field use, the three experiments presented here aim to test: 1) the CIT’s resistance to countermeasures, 2) a novel test application, designed to detect specific details from general information, and 3) whether our diagnostic methods can be improved using machine learning (ML).

In Experiment 1, we tested the impact of retroactive memory interference (RI) on the Complex Trial Protocol (CTP) version of the P300-based CIT. Because the CIT’s detection ability is based on recognition of key items, weakening the crime memory might decrease probe identifiability and reduce diagnostic power. RI research has shown that acquiring new information after encoding a memory can degrade the original memory, which suggests that RI might threaten the CIT’s accuracy. To test this, participants (Ps) completed a mock-crime, followed by either a control task or a RI manipulation task. Both the control and RI groups were subdivided into three time delay conditions: 1/3 of Ps immediately completed the task and CIT, another 1/3 completed the task and returned a week later for the CIT, and the remaining Ps
completed both the task and CIT a week later. Results showed that, while the CIT effect (the probe-irrelevant difference upon which diagnosis is based) was obvious in all six group X time delay subconditions, its strength did not differ based on group or time delay, as was the case for target response error rates and P300 latencies. The only outcome of interest to vary by group or time delay was behavioral response times: both probe and Iall responses were delayed in the SG group.

In Experiment 2, we explored the viability of the CTP when applied as a “searching” CIT (SCIT). In the field, investigators often have trouble extracting details from individuals known to be concealing information. Here, we explored the value of the CTP when used to narrow down from general to more specific information, using a mock-crime scenario. Ps randomized to a simple guilty group stole a piece of jewelry and then hid it inside a small container, located inside a larger container, while Ps in the innocent group completed a control task. All Ps then completed three SCIT blocks, the first testing for knowledge of the jewelry item, the second for knowledge of the large container, and the third testing for the small container in which the jewelry was ultimately hidden. Analyses were conducted in two ways first, comparing the known probe to irrelevant stimuli, and next, assuming that the largest P300 response belonged to the probe. Results showed that hit rates were high using the known probe (Block 1= 97%; Block 2=87%; Block 3=81%; Overall=84%), but not the “blind” probe (Block 2=42% Block 3=29%; Overall=68%) approach.

In Experiment 3, we compare the accuracy of the bootstrapping method currently used to diagnose concealed information to an alternative ML method. Results showed the CIT effect in the guilty group but not the innocent group, and comparison of diagnostic methods revealed a seemingly higher hit rate (80% vs. 73%) and area under the curve (AUC) (.872 vs. .712) using
bootstrapping compared to ML, although standardized $z$-scores did not provide evidence suggesting the superiority of one approach over the other ($Z=1.06, p>.2$, two-tailed).

Across studies, hit rates were high when using our current methods, supporting the field readiness of the CTP. In Experiment 1, the RI manipulation did not threaten diagnostic accuracy, providing further evidence for the CTP’s countermeasure resistance. In Experiment 2, while our findings do not support that the SCIT was an effective tool for narrowing from general to specific details using traditional diagnostic methods, perfect discrimination between guilty and innocent groups was achieved by summing the number of bootstrapped iterations probe>irrelevants on known Block 1 and blind Blocks 2 and 3. In Experiment 3, results offered no evidence that the ML algorithm increased diagnostic accuracy of the CTP, and thus do not support that the ML method should replace the currently used bootstrapping method. Limitations and future directions are discussed, along with possible explanations for and consequences of our findings. Taken together, these studies address questions posed by the academic and the applied communities and provide evidence that further prepares the CTP for applied use.
# TABLE OF CONTENTS

**CHAPTER 1 THE EFFECT OF RETROACTIVE MEMORY INTERERENCE ON THE COMPLEX TRIAL PROTOCOL (CTP)**

- Introduction ........................................................................................................................................ 11
- The Concealed Information Test ........................................................................................................ 11
- Generalizability of CIT Research Time Delay and Countermeasures ............................................. 11
- Current Study ..................................................................................................................................... 14
- Hypotheses ......................................................................................................................................... 15

**Method** .......................................................................................................................................... 17

- Participants ......................................................................................................................................... 17
- Procedure Overview .......................................................................................................................... 18
- Mock-Crime ......................................................................................................................................... 18
- Interference/Control Task ................................................................................................................... 18
- Data Acquisition ............................................................................................................................... 19
- P300-based CIT ................................................................................................................................... 19
- Post-experiment questionnaire ........................................................................................................... 20

**Data Analysis Plan** .......................................................................................................................... 21

- Behavioral Response Times (RTs) and Response Errors .................................................................... 21
- ERP Analysis Plan for ERP-based CIT ............................................................................................... 21
- Intraindividual Diagnosis ................................................................................................................... 22
- Intergroup Differences ....................................................................................................................... 23

**Results** ............................................................................................................................................ 24

- Behavioral .......................................................................................................................................... 26
CHAPTER 2 A PRELIMINARY TEST OF THE “SEARCHING” P300-BASED CONCEALED INFORMATION TEST (CIT) UNCOVERING SPECIFIC DETAILS FROM MORE GENERAL INFORMATION USING THE COMPLEX TRIAL PROTOCOL (CTP)........................................................................... 43

Background.................................................................................................................. 43
The Concealed Information Test (CIT) ....................................................................... 43
P300-Based CIT............................................................................................................. 43
Three Stimulus Protocol (3SP)................................................................................ 43
Complex Trial Protocol (CTP).................................................................................. 44
The Searching Concealed Information Test (SCIT)..................................................... 44

Method....................................................................................................................... 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootstrapping</td>
<td>84</td>
</tr>
<tr>
<td>ML</td>
<td>85</td>
</tr>
<tr>
<td>ROC Analysis</td>
<td>85</td>
</tr>
<tr>
<td>Discussion</td>
<td>86</td>
</tr>
<tr>
<td>Conclusions and Future Directions</td>
<td>88</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>90</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>96</td>
</tr>
</tbody>
</table>
LIST OF TABLES AND FIGURES

CHAPTER 1
Figure 1. Study design................................................................................................................15
Figure 2. Complex Trial Protocol trial design..............................................................................20
Table 1. Outcome measures of interest, presented by group and by combined groups…………25
Figure 3. Probe and combined irrelevant response times, presented by group.........................27
Figure 4. Grand average ERPs at Pz.................................................................................................29
Figure 5. Probe and combined irrelevant P300 amplitudes, presented by group.........................31
Figure 6. Probe and combined irrelevant P300 latencies, presented by group..........................31

CHAPTER 2
Figure 1. Grand average innocent and guilty ERPs..........................................................57
Table 1. Outcome measures of interest presented by analysis method.....................................60
Figure 2. Probe and irrelevant response times and P300 latencies and amplitudes..................61
Table 2. Individual bootstrapped iterations where probe>combined irrelevant amplitude........62
Table 3. Hit rates for both the known probe and blind probe method.....................................63
Figure 3. Receiver operating characteristic (ROC) analyses and areas under the curve (AUC)64

CHAPTER 3
Figure 1. Grand average innocent and guilty ERPs..........................................................83
Figure 2. Computed grand average p-p P300 values for probes and combined irrelevants.......84
Table 1. Values obtained by bootstrapping and machine learning procedures........................85
Figure 3. Receiver operating characteristic (ROC) analyses and areas under the curve (AUCs)86 for the bootstrapping and machine learning (ML) methods..................................................86
CHAPTER 1 THE EFFECT OF RETROACTIVE MEMORY INTERERENCE ON THE COMPLEX TRIAL PROTOCOL (CTP)

Introduction

The Concealed Information Test (CIT)

The Concealed Information Test (CIT) is designed to detect memory traces that imply recognition of items that only someone linked to a crime would recognize. During a CIT, subjects are exposed to irrelevant items and probes—the items of interest (e.g., murder weapon, names, etc.) that should be familiar, and thus distinguishable from other items, only to knowledgeable participants (Ps). Because the saliency of probes should be enhanced due only to recognition, a physiological difference between probes and irrelevants—the “CIT effect”—is expected only among knowledgeable suspects. This method has a firm theoretical foundation and has garnered more empirical support than other approaches used to detect recognition (Verschuere & Ben-Shakar, 2011; Ben-Shakar, 2012).

Although the CIT has been extensively studied using a variety of physiological measures, including skin conductance, respiration line length, and heart rate, a meta-analysis comparing these autonomic nervous system (ANS) approaches to the P300 event-related potential (ERP) found that the neural measure outperformed the others (Cohen’s $d=1.89$ for P300, and $d=1.55$, 1.11, and 0.89 for SCR, RLL, and HR, respectively$^1$; Cohen, 1988; Meijer et al., 2014). In short, P300 amplitude is a robust indictor of concealed information because it occurs involuntarily, even when knowledge is behaviorally denied, when meaningful items (i.e., key crime items) are presented amongst less salient ones.

Generalizability of CIT Research Time Delay and Countermeasures

---

$^1$ According to Cohen, 0.2 can be interpreted as a small effect, 0.5 as a medium effect, and 0.8 as a large effect.
While even the least advantageous measure mentioned above (heart rate) yielded a large effect size, the results included in the meta-analysis were obtained under ideal laboratory conditions in which Ps’ memories of the crime-related items were strongest and thus easiest to detect. For example, typically, after ensuring the probe item had been encoded during the mock-crime, Ps immediately completed a CIT. However, it remains unclear how well findings obtained in the lab would generalize to field settings in which probe items may be unrehearsed and CITs may not be administered until long after the crime.

Another factor limiting the generalizability of CIT research is that, unlike most laboratory Ps, guilty suspects in the real world are highly motivated to evade detection and are likely to attempt countermeasures (CMs) in an effort to beat the test and avoid punishment. Studies testing the effect of CMs on the CIT have used both physical (e.g., instructing P to bite their tongue) and mental (e.g., think of mother’s maiden name) attempts to thwart the test, and typically instruct Ps to apply the CM when presented with particular irrelevant items, thus making them meaningful and producing a physiological response that reduces the CIT effect. However, countering irrelevant stimuli requires some level of preparation and sophistication since Ps must understand the premise of the CIT and plan their responses accordingly if they are to effectively apply CMs. While Ps in laboratory studies are often coached on the theory underlying the CIT and given practice applying their assigned CMs, these methods may compromise the generalizability of results to the field. In the real world, for a suspect to effectively apply CMs, they would have to expect that they are about to undergo a CIT, understand on their own the theory underlying the test, and either accurately predict the selected irrelevants (so they can assign and practice CMs) or be clever enough to quickly generate and apply CMs.
While most CM research has focused on narrowing the CIT effect by increasing Iall responses, CMs that instead focus on reducing the probe response have received less empirical attention. Such an approach might be advantageous because guilty Ps already know the crime-relevant item and might thus plan how to counteract its effects more readily than unknown irrelevant items. One probe-focused CM that has been tested using P300 is voluntary memory suppression, which instructs Ps to prevent crime-relevant memories from coming to mind during the test, without engaging in self-distraction. When testing suppression as a CM using episodic probes, results suggested no differences in the CIT effect between the experimental and control groups (Ward & Rosenfeld, 2017). However, when autobiographical semantic probes were instead tested, results suggested that suppression actually elevated probe P300s (Rosenfeld et al., 2017). This probe enhancement can be explained by work on thought avoidance, which shows that instructions to avoid thoughts (e.g., “don’t think about a white bear”) often paradoxically increase their frequency (e.g., Wegner, Schneider, Carter, & White, 1987; Wegner, 1989).

In Rosenfeld et al. (2017), Ps were instructed to suppress the memories related to the very information they were being presented with, which seemed to increase attention to the probe, resulting in larger P300 responses. However, instead of actively suppressing during the whole experiment, which unintentionally directs one’s focus toward the memory being tested, an alternative probe-focused CM tactic that might prove more advantageous is memory degradation. To further investigate this type of CM, Gronau et al. (2015) employed a novel approach that attempted to degrade the original memory of the mock-crime through retroactive interference (RI), which involved Ps imagining an alternate crime. They reasoned that, since the efficacy of the CIT depends upon an intact memory of crime-relevant information, detection of concealed information among CM users would suffer. Indeed, previous research on RI shows that learning
new information after the encoding of some initial information damages the original memory (e.g., Barnes & Underwood, 1959), especially when the old and new memories share retrieval cues that serve to increase their competition (Anderson & Neely, 1996). Using a mock-crime scenario and an ANS-based CIT, Gronau et al. (2015) found that the RI manipulation, which involved learning the details of a crime that did not actually occur, reduced memory of crime details and impaired identification of concealed information based on SCR, but not respiration measures. Additionally, their results showed that while delaying the CIT impaired memories of the initial crime’s details, it did not influence physiological responses.

Current Study

While CMs have been shown to compromise the validity of some ANS and traditional ERP-based CITs, the Complex Trial Protocol (CTP; Rosenfeld et al., 2008) version of the P300 brainwave-based CIT has demonstrated CM resistance and is thus the most viable for field use. Given that time and RI are known to decay memory traces upon which the CIT depends for concealed knowledge identification, it is important that these factors are explored using the CTP. While previous research has shown that knowledge of crime relevant items can still be accurately detected, even after a month delay (e.g., Hu & Rosenfeld, 2012), it is unclear how competing false information might influence test results, and CMs like RI—which attempt to make the probe less salient rather than enhancing the significance of irrelevant items—have not been thoroughly testing using this protocol. Thus, the purpose of the current study is to replicate Gronau and colleagues (2015), but using the most CM-resistant protocol the P300-based CTP.

The study’s design (see Figure 1) included both a simple guilty (SG) control group that completed a filler task and an RI group that received the interference manipulation task. Ps in these two groups were tested in one of three time delay conditions. After committing the mock-
crime, the first group immediately completed their assigned task and the CIT, while the second group immediately completed their task but returned a week later for their CIT, and the third group completed both their task and the CIT a week later.

**Figure 1. Study design. SG=Simple Guilty control group, RI=Retroactive Interference experimental group.**

<table>
<thead>
<tr>
<th>Task &amp; CIT Immediate</th>
<th>Task Immediate/CIT Delayed</th>
<th>Task &amp; CIT Delayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG (n=19)</td>
<td>SG (n=18)</td>
<td>SG (n=18)</td>
</tr>
<tr>
<td>RI (n=18)</td>
<td>RI (n=20)</td>
<td>RI (n=17)</td>
</tr>
</tbody>
</table>

**Hypotheses**

While we expected a pronounced CIT effect in all Ps (since all completed the mock-crime) and similar guilty knowledge detection rates (i.e., correct “hits”) in both the RI and SG groups due to ceiling effects, it is less clear whether the RI and SG groups will differ on other indices of concealed knowledge, including probe and lall amplitudes, the magnitude of the CIT effect (as measured by actual voltage as well as number of bootstrapped iterations where probe>combined irrelevants), and behavioral response times (RTs). For example, a memory competition hypothesis would predict reduced probe but not irrelevant P300 amplitudes in the RI group compared to the SG group, since the interference probe is creating competition only in RI subjects. Because increased P300 latencies and behavioral RTs are indicative of increased processing difficulty (Duncan-Johnson, 1981), probe-specific decreases in these measures (relative to the SG group) would provide further support for the competition hypothesis. These findings would suggest a weakened memory for the real crime.

While the goal of RI was to see if it could serve as an effective CM by reducing probe responses, it is also possible that the manipulation could instead increase Ps attention toward the
probe and enhance its corresponding response, as in the suppression study cited above by Rosenfeld et al. (2017). Whether or not enhancement occurs could depend on how the RI manipulation functions. If it succeeds at degrading the memory of the crime, or if the manipulation induces cognitive load, one would expect to see reduced probe P300s. However, if those who imagined an alternative RI scenario exhibit larger P300s, it would suggest that the manipulation increased the saliency of the probe, possibly through linked associations between the real and RI probes. That is, the cognitive association between the presented real probe and the imagined interference probe may make the real probe doubly meaningful. For example, when presented with the watch (s)he stole, a P may remember the real crime as well as the interference crime in which they imagined stealing cufflinks. Put differently, it could be that the RI task functions to increase probe P300 in much the same way CMs targeting Iall (e.g., “when you see irrelevant I, think of what you ate for breakfast”) do. In addition to a probe-specific increase in amplitude, which would suggest increased attention to the probe, an enhancement hypothesis would be further supported by increased latencies and RTs in the RI group.

Regarding the effect of time delay, we expected that memory traces of the mock-crime would remain strong enough after only a week’s delay to allow for an obvious CIT effect. Because we expected high hit rates in all groups, we did not anticipate significant differences in this outcome, regardless of the various time delays in the RI/control tasks and CIT administration and despite the findings of Gronau et al. (2015). However, while differences in the CIT effect were not expected for the immediate (SG and RI task and CIT) and combined delay conditions (SG and RI task immediate and CIT delayed, task and CIT delayed), the memory deterioration that occurs naturally over time could be strong enough to result in smaller probe amplitudes, as well as longer probe P300 latencies and RTs in the delayed CIT groups. If this were the case, one
might find group differences in the number of bootstrapped iterations where the probe amplitude exceeds that of Ialls. If the RI effect is larger in the delay group that received the manipulation immediately before the CIT (i.e., delayed task and CIT condition) than the group that received the RI manipulation during their first session (i.e., task immediate, delayed CIT), it would support that “refreshing” the deteriorated memory of the probe with a competing (imagined) probe is more effective than trying to replace a freshly acquired memory.

**Method**

**Participants**

One-hundred and forty-nine Ps from the Psych110 pool were randomly assigned to one of six groups (see Figure 1) and participated for class credit. These six conditions reflect variations in the a) memory manipulation (i.e., retroactive interference vs. control), b) when the interference manipulation/control task were administered (i.e., immediately following the mock-crime vs. one week later), and c) when the CIT was administered (i.e., during the same session as the mock-crime vs. a week later). So, within each of the retroactive interference and control conditions, two-thirds of Ps completed the interference/control task in the same session as the mock-crime while the remaining Ps completed it in another session 6-8 days later. Additionally, one-third of Ps completed the CIT in the same session as the mock-crime while the remaining two-thirds underwent the test the following week, resulting in the following groups for each the SG and RI conditions 1) task and CIT immediate, 2) task and CIT delayed, 3) task immediate and CIT delayed. Varying the time between the mock-crime, control/interference task, and CIT allowed us to examine the possible interaction between the interference manipulation and memory deterioration that can be attributed to time.
Procedure Overview

All Ps committed a mock-crime, followed by either a control or the RI task (in the same session as the crime or a week later), the P300-based CIT (at either session 1 or 2), and finally, a post-experiment questionnaire.

Mock-Crime

Upon their arrival and consent, Ps were assigned to either the RI or SG control condition. They were then given mock-crime instructions directing them to use an enclosed key to open an office containing a desk (see Appendix A). In the top drawer of this desk, they were told they would find an item in a yellow envelope. They were instructed to take the item and hide it in a designated mailbox in the building’s basement, and then return to the lab. In order to ensure that the crime-relevant probe was not inherently more meaningful unintentionally, the hidden item (watch, bracelet, and necklace) was alternated across Ps.

Interference/Control Task

Upon returning to the lab, two-thirds of Ps immediately completed the RI or control task while the remaining third returned to the lab in 6-8 days for their second session, during which they completed their task and the CIT.

Because false memories that share the same cues as true memories create the most competition and pose the most severe CM threat, we instructed Ps in the RI condition to imagine the mock-crime that they committed, while only replacing specific details (see Appendix B for interference and control scripts). Namely, Ps received instructions asking them to instead imagine a changed crime-relevant office number, type of furniture, location of drawer, envelope color, and stolen jewelry item. The RI probes that Ps were asked to imagine were alternated between Ps. After the interference task, Ps were asked to recall the RI items to ensure they were
encoded. Instead of imagining an alternative crime, Ps in the control condition completed a word find puzzle for five minutes to control for the passage of time. The puzzle was a letter matrix, among which Ps were instructed to find weather themed words (e.g., fog, snow, cloud) from a provided list. Ps in both groups were told that focusing on their task might help them pass the brainwave-based test designed to identify the stolen item, and that passing the test would result in a $5 reward. After their task, one third of Ps completed the CIT in the same session while the remaining two thirds of Ps returned the following week.

Data Acquisition

After the mock-crime and RI/control task, continuous EEGs were collected using Ag/AgCl electrodes attached to the scalp midline at sites Fz, Cz, and Pz, and referenced to linked mastoids. Electro-oculogram (EOG) was recorded referentially with an electrode placed above the left eye to record eye movements and blinks. Electrode impedance remained under 5 kΩ, the criterion for artifact rejection was 70 µV, and eye blink artifacts were corrected using the Semlitsch (1986) method. The 19 channel Mitsar amplifier passed signals with a 30 Hz low pass and a 0.16 Hz high pass filter setting, and output passed through a 16-bit Mitsar A/D converter, sampling at 500 Hz. For display and analyses, single sweeps and averages were filtered off-line to remove higher frequencies, with the digital filter set to pass frequencies from 0 to 6 Hz using a Kaiser (alpha=1.8) filtering algorithm.

P300-based CIT

After instructing them how to respond (see Appendix C), Ps then completed the CTP version of the CIT, which is divided into two parts (see Figure 2). For each trial, an initial fixation cross “+” (200 ms) was replaced with either a probe or an irrelevant (300 ms), followed by 1100 ms of observation, and finally a randomly varying interstimulus interval of 50-200 ms.
(with fixation “+”). This was then repeated in the second part of the trial, except with the target and nontarget stimuli. In the first part of the trial, the probe was presented among six other images from the same category that were not involved in the crime (i.e., other jewelry items; Ialls), and each stimulus required a simple “I saw it” response with the left hand, regardless of whether it was a probe or an irrelevant. The subsequent response depends on stimulus type, and Ps pressed the right button on the right hand mouse when presented with a target (i.e., “11111”) and the left button when presented with a nontarget (i.e., “22222…55555”). Each run consisted of 210 trials and Ps were stopped about every 50 trials and asked to recall the last image they saw, so as to force their attention.

*Figure 2. Complex Trial Protocol trial design. Pr/Iall=probe/irrelevants; NT=nontarget; Tar=target*

Post-experiment Questionnaire

Following the CIT, Ps completed a post-experiment questionnaire inquiring about their focus on the control/RI task while completing it as well as during the CIT, how often crime memories arose when they saw the probe, whether they attempted additional CMs, and their motivation for and confidence in beating the test (see Appendix D). Although we believed that
the memory of the real probe would remain strong enough to allow for detection using P300, even with the RI manipulation and after a delay, it is less clear how well memories for peripheral crime-related details (e.g., room number of crime; see Appendix D) might be influenced by RI.

To further investigate how group membership and when the task was administered might impact Ps’ general crime memory, we asked those who returned for a second appointment additional questions regarding their memory of the crime.

**Data Analysis Plan**

*Behavioral Response Times (RTs) and Response Errors*

Since probe response times (RTs) and errors often increase when cognitive resources are taxed, as when applying CMs, (Verschuere, 2011; Rosenfeld & Labkovsky, 2010) these behavioral measures were assessed. Because probes and Ialls require a single response, errors were based on target responses in the second half of the trial. Response errors included uncorrected wrong target responses as well as target responses that were skipped.

*ERP Analysis Plan for ERP-based CIT*

P300 was assessed at Pz, where it is typically largest (Johnson, 1993), and was measured peak-to-peak (p-p), since this method has proven at least 25% more accurate than baseline-to-peak calculations (Soskins, Rosenfeld, & Niendam, 2001). Across all conditions, in cases where Ps fell short of being identified as knowledgeable using Pz P300, the Cz site was used if the p-p probe amplitude was >1μV larger than at Pz.

To compute p-p amplitudes, the algorithm found the maximally positive 100ms segment following the stimulus within 350-750ms, the midpoint of which is P300 latency. The average of the maximally negative 100ms segment in the time following P300 latency to 1300ms was then found and subtracted from the average of the highest positive 100ms segment’s value to arrive at
the p-p amplitude. As recommended by Keil et al. (2014), look windows were chosen using a grand average including all Ps, and each ERP was checked so windows could be adjusted to include P300 peaks occurring outside the default windows. Since recognized probes often take longer to process, we also checked for differences in P300 probe and irrelevant latencies.

Three measures that gauge the strength of the CIT effect are also reported, including the μV difference between the probe and combined irrelevant (“Iall”) amplitudes (probe-Iall), number of bootstrapped iterations where the probe amplitude is greater than that of the combined irrelevants (probe>Iall), and the average difference between Pr and Iall bootstrapped mean amplitudes.

*Intraindividual Diagnosis*

For each individual, the p-p bootstrapped amplitude difference method was used to determine if their probe response was truly larger than the response to the combined irrelevants. Since single EEG sweeps are noisy, it would be preferable to compare multiple probe and irrelevant responses. However, this is not feasible since administering the test to each P several times could produce unintended psychological effects that would compromise the interpretation of results (e.g., habituation, fatigue, irrelevants gaining relevance through repeated exposure). To get around this, Farwell and Donchin (1991) introduced bootstrapping (Efron & Tibshirani, 1994) to the field. This method works by generating distributions that simulate what results would be if a single P completed one test several (e.g., 100) times, and lets one determine with a known level of confidence (e.g., 90%) that Pr>Iall, within an individual.

To carry out the bootstrapping procedure, we first randomly sampled 30 probe and 30 irrelevant waves (with replacement) from the original set of sweeps and used them to create probe and irrelevant ERPs. Next, probe and irrelevant average amplitudes were computed from
each ERP, and a Pr-Iall difference value was calculated. This process was completed 100 times and the 100 Pr-Iall difference values were placed into a distribution. If the Pr>Iall—meaning the difference value >0—in ≥85% of comparisons, a “knowledgeable” diagnosis was given. While a 90% cutoff is typical of research using semantic autobiographical probes, we selected an 85% criterion since probe knowledge here was gained during an episodic mock-crime, and was thus less well-rehearsed and so expected to produce smaller P300s than semantic details (e.g., name, phone number; Rosenfeld, Shue, & Singer, 2007).

The final measure of the CIT effect produced by the bootstrapping procedure is the difference between estimated probe and irrelevant mean amplitudes (“bootstrapped amplitude mean difference;” BAMD). As explained above, when bootstrapped probe and Iall mean amplitudes are computed, their differences are added to a distribution containing 100 difference values after 100 iterations of the bootstrapping process. The BAMD outcome is based on the probe and irrelevant amplitude values used to calculate each difference value in the distribution, which we used to estimate the population Pr-Iall mean difference.

While the number of iterations where the Pr>Iall is used for individual diagnosis, the BAMD measure can offer further information to aid in diagnostic decision making. To clarify the difference in these bootstrapping outcomes, consider two guilty Ps, both with 100% of Pr vs. Iall comparisons where the Pr>Iall. While both Ps would be binary “hits,” number of iterations where Pr>Iall fails to provide as much information about the degree of the CIT effect as the BAMD. For example, one may have a probe-Iall difference of 5 microvolts (µV) and the other 15µV, and both would produce 100% Pr>Iall iterations.

Intergroup Differences
Group differences in behavioral outcomes (i.e., RTs and error rates), post-experiment responses, and the CIT effect were also explored, as measured by probe-Iall voltages and the two bootstrapped variables described above, using analyses of variance (ANOVA) methods and t-tests. Partial eta squared ($\eta_p^2$) and Cohen’s $d$ are reported as estimates of effect size (Cohen, 1969; Cohen, 1988), and JZS Bayes Factors (BFs, scaled r=.707; Rouder et al., 2009; calculated at http://pcl.missouri.edu/bayesfactor) are reported as directly interpretable odds ratios, stating the likelihood that group means are truly different (favoring the alternative [alt.] hypothesis) or that they do not differ (favoring the null hypothesis). To illustrate, if BF=1.0, the null and alt. hypotheses are equally likely. Chi-square tests and Fischer’s Exact tests were also used to compare group hit rates.

Results

After excluding Ps who endorsed employing a CM other than RI (or any CM in the SG group; $n=12$), those who failed to follow instructions (e.g., got >1 recall wrong, made >½ target commission errors; $n=9$), and those with excessive artifacts ($n=6$) or technical/software errors (e.g., corrupted files, no detectable P300, >35 µV probe response; $n=12$), one hundred and ten data sets remained (see Figure 1; SG task and CIT immediate $n=19$; task immediate/CIT delayed $n=18$; task and CIT delayed $n=18$; RI task and CIT immediate $n=18$; task immediate/CIT delayed $n=20$; task and CIT delayed $n=17$). For the five Ps not diagnosed as knowledgeable with Pz P300 and who had Cz p-p amplitudes that were >1µV larger, Cz was instead used. Each probe (i.e., the necklace, bracelet, or watch) was used at least four times in each subcondition and >30

---

2 The benchmark criteria for $\eta_p^2$ is 0.01 for a small effect, 0.06 for a medium effect, and 0.14 for a large effect.
3 Bayes Factors not reported if $p>.2$
4 Fischer’s exact test was used when making comparisons between <5 groups and chi-square is reported for comparisons between all six subgroups.
times overall. Behavioral RTs and target error rates are reported, as well as P300 latencies and the following amplitude-related outcomes a) probe vs. Iall amplitude, b) number of bootstrapped iterations where the probe amplitude exceeds that of the combined irrelevants (“Iall”; probe>Iall bootstrapped iterations), and c) average difference between bootstrapped probe and Iall mean amplitudes (“bootstrapped amplitude mean differences”). Hit rates are also reported, and responses on post-experiment questionnaires are compared.

Table 1. Outcome measures of interest, presented by group (Simple Guilty [SG] and Retroactive Interference [RI]) and by combined groups. Probe (Pr) and Irrelevant (Iall) averages are displayed in milliseconds for response times (RTs) and latency, µV for amplitude, Pr-Iall, and bootstrapped (BS) Pr-Iall mean amplitude differences, and by average frequency for errors and number of BS iterations where Pr—Iall amplitude.

<table>
<thead>
<tr>
<th>Group</th>
<th>Task Timing/CIT Timing</th>
<th>n</th>
<th>RTs</th>
<th>Errors</th>
<th>Latency</th>
<th>Amplitude</th>
<th>Pr—Iall</th>
<th>Pr—Iall BS Iterations</th>
<th>Pr—Iall BS Ave. Amp. Diff.</th>
<th>Hit Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>Immediate/Immediate</td>
<td>19</td>
<td>3.63</td>
<td>Pr=522.32, Iall=513.26</td>
<td>Pr=17.56, Iall=11.2</td>
<td>6.36</td>
<td>91.21</td>
<td>6.13</td>
<td>73.68%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Immediate/Delay</td>
<td>18</td>
<td>4.54</td>
<td>Pr=477.32, Iall=462.21</td>
<td>Pr=15.67, Iall=10.71</td>
<td>7.19</td>
<td>93.39</td>
<td>7.2</td>
<td>94.44%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delay/Delay</td>
<td>18</td>
<td>5.34</td>
<td>Pr=448.22, Iall=439.39</td>
<td>Pr=16.72, Iall=9.11</td>
<td>7.61</td>
<td>92.36</td>
<td>7.36</td>
<td>94.44%</td>
<td></td>
</tr>
<tr>
<td>RI</td>
<td>Immediate/Immediate</td>
<td>18</td>
<td>4.94</td>
<td>Pr=595.33, Iall=563.78</td>
<td>Pr=17.91, Iall=9.5</td>
<td>8.41</td>
<td>90.78</td>
<td>7.8</td>
<td>88.89%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Immediate/Delay</td>
<td>20</td>
<td>6.45</td>
<td>Pr=425.65, Iall=413.1</td>
<td>Pr=17.75, Iall=11.46</td>
<td>6.29</td>
<td>89.6</td>
<td>6.65</td>
<td>90.00%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delay/Delay</td>
<td>17</td>
<td>5.33</td>
<td>Pr=368.18, Iall=368.65</td>
<td>Pr=14.65, Iall=9.28</td>
<td>5.36</td>
<td>89.65</td>
<td>5.15</td>
<td>76.47%</td>
<td></td>
</tr>
<tr>
<td>SG + RI</td>
<td>Immediate/Immediate</td>
<td>37</td>
<td>4.27</td>
<td>Pr=429.92, Iall=418.76</td>
<td>Pr=17.73, Iall=10.37</td>
<td>7.36</td>
<td>91</td>
<td>6.94</td>
<td>81.08%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Immediate/Delay</td>
<td>73</td>
<td>4.07</td>
<td>Pr=423.15, Iall=415.4</td>
<td>Pr=16.83, Iall=10.21</td>
<td>6.62</td>
<td>91.49</td>
<td>6.61</td>
<td>89.04%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Immediate/Delay</td>
<td>38</td>
<td>4.37</td>
<td>Pr=435.42, Iall=424.95</td>
<td>Pr=17.85, Iall=11.14</td>
<td>6.71</td>
<td>91.39</td>
<td>6.91</td>
<td>92.11%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Delay/Delay</td>
<td>35</td>
<td>3.74</td>
<td>Pr=409.83, Iall=405.03</td>
<td>Pr=15.71, Iall=9.2</td>
<td>6.52</td>
<td>91.6</td>
<td>6.29</td>
<td>85.71%</td>
<td></td>
</tr>
</tbody>
</table>

Each dependent measure underwent a 2x3 between-subjects ANOVA with group (SG vs. RI) and time delay (task and CIT immediate; task immediate/CIT delayed; task and CIT delayed) as the two factors. Next, like Gronau et al. (2015), we conducted two sets of planned contrasts.
The first set further examined the main effect of time delay by combining the SG and RI group
and comparing a) the immediate conditions (task and CIT) with the two delay conditions
combined (task immediate/CIT delayed and both task and CIT delayed), and b) the two delay
conditions to each other. The second set of contrasts was carried out to explore whether
differences in results related to the time delay conditions varied based on group (SG vs. RI).
Because we did not have clear directional hypotheses about how the effects of the RI
manipulation and time might impact our findings, two-tailed tests are reported.

Behavioral Results

Response Times To examine how the RI manipulation and time delay effects might have
impacted behavioral RTs, we conducted a 2 (group SG vs. RI) X 3 (time delay task and CIT
immediate, task immediate/CIT delayed, both task and CIT delayed) X 2 (stimulus probe vs.
Iall) mixed ANOVA. While those in the SG group took longer to respond and there was a main
effect of group (see Figure 3; \( F(1,104) =7.745, p=.006 \)), favoring the alternative (BF=4.093) and
reflecting a medium effect size (\( \eta^2=.069 \)), both the main effect of time (\( F(2,104)=.391, p=.677, \)
\( \eta^2=.007 \)) and the interaction of group and time (\( F(2,104)=.872, p=.421, \eta^2=.016 \)) were not
significant. Results also showed a main effect of stimulus type (Pr=425.43 vs. 416.53ms;
\( F(1,104)=9.565, p<.01, BF=9.291, \) alt.), with a moderate effect size (\( \eta^2=.084 \)). However, the
interactions between stimuli and group (\( F(1,104)=.508, p=.478, \eta^2=.005 \)) and between stimuli
and time delay (\( F(2,104)=.503, p=.606, \eta^2=.01 \)), were not significant. This was also the case for
the triple interaction (stimuli X group X time) (\( F(2,104)=.566, p=.569, \eta^2=.011 \)).

Given the significant main effects of group and stimulus, follow-up tests were conducted
to further examine how the RI manipulation might have differentially impacted probe and
irrelevant RTs. Results showed that both probe (SG=457.64ms vs. RI=392.22ms;
Probe responses took longer than irrelevant responses, that probe and Iall RTs were longer in the SG compared to the RI group, and that time delay did not impact RTs.

Figure 3. Probe (Pr) and combined irrelevant (Iall) response times, presented by group (Simple Guilty [SG]; Retroactive Interference [RI]).

Target Response Errors Across subconditions, average target response errors were low (see Table 1) and ranged from about 3 to 6, suggesting no group or timing differences in target response errors. Target response errors were examined with a 2 (group) x 3 (time delay) ANOVA (see Table 1 for averages). As expected, both main effects failed to reach significance (group $F(1,109) = 1.499, p = .224, \eta^2 = .014$; time delay $F(2,109) = .197, p = .822, \eta^2 = .004$), as did their interaction ($F(2,109) = .818, p = .444, \eta^2 = .015$).

ERP Qualitative Results
Grand average ERPs at Pz are presented in Figure 4. In columns 1-3, the two top rows show averages for each SG and RI subcondition (6). In the final columns of the top two rows, ERPs representing the two combined conditions that received the CIT on their second visit (see column 4; task immediate/CIT delayed, task and CIT delayed) are displayed, alongside the grand average for each group, which included all three timing subconditions combined. In the bottom row, ERPs reflect collective responses of both the SG and RI groups. Across groups, an average of 27 probe sweeps were analyzed. As expected since all Ps completed the mock-crime, visual review of each grand average shows an obvious CIT effect. However, since both probe and irrelevant amplitudes appear roughly similar in each, the CIT effect does not appear to be particularly pronounced in any of the subconditions or their combinations. The grand averages below also show that probe and lall latencies and morphology remained consistent, regardless of group and timing delay.
Figure 4. Grand average ERPs at Pz. Results are presented by group and with both groups combined (Simple Guilty [SG]; Retroactive interference [RI]), as well as by time delay condition (I=Immediate; D=Delay). ERPs in column 4 represent both time delay conditions in which the CIT was delayed, while column 5 represents all time delay conditions, by group.

<table>
<thead>
<tr>
<th>Time Delay (Task/CIT)</th>
<th>I/I</th>
<th>I/D</th>
<th>D/D</th>
<th>I/D+D/D</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
</tr>
<tr>
<td>RI</td>
<td><img src="image6" alt="Graph" /></td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
</tr>
<tr>
<td>SG+RI</td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
<td><img src="image13" alt="Graph" /></td>
<td><img src="image14" alt="Graph" /></td>
<td><img src="image15" alt="Graph" /></td>
</tr>
</tbody>
</table>
**ERP Quantitative Results**

*Probe and Irrelevant Amplitudes* Across subconditions, probe-irrelevant amplitudes—the CIT effect—ranged from 5.36µV (in the RI group with both task and CIT delayed) to 8.41µV (in the RI group with both task and CIT immediate) (see Table 1). As shown in Figures 3 and 4, while probe>Iall in both the SG and RI groups and in all six subconditions, probe and irrelevant p-p amplitudes were similar. For the SG group, average probe and Iall amplitudes were 17.42µV and 10.38µV respectively, and for the RI group, the probe amplitude was 16.84µV and Iall amplitude was 10.15µV. To investigate the CIT effect, a 2 (group) X 3 (time delay) x 2 (stimulus type) ANOVA was conducted on P300 amplitudes. Results confirmed a main effect of stimulus type, and that the probe was much larger than Iall (Pr=17.13µV vs. Iall=10.26µV; F(1,104)=231.393, p<.001, η²=.69, BF=3.318815e+25, alt). However, neither main effect of group (F(1,104)=.157, p=.693, η²=.002) or time condition (F(2,104)=1.103, p=.336, η²=.021) reached significance. All interactions between group, time delay, and stimulus type also failed to significantly impact amplitudes The group X time (F(2,104)=.093, p=.911, η²=.002), stimuli X group (F(1,104)=.166, p=.684, η²=.002), and stimuli X time (F(2,104)=.348, p=.707, η²=.007) interactions were not significant, as was the case for the triple interaction (F(2,104)=1.954, p=.147, η²=.036, BF=3.677, null). Taken together, results show that probes>Iall in each of the six subgroups, that SG and RI probe and irrelevant amplitudes were similar, and that amplitudes did not vary by subcondition.

*Figure 5.* Probe (Pr) and combined irrelevant (Iall) P300 amplitudes, presented by group (Simple Guilty [SG]; Retroactive Interference [RI]).
Latency

Average latencies and are displayed in Table 1 and plotted in Figure 6. Another 2 (group) X 3 (time delay) X 2 (stimulus type) ANOVA was conducted on P300 latency, and no meaningful latency differences were found between groups ($F(1,104)=2.527, p=.115, \eta^2=.024, BF=2.796, \text{null}$) or based on time delay condition ($F(2,104)=1.662, p=.195, \eta^2=.031, BF=4.3, \text{null}$). This was also the case for the group X time interaction ($F(2,104)=1.986, p=.142, \eta^2=.037, BF=3.62, \text{null}$). Across groups, probe and Iall P300s occurred at 563 and 559ms respectively, and their latency differences were not significant ($F(1,104)=.367, p=.546, \eta^2=.004$). Average probe latency was 548ms for the SG group and 578ms for the RI group, and for Iall, latencies were 546ms and 572ms for the SG and RI groups respectively. Results showed that latency was uninfluenced by group ($F(1,104)=.137, p=.712, \eta^2=.001$), time delay ($F(2,104)=2.281, p=.107, \eta^2=.042, BF=3.144, \text{null}$), or their interaction ($F(2,104)=.407, p=.667, \eta^2=.008$).

Figure 6. Probe (Pr) and combined irrelevant (Iall) P300 latencies, presented by group (Simple Guilty [SG]; Retroactive Interference [RI]).
Bootstrapping Outcomes The number of bootstrapped iterations where probe>Iall and bootstrapped amplitude mean differences (BAMD) appear in Table 1 and were analyzed using 2 (groups) X 3 (time delay) ANOVAs. The SG and RI groups averaged 93 and 90 probe>Iall iterations, respectively, and the evidence again suggests an obvious CIT effect in each of the six conditions, with probe>Iall iterations ranging from 90 to 93. Again, findings failed to show main effects of group ($F(1,109)=1.415, p=.237, \eta^2=.013$) or time delay condition ($F(2,109)=.025, p=.976, \eta^2<.001$), or their interaction ($F(2,109)=1.842, p=.164, \eta^2=.033, BF=3.879, \text{null}$).

The bootstrapped amplitude mean difference measure was subjected to the same analysis as above and also failed to show significant effects of group ($F(1,109)=.196, p=.659, \eta^2=.002$), time delay condition ($F(2,109)=.308, p=.736, \eta^2=.006$), or their interaction ($F(2,109)=1.842, p=.164, \eta^2=.033, BF=3.879, \text{null}$).
Hit Rate Results

Detection accuracy sensitivity was 87% in the SG group vs. 85% in the RI group, for an overall 86% hit rate. Of the 15 total misses, 10 had >75 probe-Iall bootstrapped iterations. As displayed in Table 1, across subconditions, hit rates ranged from 74% (in the SG task and CIT immediate condition) to 94% (in both the SG task immediate/CIT delayed and the task and CIT delayed conditions). A chi-square test was carried out to determine if the observed frequency of misses varied significantly from what would be expected if misses were distributed evenly across the subgroups, and results showed no association between subcondition and diagnostic accuracy ($\chi^2[5, N=110]=6.325, p=.29$).

Contrasts

Planned comparisons were then performed to further explore the possible effect of time delay and its interaction with group (SG vs. RI). The contrasts designed to assess the main effect of time failed to produce statistically significant results. In the first comparison, no differences were found between the group that received both the task and the CIT in one session compared to those who were administered the CIT at a second session, and $p$-values for all dependent variables of interest >.1. When comparing the Ps in the two delay conditions that completed the CIT on their second visit (task immediate/CIT delayed vs. task and CIT delayed), the sole outcome with $p<.1$ was Iall latency. However, this finding fell short of significance ($t(5,104)=1.884, p=.062 \ d=.459$) and was not supported by BF (=1.72, null).

The two contrasts designed to examine the interaction between group and time delay also fell short of significance. Comparison of the task/immediate to the two CIT delay conditions revealed only one outcome with $p<.1$, suggesting that probe latency tended to be longer ($t(5,104)=1.819, p=.072 \ d=.082$). However, the BF favored the null hypothesis nearly two to one
(BF=1.928, null) in the delayed conditions, suggesting no difference between probe and Iall latency. Finally, the direct comparison of the two delay conditions (task immediate/CIT delayed vs. both task and CIT delayed) showed no differences in any of the outcome variables of interest (all \( p \)-values>.2). Using Fischer’s exact test to assess hit rates, no differences were found between the immediate and combined delayed conditions (\( p=.2563 \)) or between the two delayed conditions (\( p=.4681 \)). Thus, the evidence provided by the planned comparisons offers no evidence that the main effect of time delay or the interaction of group and time delay significantly impacted outcomes.

**Post-Experiment Questionnaire**

To further explore possible explanations for our ERP findings and the group differences in RTs, we conducted \( t \)-tests on post-experiment questionnaire responses. While the two groups reported no differences on their level of focus on the task while completing it \((t(1,91)=1.324, p=.189, d=.278, BF=3.752, \text{null})\), during the CIT, the RI group endorsed focusing more on their task \((t(1,91)=3.707, p<.001, d=.77, BF=58.239, \text{alt})\), but also thinking about their real crime more often when presented with the probe \((t(1,91)=2.535, p=.013, d=.524, BF=2.349, \text{alt})\). Although both groups reported being equally motivated to beat the test \((t(1,91)=.246, p=.806, d=.045, BF=8.469, \text{null})\), the RI group also reported being more confident in succeeding \((t(1,91)=2.589, p=.011, d=.535, BF=2.663, \text{alt})\).

Because the only crime-related stimuli that Ps are exposed to during the testing phase are the stolen items, it is possible that the memory of less central details might be impacted by the RI manipulation in ways that are undetected by the CIT. To further investigate how the RI manipulation impacted memory of the crime in general over time, we compared subconditions that completed the CIT after a delay on how much they believed they forgot about the real crime
and on how many key details (e.g., room number of crime) they actually misremembered. A 2 (group SG vs. RI) X 2 (time delay task immediate and CIT delayed vs. both task and CIT delayed) ANOVA revealed that although there was no significant effect of group on the amount of detail Ps endorsed forgetting by their second appointment ($F(1,55)=.254, p=.616, \eta^2=.005$), there was an effect of delay, such that those who completed the task immediately reported thinking they had inferior memory for crime details than those who completed the task after a delay ($F(1,55)=8.476, p=.005, \eta^2=.134, BF=6.339$, alt). The interaction between group and time delay was not significant ($F(1,55)=.628, p=.432, \eta^2=.011$). Despite thinking they had forgotten more crime details however, those who competed the task immediately did not actually forget more than their counterparts who completed their task a week later ($F(1,55)=.502, p=.482, \eta^2=.009$). There were also no differences in the number of crime details forgotten by the second session between the SG and RI groups ($F(1,55)=.139, p=.71, \eta^2=.003$), and the interaction also failed to reach significance ($F(1,55)=.425, p=.517, \eta^2=.008$). Thus, although Ps who completed their task during their first appointment thought they had forgotten more crime details by their second visit than those who completed their task immediately preceding the CIT, they had not actually done so. Additionally, the RI manipulation seemed to have no impact on Ps’ judgments about how much they had forgotten, or on their actually memory performance.

Discussion

Because memory of the crime relevant details is necessary for an effective CIT, our goal was to determine how factors related to memory quality—namely, quality over time and especially competing memories—might impact the validity of the P300-based CIT. Our central question concerned whether an RI manipulation, which was intended to degrade the memory of the original crime and involved learning false information, might reduce detection of crime-
related memories using the CM-resistant, P300-based CTP. Additionally, we explored how administering the RI manipulation and CIT at varying times might impact results.

Behavioral RTs and target error rates were examined, along with latency and amplitude-based indices of the CIT effect (probe vs. Iall amplitude, number probe>Iall bootstrapped iterations, BAMD, and hit rates). Because all Ps completed the mock-crime and P300 is a reliable index of recognition, we expected a pronounced CIT effect and high hit rates across the board. Since we believed memory traces would remain strong enough for Ps to recognize probe items, regardless of the RI manipulation or when the task and CIT were administered, we did not expect outcomes to differ between subconditions. Findings confirmed a CIT effect in both the SG and RI groups, as well as for each subcondition that varied in terms of when the memory manipulation task and CIT were conducted. While the CIT effect was prominent in each of the six subgroups, none of the amplitude-based outcomes varied between them or were affected by time, as was the case for P300 latency, target response error frequency, and hit rates. The only CIT outcome of interest to differ between groups was RTs, which were longer for both the probe and Ialls in the SG group compared to the RI group and did not vary based on time-delay condition. Hit rates were 86% overall using a diagnostic criterion of $\geq 85\%$ probe>Iall bootstrapped iterations, and among the 15 misses seen, 5 had $\geq 80\%$ probe>Iall iterations, and another 5 had probe>Iall bootstrapped iteration scores between 75%-80%.

As previously mentioned, a competition hypothesis would predict that the newly introduced RI memory would compete with and degrade the original crime memory. If this were the case, P300 reductions should be probe-specific since it was the only item presented during the CIT that was replaced with a false detail during the RI task. Additionally, a competition hypothesis would predict reduced saliency for the probe in the RI group compared to the SG
group, which would likely be reflected in shorter probe P300 latencies and behavioral RTs.

Given that probe and Iall amplitudes and latencies were nearly equal across conditions, the ERP data do not provide support that RI created competing memories. While RTs were shorter overall in the RI group, this was not specific to the probe, and so the behavioral data also do not support that the RI task competed with the memory of the real probe.

In contrast, an enhancement hypothesis, which predicts an increased focus on the probe, would be supported by elevated probe responses in the RI group, possibly through linked associations. For example, when told to imagine they stole cufflinks instead of a watch, Ps may form an association between the items, such that when the watch appears during the CIT, they automatically think of the crime they committed as well as the imagined crime. Because the association adds meaning to the probe, its response would be expected to increase, along with latencies and RTs, which are typically longer under increased cognitive load (as when considering both crimes during the CIT). While amplitude-based outcomes do not seem to support the enhancement hypothesis, post-experiment questionnaire responses offer possible interpretations for our ERP findings.

RI Ps reported thinking more often about both their RI task and the real crime than the SG group during the CIT, which in light of paired-association research, could be taken as support that Ps formed a cognitive link between the real and imagined probes. However, the ERP evidence does not suggest enhanced probe amplitude, latencies, or RTs in the RI group. While we originally assumed increased focus on the RI task during the CIT would lead to an increased probe response through linked associations, this would only be expected if Ps endorsed focusing specifically on the RI probe. Instead of asking if they focused on the RI probe specifically, we instead asked how often Ps thought of the imaginary crime in general. Since the RI group
endorsed thinking of both their assigned task and the real crime more often than the SG group, one possibility is that they thought of their RI crime in a general way, as opposed to picturing the interference probe specifically. If this were the case, the expected amplified response associated with thinking of the real crime more often when exposed to the probe might be ameliorated by thoughts of the RI crime, which might have weakened memory traces or reduced P300 amplitude due to the dual-task of focusing on the crime and the CIT (Donchin, Kramer, & Wickens, 1986; Polich, 2007). A subtly different alternative explanation for our ERP findings is that RI Ps did draw cognitive associations between the real and RI probe that resulted in increased probe amplitude, but that these increases were counteracted by the cognitive load imposed by considering both crimes while completing the CIT task.

Although P questionnaire responses help shed some light on ERP findings, they also raise questions regarding our RT results. Because the questions regarding focus during the CIT imply that RI Ps were more focused on their interference task than SGs were on their control task when performing the CIT itself, it is puzzling that their RTs are also shorter, since divided attention typically increases RTs (e.g., Ninio & Kahneman, 1974). While RT differences serve as a manipulation check by showing that the RI task did indeed produce differences in the experimental group, the underlying causes for the difference are unclear. One possibility is that, while RI Ps endorsed thinking more about their task during the CIT, these interference memories actually increased their attention toward the real crime, which is supported by their reporting that they thought more about the real crime during the CIT. In theory, the linked associations between the imagined and real crime details in the RI group might function to increase their attention toward stimuli. While this increased attention might be reflected in their shorter RTs,
this increase would also produce larger P300s to the probe, that might be counteracted by the increased cognitive effort involved in simultaneously considering both crimes.

Limitations and Future Directions

During both the RI and SG task instructions, Ps are instructed that focusing on their task might help them beat the brainwave-based test they would later take. However, the instructions were not clear as to whether they should focus on the task only while completing it or if they should focus on their task during the CIT as well. While the RI group reported focusing significantly more on their task of remembering the alternate crime than the SG group focused on their control task while completing the CIT, it might more closely approximate real-life circumstances to compare the RI group to a control group that did not receive a filler task upon which they might focus during testing.

Another factor that may compromise the generalizability of our findings is the limited time between the mock-crime and CIT administration. In the field, although subjects may be interviewed in the week following the crime in question, the delay is often much longer, and could be many months or even years. Although the CTP has proven effective when administered even a month after a crime (Hu & Rosenfeld, 2012) and there is no evidence here that the time delay degraded memories of the original crime, administering the CIT long after the crime may produce different results. For example, delaying the CIT by several months might degrade memories further due to natural decay or because other interfering events (e.g., other thefts) might occur.

In the current study, Ps were instructed to rehearse the RI task only during their session. However, research showing that recall of an existing memory decreases as a function of how frequently a competing memory is rehearsed over time (e.g., Slamecka, 1960) suggests that
instructing Ps to repeatedly rehearse their RI crime might produce differences between the SG and RI groups. Thus, to more rigorously test the viability of the RI interference CM as it might be used in the field, future work should instruct Ps to rehearse their RI crime many times over the course of several days or weeks to establish how practice and longer delays between the crime and CIT might impact findings. While we found no differences here between Ps in the conditions that completed the CIT during a second session and either imagined the RI crime at the previous session or immediately before the CIT, these findings might also change if the delay between the task and CIT were longer, and may also be influenced by rehearsal. For example, if someone committed a crime and rehearsed it repeatedly, and then imagined a RI crime, results could be much different than someone who committed a crime, and then immediately thought of and repeatedly rehearsed an alternate crime.

An additional consideration that might limit the generalizability of our results is that RI Ps were asked to only imagine their alternate crime. Given that memories for sensorimotor rich events that have actually been carried out are stronger than for situations that have only been imagined (Johnson et al., 1988) and that actual criminals often have real criminal experiences from which they can draw, future work might consider having Ps physically carry out their alternate task or instruct them to imagine a task they actually completed. Relatedly, since criminals have often completed several offenses, future work is needed to determine how multiple sources of RI might impact findings. Additionally, research suggests that memories are often better recalled in the context in which they are learned (Godden & Baddeley, 1975), and although the mock-crime and CIT were completed in different offices, they were still within the same lab space, which could have impacted results. Future work might compare conducting the mock-crime and CIT in the same location, as is typically done in lab studies, to conducting the
crime and the CIT in separate locations, as in applied settings, to determine if this factor impacts test results.

Conclusions

Overall, our findings support that different mechanisms underlie ERP, ANS, and behaviorally-based measures during the CIT. Regarding RTs, our findings, that responses were impacted by the RI manipulation while ERPs were not, support previous work suggesting their nonoverlapping mechanisms for memory detection (Hu & Rosenfeld, 2012). In Gronau et al. (2015), while respiration line length was unaffected by RI, skin conductance responses were attenuated, as in other studies demonstrating the resilience of respiration compared to electrodermal measures when mental CMs were employed (Ben-Shakhar & Dolev, 1996; Honts et al., 1996). One factor that limits the comparability of ANS and P300-based CITs is their inherent difference in questioning technique. For example, Gronau and colleagues (2015) presented Ps with five questions about their crime (e.g., jewelry stolen, envelope color), followed by five possible answers (e.g., necklace, bracelet, ring, etc.), while we asked only one question (i.e., what jewelry item was stolen?). While we used the same RI technique that targeted each of the probe items Gronau (2015) tested—telling RI Ps to imagine a different jewel, office, envelop color, etc.—the testing for multiple probes may have enabled the electrodermal RI effect in their study since Ps’ focus may been spread across remembering multiple interference probes, which would presumably draw attention away from the true probe items. In our study, although Ps imagined that they stole a different item from a different office out of a different color envelope, they may not have focused on the imagined office or envelope color as much during the subsequent CIT, during which they were presented only with jewelry items. One possibility is that the other peripheral details (e.g., office number, envelope color) replaced during the RI
manipulation were no longer in the forefront of their memory, and were therefore not as influential as in the ANS-based CIT.

Taken together, while the theoretical basis for the disparity between our RT and ERP findings remains unclear, the evidence suggests that the RI manipulation tested here does not pose a threat to the diagnostic accuracy of the CTP. However, before RI can be discounted as a CM technique, it should first be tested under more optimal conditions, including after an extended time delay, with more rehearsal, and using real sensorimotor rich RI memories. Taken together, the results presented here suggest that, unlike RTs and some autonomic measures (e.g., SCR), P300—a good sign of recognition—is unaffected by RI. In sum, our findings underscore the CM resistance of the CTP and provide further support for its field use.
CHAPTER 2 A PRELIMINARY TEST OF THE “SEARCHING” P300-BASED CONCEALED INFORMATION TEST (CIT) UNCOVERING SPECIFIC DETAILS FROM MORE GENERAL INFORMATION USING THE COMPLEX TRIAL PROTOCOL (CTP)

Background

The Concealed Information Test (CIT)

The concealed information test (CIT; Lykken, 1959) has garnered more empirical support than other methods of recognition detection and has a firm theoretical foundation (Verschuere & Ben-Shakar, 2011; Ben-Shakar, 2012). This approach works by identifying who has a differential physiological response to a critical item of interest (a “probe”) when it is presented among other similar items not related to the crime (i.e., irrelevants); because guilty individuals recognize the probe, they react differently to it, whereas innocent individuals with no particular memory of the relevant item react similarly to all stimuli.

P300-Based CIT

Although the CIT has been tested using a variety of autonomic nervous system (ANS) measures, P300-based protocols have proven the most efficacious (Ben-Shakhar, 2012; Meijer et al., 2014). Because the P300 event-related potential (ERP) is evoked involuntarily when a rare and meaningful stimulus (i.e., a probe item) is recognized (Johnson, 1988), this approach is preferable to measures that are more easily controlled by those attempting to beat the test by applying countermeasures.

Three Stimulus Protocol (3SP)

The original P300-based CIT—the three stimulus protocol (3SP)—exposed Ps to either a probe, an irrelevant, or a “target” on every trial. Since probes and irrelevants require the same response, the target becomes meaningful because it requires a unique response. While the 3SP
can successfully diagnose cooperative Ps, applying countermeasures (e.g., biting one’s tongue) during an irrelevant stimulus increases its response and narrows the probe-irrelevant difference—the “CIT effect”—upon which diagnosis is based (Rosenfeld et al, 2004; Mertens & Allen, 2008). The 3SP’s design also consumes undue cognitive resources, thereby suppressing P300 and detection accuracy, by forcing a target decision on every trial (Donchin, Kramer, & Wickens, 1986; Polich 2007).

*Complex Trial Protocol (CTP)*

The Complex Trial Protocol (CTP; Rosenfeld et al., 2008; see Figure 2 in Chapter 1) was designed to address the shortcomings of the 3SP, and is called such because it divides each trial into two parts in the first, Ps are exposed to either a probe or an irrelevant and make the same left-hand response for both, simply acknowledging that it was seen. Next, either a target (e.g., “11111”) or a nontarget (e.g., “22222,” “3333”) appears, each requiring a different right-hand response. Separating the implicit probe/irrelevant judgment from the behavioral target/nontarget decision has been hypothesized to maximize the probe’s P300 response, facilitating recognition detection.

*The Searching Concealed Information Test (SCIT)*

In a standard CIT, investigators know which item is the probe and has the goal of determining which suspect recognizes it. However, in cases when the crime-related item is unknown, a standard CIT is not possible. Take for example, a scenario in which a known terrorist is withholding information about where their base of operations is located, or a case when someone admits to having planted a bomb but will not reveal where. In cases when investigators are confident a suspect is concealing information but not of its specifics, a “searching” CIT (SCIT) may help identify critical details.
In a SCIT, a subject is presented with possible alternatives, as in a standard CIT, but instead of identifying knowledge based on an enhanced response to a known probe, the item that evokes the largest response is assumed to be the probe. One limitation of this approach is that it depends on the probe being included in the stimulus set, and so a false negative result is likely in cases when it is not. However, when investigators have enough information to be confident that the probe is included, the SCIT can be a useful tool for uncovering crime-related information.

While the searching version of the CIT has great applied potential and is currently used by Japanese police, it has not received extensive empirical attention. However, the limited existing research suggests the potential of the test when utilized this way. For example, Meijer et al. (2010) reported that, using skin conductance (SCR) group averages, they could correctly identify three critical items in groups that had either planned or completed a mock-crime. Meixner and Rosenfeld (2011) expanded on this research by instead attempting to identify critical items at the individual level using the P300 measure. With limited a priori knowledge of the probe, they correctly identified 10 of 12 knowledgeable suspects (with no false positives) and 21 of 36 items of interest by comparing each P’s two largest responses. While these studies demonstrate how details that might help prevent crimes can be detected without knowledge of probe-specifics, the stimulus sets used were known to include the critical item.

In an effort to improve the generalizability and applied value of the SCIT, we modeled real-life situations when investigators have limited information from which to form CIT stimulus sets by selecting stimuli that allow a maximum number of possible alternatives to be tested. Thus, the aim of the current study is to test the accuracy of the CTP when used to identify specific crime-relevant details after first narrowing down possible alternatives from a broader set of possibilities.
Method

Participants

Forty-three participants (Ps) were recruited from the Northwestern Psych110 and paid participant pools and were randomly assigned to either a simple guilty (SG) or an innocent control (In) group. A third of the Ps completed the experiment at a satellite lab location. All Ps had normal or corrected to normal vision, reported no neurological or psychological abnormalities, consented to participation, and were paid $10.

Mock-Crime/Innocent Control Task

Ps completed either a mock-crime or innocent control task (see Appendix E), followed by three SCIT blocks, and a post-experimental questionnaire (see Appendix H). SG Ps received written mock-crime instructions detailing their task after taking a jewelry item from a desk in the lab, they were to enter another office, where they would select and open one of the seven large containers (e.g., briefcase, duffle bag; see Appendix G) they would see. There, they would find seven small containers (e.g., gift box, jar, etc.), one of which they should select to place the jewelry inside (see Appendix E for illustration). Finally, Ps put the small container (now holding the jewelry), back inside the larger container. Innocent Ps instead opened the desk drawer to find a piece of paper and a pen with which to sign their name. After their task, all Ps returned to the testing room for the SCIT.

Data Acquisition

Ps were seated about a meter from the stimulus display screen, and electrodes were

5 The second lab location was included to aid in recruitment efforts. Participants from the two locations did not differ significantly on reported outcome variables.
6 Probe item was randomly assigned (necklace, watch, or bracelet) and the final dataset included at least five of each.
applied as Ps read the SCIT instructions. EEG was recorded with Ag/AgCl electrodes attached midline at Fz, Cz, and Pz, sites, electro-oculogram (EOG) was recorded referentially with an electrode above the left eye to collect eye movements and blinks, and were grounded to a forehead electrode, referenced to linked mastoids. Impedance remained under 5 kΩ, blink artifacts were corrected using the Semlitsch method (Semlitsch et al., 1986), and trials with artifacts >90 µV were excluded. The 19 channel Mitsar amplifier (Model 201) used a 30 Hz low pass and a 0.16 Hz high pass filter setting, and output passed through a 16-bit Mitsar A/D converter sampling at 500 Hz. For display and analyses, a Kaiser (alpha= 1.8) filtering algorithm (with the digital filter passing frequencies from 0-6 Hz), was used off-line to remove higher frequencies in single sweeps and averages.

Complex Trial Protocol (CTP) Blocks

Ps in both conditions received the same instructions, which explained the premise of the test and how they were to respond behaviorally. As detailed in Appendix F, Ps read that a jewelry item was stolen and hidden in the lab, and that the brainwave-based test was designed to tell us who took the item and where they put it by showing us who recognized key items. Innocent Ps read that their goal was to be determined innocent by the test, while guilty Ps were told to try to beat the test, but were not given a strategy to try.

After confirming they understood how to respond to the stimuli using the mice and being told they would be stopped randomly and asked to recall the last item they saw in order to ensure their attention, Ps were shown all the stimuli so that they could identify all the items in the test to follow. Each item, accompanied by a label stating its name, was then displayed for 5 sec. After seeing the series of images, Ps who responded “no” when asked if they could identify what each
item was were shown the series of images again. Each stimulus set for a given block was previewed just before that block.

This pretest stimulus exposure step was included to help control for possible effects caused by confusion, which could impact results in unpredictable ways. Because stimuli are presented so quickly, guilty Ps might not recognize the probe image as the item they stole, or items might be misidentified, especially by innocent Ps who cannot draw from an actual memory of the mock-crime. Additionally, if a P is worried about being quizzed after an ambiguous item, they might focus on it to try to figure out what it is, and the resulting probe-like response could reduce the critical probe-irrelevant amplitude difference (Donchin, Kramer, & Wickins, 1986) in guilty Ps and also lead to false positives. Guilty Ps', worry over failing a quiz might also produce cognitive load, which could also reduce the probe’s P300. Previewing stimuli might also prevent Ps from forming unintended associations between items. For example, if someone thinks they are seeing two bracelets among other distinct pieces of jewelry when one of the items is actually a necklace, their association between them might lead the pair to stick out, causing P300s for both and/or diverting attention away from the real probe.

In an attempt to establish who had knowledge of the lab crime, stimuli in SCIT Block 1 consisted of photographs of jewelry items. This models real-life circumstances where the first step would be to confirm a suspect’s knowledge of a critical item that is already known to officials. The next two blocks tested for knowledge of the “core probe’s” (i.e., the jewelry’s) hiding spot, narrowing down from the larger (Block 2) to the smaller (Block 3) container. Each block consisted of 210 trials and ran about 15 minutes.

Images (on a white background) of jewelry and the containers, the target (“11111”), and the nontargets, (“22222”…”55555”) were presented on a black screen. Each trial (see Figure 2 in
began with a fixation cross “+” in the middle of the screen (200 ms), followed by either a probe or irrelevant (300 ms), and an observation period (1100 ms). After a randomly varying interstimulus interval (50-200 ms with a fixation “+” presented), this sequence was repeated in the second part of the trial, using targets and nontargets (at a 20%-80% ratio). To prevent Ps from establishing patterns, which would draw their attention away from the probe, the target and each of the four nontargets followed each irrelevant and the probe with the same frequency. As described above, Ps made a single left-hand “I saw it” response for probes and irrelevants, followed by right-hand response that varied for targets and nontargets.

Post-Experiment Questionnaire

After the CIT, a post-experiment questionnaire inquired whether Ps attempted countermeasures and how confident they were that the test could correctly determine their guilt or innocence. To evaluate how well Ps’ subjective judgments (of what items were particularly salient for them) matched their objective P300 outcomes, we also asked Ps if any of the stimuli stood out to them among the others, and if so, why.

Data Analysis Plan

Known and Blind Probe Analyses

Key dependent variables were calculated using both a “known probe” and a “blind probe” approach. Analyses were first conducted as in a standard CIT, comparing the known probe to Iall, to confirm acceptable accuracy rates using this vetted procedure. Because field investigators should establish recognition of an initial item of interest (i.e., the core probe) before probing for further details, Block 1 was analyzed only using the “known probe” approach. Next, using the “blind probe” approach, we analyzed the data as if we were blind to where the jewelry was hidden, treating the stimulus with the largest P300 as the probe and comparing it to the
combined remaining stimuli, assumed to be irrelevants. For innocent Ps, the irrelevant to be treated as the probe for analyses purposes was randomly selected\(^7\).

Given the high diagnostic accuracy of previous studies probing for similar information when the probe was known to researchers and when Ps completed multiple blocks (Rosnefeld et al., 2008; Ward & Rosenfeld, 2017), we expected high accuracy (e.g., >80% overall diagnostic accuracy over the three blocks) using the known probe method. However, it is less clear what the results of the blind probe analyses might reveal. Take for example, a guilty P with an elevated response to both the probe and an irrelevant When the probe is known, it is compared to the randomly occurring large irrelevant combined with several (five here) others, its heightened response might be compensated for by one of the smaller irrelevants, allowing the probe to not be called. However, using the blind probe approach, the large irrelevant would be mislabeled as the probe, and if it were enough larger, it could result in a false probe identification within a guilty individual. Likewise, an enhanced irrelevant response might be mitigated by other irrelevants in an innocent P, but when selected as the probe, might result in a false positive.

*Behavioral Response Times (RTs) and Response Errors*

Given that response times (RTs) and errors are often elevated for probes and when cognitive resources are depleted, as when applying countermeasures, (Verschuere, 2011; Rosenfeld & Labkovsky, 2010; Gronau, Ben-Shakhar, & Cohen, 2005) these two behavioral measures were explored. RTs were examined by block, using both the known and blind probe methods. Three types of errors were explored omission errors (i.e., no target response was made), commission errors (i.e., an initial incorrect target response was corrected), and regular

\(^7\) Using the known probe approach for Block 1, irrelevant items assigned probes were the same as in the SG group. For Blocks 2 and 3, three stimulus sets were randomly assigned.
errors (i.e., an incorrect target response that was not corrected). Because probes and irrelevants require the same response, errors were based on target responses during the second part of the trial, and were analyzed by block.

**ERP Analysis Plan for ERP-based CIT**

P300 was evaluated at Pz, where it is typically largest (Johnson, 1993), using the peak-to-peak (p-p) measurement method, which has been shown to produce results that are at least 25% more accurate than baseline-to-peak calculations (b-p; Soskins, Rosenfeld, & Niendam, 2001). To calculate p-p amplitudes, the algorithm searched for the maximally positive 100ms segment occurring in the 300-800 ms following the stimulus, the midpoint of which is P300 latency. Searching from this point to 1300 ms, the algorithm then found the average of the maximally negative 100 ms segment and subtracted it from the average of the maximally positive 100ms value to determine the p-p amplitude. These search windows were established using a grand average including Ps in both conditions, as recommended by Keil et al. (2014), and individual ERPs were checked so minimal adjustments could be made to capture P300 peaks outside the default windows. Given that recognized items can take longer to process, we also looked for differences in P300 latency, by block.

Indices of the CIT effect, as measured by a) probe amplitude minus the combined irrelevant (“Iall”) amplitude (probe-Iall), b) number of bootstrapped iterations where the probe amplitude exceeds that of the combined irrelevants (probe>Iall), and c) the average difference between probe and Iall bootstrapped mean amplitudes, are reported below, both individually and by group. Analyses were conducted for each block.

**Intraindividual Diagnosis**

We used the p-p bootstrapped amplitude difference method to establish if an individual’s
response to the probe was truly larger than to irrelevants. Because single EEG sweeps are noisy, comparing multiple probe and irrelevant responses averages for each P is more desirable, but is impractical since having Ps complete each test several times could lead to psychological effects that could complicated the interpretation of results (e.g., exhaustion, habituation, irrelevants gaining meaningfulness after repeated exposure). Farwell and Donchin (1991) offered bootstrapping (Efron & Tibshirani, 1994) as a solution because it produces distributions that mimic what CIT results might look like if a single P completed the same test many (e.g., 100) times, this method allows one to state with a chosen level of confidence (e.g., 90%) that probe>Iall, within an individual.

We conducted the bootstrapping procedure by first making probe and irrelevant ERPs, created by randomly sampling 30 probe and 30 irrelevant waves (with replacement) from the original set of sweeps. Average probe and irrelevant amplitudes were then calculated from the ERPs, and a probe-Iall difference value was computed. After 100 iterations of this process, the resulting probe-Iall difference values were used to create a distribution, and a diagnosis of “knowledgeable” was given if the probe>Iall (i.e., difference values were >0) in at least 80% of comparisons.

While this 80% criterion is lower than the 90% cutoff often used when probing for semantic autobiographical material, it was chosen here for two primary reasons. First, since probe knowledge was acquired during a mock-crime, critical items were expected to produce smaller P300s than well-rehearsed semantic details (e.g., name, birthdate) which are often tested for and easily detectable with a more stringent cutoff (Rosenfeld, Shue, & Singer, 2007). Secondly, consideration of how probing for details over multiple blocks—as required for the SCIT to extract specific details from general information—might influence test results, also
supports that a lower criterion could increase correct detections without compromising the test’s specificity. Put differently, while an innocent P might show an elevated response on one block, it is less likely they would find multiple probes meaningful.

In addition to the number of iterations where the probe>Iall, bootstrapping produces another measure of the CIT effect. Bootstrapped probe and Iall mean amplitudes are calculated 100 times, and while their differences are added to the distribution that serves as the benchmark for individual diagnosis, the values used to calculate each difference can be used to estimate the population mean difference. To illustrate the difference in these bootstrapping outcomes, consider two Ps, both with probe responses that exceed Ialls’ on 100% of comparisons. While the CIT effect would be confirmed for both, as measured by number of probe>Iall iterations, the degree of the effect is reflected in the difference between the bootstrapped probe and irrelevant mean amplitude values. For example, Ps with probe-Iall differences of 5 and 15 microvolts (µV) would both produce 100% probe>Iall iterations. An individual diagnosis based on the bootstrap test was made for each block, as well as an overall cumulative judgment (explained below).

**Intergroup Differences**

Group differences in behavioral outcomes and in the CIT effect, as measured by the two bootstrapped variables explained above and probe-Iall voltages, were assessed at the Pz site for each block using analyses of variance (ANOVA) methods and t-tests. Partial eta squared (ηp²) is reported as an effect size estimate (Cohen, 1969⁸), and JZS Bayes Factors (BFs, scaled r=.707; Rouder et al., 2009; calculated at http://pcl.missouri.edu/bayesfactor ) are reported as directly interpretable odds ratios, stating the likelihood that the group means are actually different

---

⁸ The benchmark criteria for ηp² is 0.01 for a small effect, 0.06 for a medium effect, and 0.14 for a large effect.
(favoring the alternative [alt.] hypothesis) or that they truly do not differ (favoring the null hypothesis). For example, if BF=1.0, the null and alt. hypotheses are equally likely.

While continuous measures of the CIT effect are informative, an ultimate “knowledgeable” vs. “unknowledgeable” diagnosis is necessary when deciding whether to pursue information gained during a SCIT block. Here, the inclusion of an innocent group (so false positive rates could be computed) lends the data to be analyzed in a way that models situations where investigators are unsure which suspects are involved in a crime. Thus, to see if we could identify details after first establishing who had knowledge of initial crime-related information (i.e., the jewelry), we assessed Block 1 using only the known probe method.

Hit rates were calculated for the SG and Innocent groups and compared for each block independently, as well as cumulatively. Overall cumulative hit rates were figured using both the known and blind probe methods. SG Ps were considered a miss if they a) had <80 bootstrapped iterations where the probe>Iall on any block using the known probe method, or b) showed the largest response to an item that was not the true probe on any block using the blind probe method. To approximate real life situations where innocent Ps would not go on to further blocks after being determined to be unknowledgeable in Block 1 where the probe is known, overall innocent hit rates are based on Block 1 alone. However, in many applied situations, such as when an admitted terrorist is in custody, there is no need to confirm initial knowledge before probing for details. To gauge how effective the test might be in situations when the core probe is known by investigators and suspects, we also determined cumulative bootstrap scores that included only blind Blocks 2 and 3. Finally, we examined all three blocks combined, using scores based on the largest item, whether or not it was the true probe for that individual. While this comparison includes false positives and does not offer information about probe specifics, we reasoned that
evaluating group differences in this fashion might still differentiate between guilty and innocent groups. Fischer’s Exact tests were used to compare group hit rates on each block.

While guilt was inferred here if the probe>Iall on 80+ of 100 bootstrapped comparisons so that overall hit rates incorporating all 3 blocks could be figured, this criterion can be adjusted to maximize the tradeoff between sensitivity and specificity, depending on the possible consequences of an incorrect classification. Thus, receiver operating characteristic (ROC) analyses were conducted, and area under the curve (AUC; ranging 0.5-1) values, which reflect both sensitivity (i.e., hits) and specificity (i.e., correct rejections) at each possible decision criterion, are reported. ROC analyses were performed for Block 1 (known probe), all three known blocks combined, summed blind Blocks 2 and 3, and known Block 1 combined with blind Blocks 2 and 3. A final ROC analysis was performed for blind Blocks 2 and 3 for SG Ps and known Blocks 2 and 3 for innocent Ps. This comparison was warranted since innocent Ps would not have gone on to complete the blind blocks after being deemed innocent in Block 1 where the probe was known.

Results

After excluding thirteen Ps who failed to follow instructions (e.g., got more than one recall wrong in a single block; \(n=4\)), had data with excessive artifacts (\(n=5\)) or outside the normal range (e.g., \(>35 \mu V\) probe response; \(n=3\)), or where there were software errors (\(n=1\)), thirty-one data sets remained (SG; \(n=16\), Innocent \(n=15\)). Analyses are reported using the known probe method, followed by the blind probe method (see Table 1 and Figure 3). For each block, behavioral outcomes of RTs and target error rates are reported, as well as latencies and the following indices of the CIT effect a) probe-Iall amplitude, b) number of bootstrapped iterations where the probe amplitude exceeds that of the combined irrelevants (“Iall”; probe>Iall), and c)
average difference between bootstrapped probe and Iall mean amplitudes (“bootstrapped amplitude mean differences;” BAMD). Hit rates and AUCs were then calculated as measures of the SCIT’s efficacy.

**RTs and Target Errors**

To investigate possible differences in behavioral responses, we completed a series of 2 (group SG vs. innocent) X 2 (stimulus probe vs. Iall) mixed ANOVAs on RTs, and used t-tests to explore target error rates. RTs were examined by block, using both the known and blind probe methods. While all other RT tests produced \( p > .15 \), the single test with results approaching was the Block 1 test using the known probe method mean RT for the probe was 423ms vs. 410ms for Ialls. The main effect of group was not significant (\( F(1,29)=0.00, p=.996, \eta^2<.001, BF=5.22, \) null). While the main effect of stimulus also fell short of significance, (\( F(1,29)=3.454, p=.073 \)), its effect size approached large (\( \eta^2=0.106 \)) and the BF suggested the equal likelihood of the null and alternative hypotheses (BF=1.139, null). The crossover interaction, which can be seen in Table 1 RT column and in Figure 3, was significant (\( F(1,29)=6.503, p=.016 \)), its effect size was large (\( \eta^2=.183 \)), and the BF favored the alt. at 2.981. Results of \( t \)-tests did not show differences between probe RTs (\( t[1,28]=.364, p=.719, d=0.131, BF=4.909, \) null) and Iall RTs (\( t[1,28]=.398, p=.694, d=0.143, BF=5.220, \) null). Regarding error rates, only Block 2 produced significant differences, and only for commission errors (\( t(1,28)=2.034, p=.051, d=0.334, BF=1.165 \)).

**ERP Qualitative Results**

Each block included an average of 28 probe sweeps. Grand average ERPs for each block using the known probe method are shown in Figure 1. Visual inspection of the EPRs with known probes shows an obvious CIT effect for SG but not innocent groups, across all three blocks, as expected. In each block, probes seem to have similar p-p amplitudes.
Figure 1. Grand average innocent (In; n=15) and guilty (SG; n=16) ERPs. Results were measured peak-to-peak at Pz using the known probe method.

ERP Quantitative Results

We conducted 2 (group SG vs. innocent) X 2 (stimulus probe vs. Iall) mixed ANOVAs on P300 latencies, and the same was done for amplitudes, measured p-p. These were carried out by block, using the known and blind probe methods, as were t-tests used to compare groups on the number of bootstrapped iterations where Pr>Iall and average bootstrapped mean p-p probe-Iall differences (BAMD).

Latency Group P300 latency differences between the probe and the irrelevant failed to reach significance using either the blind or known probe method. Average latencies and ANOVA results for each block are displayed in Table 1 and plotted in Figure 2.

Amplitude Known Probe Amplitude-based outcomes are also presented in Table 1 and shown in Figure 2. To investigate the CIT effect, we completed (group SG vs. innocent) X 2 (stimulus probe vs. Iall) mixed ANOVAs on p-p P300 at Pz. As expected, using the known probe
method, the main effects of group \((F(1,29)=3.706, p=.064, \eta^2_p=.113, JZS BF=1.025, \text{null})\),
stimulus type \((F(1,29)=38.577, p<.001, \eta^2_p=.571, JZS BF=222985.82, \text{alt.})\), and their interaction
\((F(1,29)=45.052, p<.001, \eta^2_p=.608, JZS BF=80695.45, \text{alt.})\) were significant (or nearly so in the
case of the group main effect) in Block 1. To see the nature of this interaction, refer to Figure 2.
Follow-up tests confirmed the CIT effect in the SG but not the innocent groups, as expected,
showing larger SG probe \((16 \mu V \text{ vs. } 9.7 \mu V; t(1,29)=3.414 p=.002 d=1.239, JZS BF=19.07, \text{alt.})\)
and equal irrelevant responses \((10.0 \mu V \text{ vs. } 9.97 \mu V; t(1,29)=.023 p=.491, d=0.008, JZS
BF=5.22, \text{null})\). In Block 2, while the main effect of group was not significant \((F(1,29)=2.979,
p=.095, \eta^2_p=.093, BF=1.39, \text{null})\), the main effect of stimulus type was \((F(1,29)=27.335,
p<.001, \eta^2_p=.485, BF=1747.71, \text{alt.})\), as was their interaction \((F(1,29)=12.524, p=.001,
\eta^2_p=.302, BF=25.51, \text{alt.; see Figure 2})\). Again, \(t\)-tests confirmed that the interaction was carried
by SG group, which showed larger probe \((13.6 \mu V [\text{SG}] \text{ vs. } 8.8 \mu V [\text{In}]; t(1,29)=2.434, p=.021
\text{d}=0.88, JZS BF=2.38, \text{alt.})\), but not Iall amplitudes \((8.8 \mu V [\text{SG}] \text{ vs. } 7.9 \mu V [\text{In}]; t(1,29)=.644
p=.525, d=0.234, JZS BF=4.31, \text{null})\). Block 3 revealed that while the main effect of group was
not significant \((F(1,29)=.331, p=0.57, \eta^2_p=.011, BF=4.48, \text{null})\), the main effect of stimulus type
was \((F(1,29)=21.708, p<.001, \eta^2_p=.428, BF=404.04, \text{alt.})\), and the interaction was not
\((F(1,29)=3.003, p=.094, \eta^2_p=.094, BF=1.38, \text{null})\).

*Amplitude Blind Probe* Because the searching function of the CIT was tested on the
second and third blocks, we also analyzed these data using the blind probe method. Results
showed that the main effect of group did not reach significance in either block (Block 2 \(p=.223;\)
Block 3 \(p=.863\)). While group differences in amplitude are typically expected, they were not
seen here since the blind probe method includes innocent false positives, which would not have
occurred in real life since innocent Ps would not have completed the blind blocks after being
determined innocent in known Block 1. While the interaction also fell short of significance in both blind blocks (Block 2 \( p = .093 \); Block 3 \( p = .21 \)), the main effect of stimulus type reached significance in both, with \( p < .001 \) for each. Effect sizes for both blocks were large and BFs definitively favored the alt. hypotheses (Block 2 \( \eta^2 = .819 \), BF=4675902314, alt.; Block 3 \( \eta^2 = .834 \), BF=16510488395, alt.).

**Bootstrapping Outcomes Known Probe** Number of bootstrapped iterations where probe>irrelevant and BAMD were also examined using the known and blind probe methods for all blocks. Individual bootstrapped iteration scores are presented in Table 2 using both the known and blind probe methods. Results were calculated for each known and blind block, as well as for blind Blocks 2 and 3 combined, and known Block 1 combined with blind Blocks 2 and 3.

Again, the evidence points to an obvious CIT effect that is limited to the SG group when the probe was known. The number of iterations where probe>irrelevant was significantly greater in the SG group compared to innocents in Block 1 (SG=93.63, In=42.73; \( t(1,29)=11.937, p < .001 \), \( d = 4.239 \), BF=12333147612, alt.) and Block 2 (SG=87.94, In=58; \( t(1,29)=4.271, p < .001 \), \( d = 1.478 \), BF=151.35, alt.), and almost significantly so in Block 3 (SG=81.63, In=62.73; \( t(1,29)=1.938, p = .062 \), \( d = 0.698 \), BF=1, null). BAMD were also significantly higher in the SG Group for all three blocks when the probe was known. For Block 1 (SG=5.8, In=-.66), \( t(1,29)=7.844, p < .001 \), the effect size was large (\( d = 2.851 \)) and clearly favored the alternative hypothesis (BF=1375797, alt.), and results were similar for Block 2 (SG=4.51, In=.7; \( t(1,29)=4.348, p < .001 \), \( d = 1.573 \); BF=183.61, alt.) and Block 3 (SG=3.5, In=1.14; \( t(1,29)=2.17 \), \( p = .038 \); \( d = 0.98 \); BF=1.47, alt.).

**Bootstrapping Outcomes Blind Probe** Examining bootstrapping outcomes using the blind probe method, only Block 2 showed marginally significant group differences, and only for the
BAMD measure. SG and Innocent group means were 5.10µV and 3.64µV respectively, 
\( t(1,29)=1.979, p=.056, \) and the effect size approached large at \( d=0.716, \) with a BF=1.06, alt.

Table 1. Outcome measures of interest presented by analysis method. Probe (Pr) and Irrelevant (Iall) averages are displayed in milliseconds for response times (RTs) and latency, µV for amplitude and bootstrapped (BS) Pr-Iall mean amplitude differences, and by average frequency for errors and number of BS iterations where Pr>Iall amplitude. SG=Simple Guilty and In=Innocent, omit=omission errors, co=commission errors, reg=uncorrected errors. Significant (or nearly significant) results are in bold. Since target responses in the second part of the CTP trial are independent of which stimulus evokes the largest P300, Xs appear in the blind error rate spaces.

<table>
<thead>
<tr>
<th>Block</th>
<th>RTs</th>
<th>Latency</th>
<th>Amplitude</th>
<th>Pr-Iall BS Iterations</th>
<th>Pr-Iall BS Ave. Amp. Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Errors: Omit/Co/Reg</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Known</td>
<td>1</td>
<td>SG: Pr=431.31, Iall=401.56</td>
<td>SG=4.06, I=4.07</td>
<td>SG=599.25, Iall=589.5</td>
<td>SG=59.63, I=42.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In: Pr=413.87, Iall=418.53</td>
<td>SG=4.08, I=4.07</td>
<td>In: Pr=543.33, Iall=555.07</td>
<td>SG=5.8, I=-0.66</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SG: Pr=360.81, Iall=357.88</td>
<td>SG=4.25, I=4.25</td>
<td>SG=550.38, Iall=568.88</td>
<td>SG=87.94, I=58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In: Pr=349.4, Iall=357.4</td>
<td>SG=4.08, I=4.08</td>
<td>In: Pr=538.27, Iall=526.27</td>
<td>SG=4.51, I=7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>SG: Pr=359.63, Iall=354.13</td>
<td>SG=4.23, I=4.23</td>
<td>SG=597.59, Iall=605</td>
<td>SG=81.63, I=62.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>In: Pr=358.73, Iall=352.33</td>
<td>SG=4.31, I=4.31</td>
<td>In: Pr=529.6, Iall=539.73</td>
<td>SG=3.5, I=1.14</td>
</tr>
<tr>
<td>Blind</td>
<td>2</td>
<td>X</td>
<td>SG: Pr=557.38, Iall=554.3</td>
<td>SG=14.34, I=8.72</td>
<td>SG=93.31, I=89.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>In: Pr=488.4, Iall=534.53</td>
<td>In: Pr=11.53, Iall=7.39</td>
<td>SG=5.1, I=3.61</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>X</td>
<td>SG: Pr=615.19, Iall=599.63</td>
<td>SG=12.18, I=6.66</td>
<td>SG=93.31, I=91.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>In: Pr=552.8, Iall=533.33</td>
<td>In: Pr=11.4, Iall=6.96</td>
<td>SG=5, I=4.11</td>
</tr>
</tbody>
</table>


Figure 2. Probe (Pr) and irrelevant (Iall) response times (RTs) and P300 latencies and amplitudes. Results are presented for the Innocent (In) and guilty (SG) groups, by block, using both the known and blind probe methods. RTs and latencies are presented in milliseconds and amplitudes are in $\mu$V.
Table 2. Individual bootstrapped iterations where probe > combined irrelevant (Iall) amplitude. Results are presented by group (P=participant, SG=Simple Guilty, In=Innocent), using both the known and blind probe analysis method. Shaded entries indicate a miss (i.e., an incorrect probe identification for SGs or a false positive for innocents).

<table>
<thead>
<tr>
<th>Block:</th>
<th>Known</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1 (known)</td>
<td>2 &amp; 3 (blind)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3 (blind)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>91</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>291</td>
<td>200</td>
</tr>
<tr>
<td>P2</td>
<td>99</td>
<td>82</td>
<td>88</td>
<td>83</td>
<td>85</td>
<td>267</td>
<td>168</td>
</tr>
<tr>
<td>P3</td>
<td>96</td>
<td>99</td>
<td>81</td>
<td>99</td>
<td>81</td>
<td>276</td>
<td>180</td>
</tr>
<tr>
<td>P4</td>
<td>100</td>
<td>98</td>
<td>86</td>
<td>98</td>
<td>92</td>
<td>290</td>
<td>190</td>
</tr>
<tr>
<td>P5</td>
<td>100</td>
<td>99</td>
<td>4</td>
<td>99</td>
<td>100</td>
<td>299</td>
<td>199</td>
</tr>
<tr>
<td>P6</td>
<td>100</td>
<td>92</td>
<td>17</td>
<td>94</td>
<td>100</td>
<td>294</td>
<td>194</td>
</tr>
<tr>
<td>P7</td>
<td>89</td>
<td>84</td>
<td>94</td>
<td>86</td>
<td>94</td>
<td>269</td>
<td>180</td>
</tr>
<tr>
<td>P8</td>
<td>90</td>
<td>87</td>
<td>87</td>
<td>96</td>
<td>91</td>
<td>277</td>
<td>187</td>
</tr>
<tr>
<td>P9</td>
<td>86</td>
<td>94</td>
<td>99</td>
<td>94</td>
<td>99</td>
<td>279</td>
<td>193</td>
</tr>
<tr>
<td>P10</td>
<td>93</td>
<td>98</td>
<td>97</td>
<td>98</td>
<td>97</td>
<td>288</td>
<td>195</td>
</tr>
<tr>
<td>P11</td>
<td>100</td>
<td>90</td>
<td>95</td>
<td>90</td>
<td>95</td>
<td>285</td>
<td>185</td>
</tr>
<tr>
<td>P12</td>
<td>95</td>
<td>81</td>
<td>85</td>
<td>81</td>
<td>93</td>
<td>269</td>
<td>174</td>
</tr>
<tr>
<td>P13</td>
<td>92</td>
<td>50</td>
<td>99</td>
<td>93</td>
<td>99</td>
<td>284</td>
<td>192</td>
</tr>
<tr>
<td>P14</td>
<td>94</td>
<td>87</td>
<td>80</td>
<td>87</td>
<td>80</td>
<td>261</td>
<td>167</td>
</tr>
<tr>
<td>P15</td>
<td>98</td>
<td>66</td>
<td>100</td>
<td>95</td>
<td>100</td>
<td>293</td>
<td>195</td>
</tr>
<tr>
<td>P16</td>
<td>75</td>
<td>100</td>
<td>94</td>
<td>100</td>
<td>87</td>
<td>262</td>
<td>187</td>
</tr>
</tbody>
</table>

|        | SG Mean: 94 | 88 | 82 | 93 | 95 | 280 | 187 |
|        | In Mean: 43 | 58 | 63 | 90 | 92 | 224 | 181 |
Qualitative Hit Rate Results

Hit rates across groups were then figured (see Table 3) and accuracy rates reflecting both the specificity and sensitivity of the test were calculated using ROC analysis, again using both the known and blind probe methods (see Figure 3). Using the known probe method, hit rates were high across all three blocks (Block 1 97%, Block 2 87%, Block 3 81%), with an overall detection rate of 84% using the cumulative rule. Using the blind probe approach, hit rates dropped to 42% and 29% for Block 2 and Block 3, respectively, resulting in an overall 68% accuracy. The overall accuracy in this case includes all innocents who scored <80 probe>lall bootstrapped iterations on Block 1 and guilty individuals whose largest response was to the true probe on blind Block 2 and 3.

Table 3. Hit rates (rounded %) for both the known probe and blind probe methods. Results are listed by block, for innocent (In; n=15) and guilty (SG; n=16) groups, as well as combined (“overall”).

<table>
<thead>
<tr>
<th>Block:</th>
<th>Known</th>
<th>Overall</th>
<th>Blind</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>Overall</td>
</tr>
<tr>
<td>SG</td>
<td>94%</td>
<td>88%</td>
<td>88%</td>
<td>69%</td>
</tr>
<tr>
<td>In</td>
<td>100%</td>
<td>87%</td>
<td>73%</td>
<td>100%</td>
</tr>
<tr>
<td>TOTAL=</td>
<td>97%</td>
<td>87%</td>
<td>81%</td>
<td>84%</td>
</tr>
</tbody>
</table>

Quantitative Hit Rate Results

Fischer’s Exact tests on hit rate differences between SG and Innocent groups showed that when the probe was known, detection accuracy did not differ between the groups on any individual block (Blocks 1-3 \(p=.1.0; \ p=1.0; \ p=.394\). However, comparing the accuracy rates of SG individuals who were a hit on all three blocks to innocent Ps who were determined innocent on Block 1 (and were thus considered overall hits) showed that accuracy was higher in the Innocent group overall \(p=.043\). When the blind probe method was used, group hit rates were
significantly different in Block 2 \((p=.029)\), Block 3 \((p<.001)\), combined Blocks 2 and 3 \((p=.018)\), and overall \((p<.001;\) using known Block 1 and blind Blocks 2 and 3).

ROC analyses were then completed to establish the diagnostic ability of the test, with and without a known probe. Five analyses were carried out 1) Block 1 (known), 2) Block 1+2+3 (known), 3) Blocks 2+3 (blind), 4) Block 1 (known)+Blocks 2 and 3 (blind), and 5) SG Blocks 2+3 (blind) vs. innocent Blocks 2+3 (known). AUCs (see Figure 4) for Block 1 (known) and for Block 1 (known)+Blocks 2 and 3 (blind) were 1.0, indicating perfect discrimination between the groups, and could not be plotted with our software. Combining all three known blocks resulted in an AUC of .987, while combining blind Blocks 2 and 3 produced an AUC of .601 (see Figure 3). Finally, comparison of blind Blocks 2 and 3 for SG Ps and known Blocks 2 and 3 for innocent Ps resulted in an AUC of .993.

**Figure 3.** Receiver operating characteristic (ROC) analyses and areas under the curve (AUC).

![ROC curves](image)

**Discussion**

While previous versions of the CIT are designed to test *who* recognizes a known item,
Our SCIT aimed to identify specific details from more general information, and is instead intended for Ps known to be concealing information. To approximate scenarios where investigators are limited by their lack of specific details but must locate something (for example, a hostage or an explosive), Ps here stole and hid a jewelry item, and then were tested to see if the SCIT could uncover details by further investigating the item with the largest P300 response, assuming it belongs to the probe.

For comparison purposes, we first completed all analyses the traditional way, using the known probe methods. As seen in Table 3, accuracy rates using this method were high, resulting and only eleven misses in 93 tests (88% accuracy). This is further supported by the highly significant effects in the key measures of CIT, including number of bootstrapped iterations where Pr>Iall and bootstrapped probe-irrelevant mean amplitude differences. Using a criterion that required SG Ps to be a hit on the true probe in all three blocks, and determined innocent Ps to be misses if they were a false positive on the first block, the known probe method correctly classified 84% of Ps. When the new blind probe method was instead performed, the test lost some diagnostic power in blind Block 2, SG hit rates fell to 63%, and dropped even further in Block 3 (56%). Even worse were the false positive rates among the innocent group in these blocks (80% and 100% respectively).

Given the high accuracy rates the test produced when not used in the searching manner and the high accuracy rates of Meixner & Rosenfeld (2011) who also assumed the largest response belonged to the probe and made no false positive judgments, it is unclear why our outcomes were poor when the probe was unknown. One possibility is that the length of the test took a toll on Ps, and while this is supported by the overall declining hit rates over the blocks, Meixner & Rosenfeld (2011) and others also had study designs with multiple blocks and did not
face this issue. Another possible explanation for the disparity between our results and the results of Meixner & Rosenfeld involves the stimuli we used. While verbal stimuli were used in that study, we instead used pictures based on findings that they produce larger P300s in the CTP (Rosenfeld, 2015). However, these findings were from a guilty group, and in recent work using images as stimuli, an innocent group has not typically been run. Thus, it may be that while images enhance the CIT effect in guilty Ps, they might also be unintentionally salient to innocents in some cases, contributing to false positives.

Considering how stimuli responses are statistically distributed can help explain the present study’s high rates of false positives in the innocent group and probe misidentification in the SG group when the blind probe method was used. One expects innocent Ps to respond similarly to all stimuli and guilty individuals to respond similarly to all items except for the probe. When the probe is known, one compares its response to that of the combined responses to the other items. So, for guilty individuals, if the largest response is not to the probe, they can still be correctly identified if the true probe is enough larger than the combined irrelevants. The known probe method also protects innocent Ps from false positives because, even if an innocent P has a fluke elevated response to one of the items (as long as it is not to probe), it would decrease the CIT effect instead of enhancing it. While the poor detection accuracy seen here when using the blind probe approach is still unclear, it highlights why it is important to test stimulus sets on a group known to be innocent so that any items that cause an elevated response can be replaced before real suspects are tested. Interestingly, post-experiment questionnaires revealed that, of the 14 Ps who responded that one of the stimuli in the set stood out for them in Blocks 2 and 3, only three identified the same item selected by the blind probe procedure. This suggests that future work should test stimuli using the CIT rather than simply asking an innocent
stimuli pilot group. Another possible approach for reducing false positives is to increase the number of irrelvants used. In cases when an irrelevant sticks out, the added irrelevants could serve to buffer its response while also broadening group differences, since P300 is known to increase as a function of rarity.

Here, we carried out a standard known probe analysis in Block 1 and then applied the blind probe approach in Blocks 2 and 3. Because of the high false positive rate when using the blind probe method, establishing that someone has initial knowledge or crime-related details before using the blind probe approach might be an important way to protect against false positives. However, this may step may not be necessary in cases when initial knowledge has been established, for example, when someone admits to a crime.

While individual detection rates using the blind probe method were not impressive, like Meijer et al. (2010), we found some evidence—in the form of bootstrap measures and probe vs. irrelevant amplitude differences—that the CIT could be useful for detecting an unknown probe in groups. Future research testing the SCIT might more directly test this approach, for example, with a paradigm that has Ps complete a crime together and then combines their results. This method could also be explored on the group level to test for crimes that have been planned and not committed, as in Meixner & Rosenfeld (2011). Importantly however, while using the known probe method in Block 1 combined with the blind probe method in Blocks 2 and 3 only resulted in finding the jewelry item in six of the 16 SG Ps, 15 of the 16 were correctly classified as guilty, even if the specific item was not correctly identified. Furthermore, the true probe was correctly identified in 10 SG Ps in Block 2 and nine in Block 3, providing evidence that the blind probe method can be useful approach to help identify items of interest.

Limitations and Future Directions
While the findings presented here leave room for improvement, given the theoretical basis for the P300-based CIT and reasoning behind applying it with the goal of searching for unknown general to unknown specific details, future efforts should be taken to refine this method before it is abandoned. One limitation of conducting lab mock-crimes, in which Ps have very limited exposure to the probe, is that they do not necessarily translate to some real situations when criminals have planned and rehearsed their crime extensively or reflected back on it after it was committed. To account for this, future tests of the SCIT might test for information that is more well-rehearsed (e.g., what neighborhood Ps live in). Future research might also directly test the SCITs accuracy on Ps who are known to be knowledgeable before testing. While we indirectly explored this by assessing cumulative hit rates in Blocks 2 + 3, future work might directly test the viability of the SCIT on those with confirmed knowledge of the crime to eliminate and undue effects (e.g., exhaustion) that testing might have. By addressing the factors that may have limited our findings, we hope that future work improves the viability of the CIT when used to search for unknown details.
CHAPTER 3 APPLYING MACHINE LEARNING TO THE CONCEALED INFORMATION TEST (CIT) COMPARISON OF DIAGNOSTIC ACCURACY USING BOOTSTRAPPING METHODS VERSUS A MACHINE LEARNING ALGORITHM USING THE COMPLEX TRIAL PROTOCOL (CTP)

Introduction

The Concealed Information Test (CIT)

The concealed information test (CIT; Lykken, 1959) is a method for detecting recognition of details that only someone related to a crime would know. During a CIT, subjects are presented with three types of stimuli. The first—the probe—is the item investigators and involved suspects know to be related to the crime (e.g., the murder weapon). Items in the next stimulus category are selected to be from the same domain as the probe and are called “irrelevants” because, although they are similar to the probe and typically require the same behavioral response as the probe, they are not related to the crime under investigation. For example, if a murder were committed with a gun, irrelevant items might include images of a knife, bat, etc. When shown an image of the murder gun (i.e., the probe) presented among images of other weapons unrelated to the crime (i.e., irrelevants), those who are unfamiliar with the crime’s details should respond similarly to all items. However, because the murder gun is distinguishable amongst the other items for knowledgeable individuals due to its familiarity and salience, individuals involved in the crime are expected show a differential physiological response to the probe (i.e., the “CIT effect). The third class of stimuli requires a different behavioral response than the probe and irrelevants, and are called “targets” because they are included to force task engagement/attention. Because its unique button response makes the target meaningful, elevated physiological target responses are expected for both knowledgeable and unknowledgeable subjects.
The CIT has a firm theoretical basis, is more empirically established than alternative methods for detecting recognition, and has been researched using a variety of physiologically measures, including skin conductance rate (SCR), respiration line length (RLL), and heart rate (HR) (Verschuere & Ben-Shakar, 2011; Ben-Shakar, 2012). While these autonomic nervous system (ANS) measures can effectively detect recognition (with Cohen’s $d$ effect sizes of $d=1.55$, $1.11$, and $0.89$ for SCR, RLL, and HR, respectively), the most robust CIT indicator of recognition is P300 amplitude ($d=1.89$; Meijer et al., 2014). Because this endogenous ERP occurs involuntarily when a rarely presented meaningful item—like a murder weapon—is recognized, it can be used to determine who is privy to a crime’s details by comparing P300 amplitude in response to probes and irrelevant items.

Concealed Information Detection Using Bootstrapping

Much of the original work using P300 to detect concealed information used t-tests to look for intra-individual differences between the means obtained from single sweep probe and irrelevant ERPs (e.g., Rosenfeld et al., 1987, 1988). Because single sweeps are very noisy and several sweeps are typically needed to detect signals unless they are very pronounced, most analyses in ERP physiological work has examined group differences by comparing mean amplitudes of grand averages, which include the averages of all participants (Ps) in a condition. For each individual to have several probe and irrelevant averages, they would need to complete the test many times, which is not feasible (due to time constraints, expense, habituation, the risk of irrelevants becoming relevant through repeated exposure, etc.). As a result of comparing noisy single sweeps, individual diagnostic rates in these original studies was low (<80%).

Farwell and Donchin (1991) solved the problem of comparing single sweep probe and irrelevants and made accurate individual diagnosis possible by applying the bootstrapping method
(Efron & Tibshirani, 1994) to P300-based concealed information detection to produce multiple probe and irrelevant averages for each individual. Using bootstrapping as we currently do, the ultimate goal is to establish if a CIT effect is present in a given individual (e.g., determine with at least a 90% confidence that the probe P300 amplitude is truly larger than the irrelevants’). To do this, distributions of probe and the combined irrelevant (Iall) averages are needed for each P. As detailed below (see “Bootstrapped Amplitude Difference Method” section), distributions are created by randomly sampling single probe and single irrelevant waves, creating a probe and an irrelevant ERP based on these sweeps, finding the probe-irrelevant difference and adding it to a distribution (consisting of 100 difference values after this process is repeated 100 times), and then drawing the lower bound of the confidence interval at the point in the distribution that makes to confidence interval >0 (e.g., a cutoff -1.29 SDs corresponds to a 90% confidence that the probe and irrelevants differ). While guilt is often inferred if the probe-Iall bootstrapped differences are >0 in 90+ of 100 iterations, this criterion can be adjusted to achieve the desired sensitivity/specificity (based on the severity of the consequences of false positives or false negatives).

Instead of diagnosing concealed information based on the number of bootstrapped iterations where the probe is larger than Iall, others have instead relied on cross-correlation coefficients (Farwell & Donchin, 1991). Essentially, this method creates bootstrapped ERPs for the probe, Iall, and the target, which are then used to determine whether the probe P300 is more similar in shape to the target or to Iall. Single sweeps from the probe are compared to those of the target and Iall, and these values are used to make two distributions probe vs. Iall and probe vs. target. This approach assumes that a greater cross-correlation between the probe and Iall (A) indicates innocence, because it suggests that the probe is not meaningful for the P, while a
greater cross-correlation between the probe and target (B) indicates guilt, because it suggests that the probe is meaningful like the target. In this case, concealed information is typically indicated if at least 90 of 100 correlation subtractions (B-A) are >0.

The cross-correlation approach has limitations however, including that it compares probes and targets, which are inherently different; probe P300s occur because an item is meaningful, while target P300s are evoked due to task-relevance, which can result in latency, phase, and other morphology differences (see Figure 2 of Rosenfeld et al., 2004). As detailed in Rosenfeld (2011), these differences make the target an inadequate index for probe comparison, and so unlike the cross-correlation approach, the bootstrapping method instead compares probe and irrelevant amplitudes.

Research comparing the bootstrapped peak-to-peak (p-p) amplitude difference method to the cross-correlation approach has supported the superiority of the former (Rosenfeld et al., 2004; Abootalebi et al., 2006; Mertens & Allen, 2008). However, these traditional time-domain based approaches are not ideal since they fail to tease apart functionally separate, yet co-occurring neural events, the summed effects of which are reflected in ERPs. Considering the limitations associated with restricting analyses to a time domain, one weakness of the bootstrapped p-p amplitude difference method is that it only looks at part of the wave, meaning that any activity occurring outside a given time frame that could improve diagnostic accuracy gets ignored. When a frequency domain has instead been used, components in different frequency ranges have demonstrated their association with discrete neural events. For example, some frequencies have shown stronger associations with information processing, and others with behavioral events (e.g., Basar et al., 2001). However, while Fourier transformation can help begin to relate frequency to functionally separate neural events, it fails to take time into account.
Concealed Information Detection Using Machine Learning (ML)

To improve the identification of separate functional components, EEG/ERP signals must be considered in dimensions that account for both frequency and time, which can be accomplished using computer algorithms. One such algorithm that incorporates both domains—wavelet transformation—can establish when and to what extent distinct events occur within a waveform (allowing multiple underlying neural events to be represented), as well as when separate frequencies occur, and how this changes over the course of the wave (e.g., see Unser & Aldroubi, 1996). Because machine learning (ML) incorporates both time and frequency domains, ML methods can be applied to recognize patterns in complex data sets, through the use of a statistical classifier and wavelet features.

To illustrate how features are selected, consider how one can tell a circle and a square apart; these objects can be broken down into discrete features, like number of sides, angles within shape, and curvature measures within shape, etc., which discriminative ML algorithms (e.g., logistic regression cost minimization) can use to find a plane or line in the feature vector space that best delineates squares from circles, based on the predefined features. Put differently, features or attributes selected because they are thought to be associated with a given output (e.g., shape type) are extracted, and statistical classifiers (i.e., functions) input this data and assign it a label output (e.g., circle vs. square, probe vs. irrelevant). A major benefit of ML is that it allows one to incorporate features or attributes which are unapparent to the naked eye, allowing the classifier to customize a unique and potentially non-linear delineation based upon the data, instead of on preconceived notions about how the data is expected to behave, which could ultimately lead to more objectivity and diagnostic accuracy.

Applied to concealed information detection, probe and irrelevant responses can be broken
down into different features or attributes, which a ML algorithm can use to find a division between the classes “guilty” and “innocent,” based on quantitative feature measures. In an attempt to do this using P300, Abootalebi et al. (2006) compared their ML pattern recognition method to the bootstrapped amplitude difference method and the cross-correlation approach and found evidence favoring either ML or the bootstrapping method (depending upon the chosen receiver operator characteristic [ROC] criterion), and that the cross-correlation method performed consistently (yet slightly) worse. Extending upon this research, Abootalebi, Moradi, & Khalilzadeh (2009) further refined their initial pattern recognition system by adding new morphological and frequency features to the previously used wavelet features. In addition to outperforming their previous attempt, their results also exceeded the bootstrapped p-p amplitude difference and cross-correlation approaches.

**Applying ML to Complex Trial Protocol (CTP) Data**

While the findings of Abootalebi and colleagues (2006, 2009) suggest that ML approaches may be preferable to the traditionally used bootstrapped p-p amplitude difference method, it is important to note that the comparisons between the methods for diagnosis discussed above were drawn from data acquired using the suboptimal “three stimulus protocol” (3SP; Rosenfeld, 2011). One shortcomings of this approach is that it combines implicit probe/irrelevant and explicit target/nontarget discrimination tasks; presenting probe, irrelevant, and target stimuli at random requires a simultaneous implicit probe/irrelevant categorization and explicit target/nontarget discrimination on every trial (since the target requires a unique response). While the probe’s amplitude typically exceeds that of the irrelevants in knowledgeable individuals, the secondary target/nontarget task unnecessarily consumes cognitive resources and diverts attention away from the primary probe/irrelevant categorization, reducing probe P300 amplitude
(Donchin, Kramer, & Wickins, 1986). In addition to this reduction of the CIT effect—the difference between probe and irrelevant responses, upon which a diagnosis is determined—another weakness of the 3SP is its vulnerability to “countermeasures,” or deliberate attempts to beat the test. Specifically, making irrelevant stimuli meaningful by applying countermeasures when they are presented increases their P300, thus also reducing the CIT effect (Rosenfeld, Soskins, Bosh, & Ryan, 2004; Mertens & Allen, 2008).

The Complex Trial Protocol (CTP; see Figure 2 in Chapter 1; Rosenfeld et al., 2008) was developed to address the shortcomings of the 3SP and separates probe/irrelevant and target/nontarget responses in three ways 1) temporally, by about a second, 2) spatially, by assigning left-hand probe/irrelevant responses and right-hand target/nontarget responses, and 3) by domain, using numbers as targets instead of simply making an irrelevant a target by assigning it significance.

These modifications increase target discriminability, which frees processing resources and allows for increased attention toward the probe, resulting in greater probe P300 amplitudes and a more dramatic CIT effect. Given the sparsity of research testing ML diagnosis methods on P300 CIT data and the notable differences between the 3SP and CTP, it unclear if ML methods might optimize concealed information diagnosis using the more sophisticated CTP version of the P300-based CIT. To that end, the goal of the current effort is to compare a discriminative ML algorithm to the currently used bootstrapped amplitude difference approach, using data acquired during the CTP.

Methods

Participants
The ML and traditional bootstrapped amplitude difference methods were tested on a dataset with 30 Ps from the Northwestern University’s Introductory Psychology pool. All Ps had normal or corrected to normal vision, reported no history of neurological or psychological abnormalities, consented to participation, and received class credit.

Procedure Overview

Ps were randomly assigned to either a simple guilty (SG; \(n=16\)) or an innocent control group \((n=14)\). After being told the premise of the CIT they would later take (see Appendix I “Pre-Task Instructions”), Ps completed either a mock-crime or an innocent control task, followed by a P300-based CIT.

Mock-Crime/Control Task

Ps read along as their participation and task, which involved entering an office to steal an item from an envelope (SG) or to simply look inside an envelope (innocents), and then returning to the lab, were explained by the experimenter (see Appendix I). Half of Ps in the SG group stole a watch while remaining Ps stole a bracelet.

P300-based CIT

After the goal of the CIT and instructions on how to respond behaviorally to stimuli during the task were described (see Appendix J), all Ps completed the CTP version of the P300-based CIT. Ps were seated three feet from a monitor, upon which images of the probe (watch or bracelet), six irrelevants (comparable jewelry items), the target (i.e., “11111”), and four nontargets (i.e., “22222…55555”), appeared. Each CTP trial (see Figure 2 in Chapter 1) began with a fixation cross “+” in the middle of the screen (200 ms), which was replaced with probe or irrelevant (300 ms), followed by 1100 ms of observation, and finally a randomly varying interstimulus interval of 50-200 ms (with fixation “+”). This was repeated in the second part of
the trial, except with target and nontarget stimuli. The target and four nontargets were each presented 20% of the time, resulting in a 20%-80% target to nontarget ratio. We also used a symmetric protocol—meaning that the target and each of the nontarget stimuli followed each irrelevant and the probe with the same frequency—in an attempt to prevent Ps from focusing their attention on detecting patterns to the target occurrences.

As described above, probe and irrelevant stimuli required the same left-hand button response, simply acknowledging that the item was seen, while right-handed response were made for target (right mouse button) and nontarget (left mouse button) stimuli. The CIT consisted of 210 trials and lasted about 15 minutes. To help force attention to the stimuli, Ps were stopped unpredictably about every 50 trials and asked to name the previous item they had seen (as warned during their instructions).

**Data Acquisition**

EEG was collected using Ag/AgCl electrodes attached midline to the scalp at Fz, Cz, and Pz, sites. Electro-oculogram (EOG) was recorded referentially with an electrode placed above the left eye to collect eye movements and blinks, and all electrodes were referenced to linked mastoids and grounded to a forehead electrode. Impedance remained under 5 kΩ, eye blink artifacts were corrected using the Semlitsch method (Semlitsch et al., 1986), and trails containing artifacts over 90 µV were dropped. The 19 channel Mitsar amplifier (Model 201) used a 30 Hz low pass and a 0.16 Hz high pass filter setting, and output passed through a 16-bit Mitsar A/D converter sampling at 500 Hz. For display and analyses, a Kaiser (alpha= 1.8) filtering algorithm (with the digital filter set to pass frequencies from 0 to 6 Hz), was used off-line to remove higher frequencies in single sweeps and averages.

**Data Analysis Plan**
ERP Measurement

P300 was analyzed at Pz, where it is typically largest (Johnson, 1993), using the p-p method, which has been shown to be 25% more accurate at detecting knowledge than the base-to-peak method (b-p; Soskins, Rosenfeld, & Niendam, 2001). To find p-p P300 values, the algorithm searched from 300-700 ms for the maximally positive 100ms segment, the midpoint of which is P300 latency. Searching from this P300 latency to 1400 ms, the algorithm also establishes a maximally negative 100 ms segment average, and then subtracts it from the maximally positive average to find the p-p amplitude value. As recommended by Keil et al. (2014), our search windows were chosen based on a grand average including all Ps, across conditions. Individual ERPs were also examined, and search windows were minimally expanded for the eleven Ps with P300 peaks outside the default windows.

Since we expected to see the CIT effect only in the SG group, we used analysis of variance (ANOVA) methods to assess for differences in probe and irrelevant responses at the Pz site, between the SG and innocent groups. Partial eta squared ($\eta^2_p$) is also provided as an effect size estimate (Cohen, 1969$^9$). Bayes Factors (JZS BF$s$, scaled $r=.707$; Rouder et al., 2009; calculated at http://pcl.missouri.edu/bayesfactor ) are reported as directly interpretable odds ratios, and test whether a true difference between group means exists (favoring the alternative hypothesis), or not (favoring the null hypothesis). For example, if JZS BF=1.0, the null and alternative hypotheses are equally likely.

Bootstrapped Amplitude Difference Method

As described above, the p-p bootstrapped amplitude difference method was used to

---

$^9$ The benchmark criteria for $\eta^2_p$ is 0.01 for a small effect, 0.06 for a medium effect, and 0.14 for a large effect.
determine if the probe’s P300 response exceeded that of the irrelevants, for each individual. In this procedure, 30 probe waves are first randomly sampled (with replacement) from the original set of 30 single probe sweeps. These single sweeps are then used to create a probe ERP, from which an average bootstrapped probe P300 value is computed (where P300 is defined as the midpoint of the largest 100 ms segment average in the 300-700 ms following the stimulus, for example). The same process is then carried out with irrelevant stimuli, and 30 of the original 180 irrelevant single sweeps are sampled and used to create an ERP, from which an average bootstrapped irrelevant value is computed. The difference between these two values—the bootstrapped probe average minus the bootstrapped irrelevant average—is then calculated. This process of a) randomly selecting 30 probe and 30 irrelevant sweeps, b) creating probe and irrelevant ERPs, and c) comparing their amplitudes, is typically carried out 100 times. The resulting 100 probe-irrelevant difference values are then used to create a distribution, from which a diagnosis is made, based on the number of iterations where the probe>Irall. For example, if the criterion for a “knowledgeable” determination is 90%+ Pr>Irall iterations, a cutoff is drawn 1.29 standard deviations below the mean, making the lower bound of the 90% confidence interval >0 (since negative values in the distribution represent the iterations where Irall>Pr).

Bootstrapping offers an alternative to gauging the CIT effect based on measures reported directly in voltages (i.e., probe-irrelevant amplitude differences). To determine if the CIT effect is present (i.e., to diagnose concealed knowledge), we typically use the criteria of 90 out of 100 iterations where the probe P300 is larger than Irall’s, measured p-p. However, this 90% criterion is arbitrary, and studies testing for episodic information (e.g., stolen jewelry, as used here), which is less salient and thus typically produces smaller P300s (Rosenfeld, Shue, & Singer, 2007) than semantic information (e.g., name, hometown), often adopt a less stringent criterion of
85% or 80%. A final and related measure of the CIT effect is the difference between estimated probe and irrelevant mean amplitudes; during the bootstrapping procedure, P300 amplitude averages are produced for both probes and irrelevants during each iteration, and the average of these sample means can be used to estimate a probe and irrelevant population mean difference.

**ML Method**

The discriminative ML algorithm tested here was modeled from Abootalebi, Moradi, & Khalilzadeh (2009). However, instead of using linear discriminant analysis, we used logistic regression. For a binary classification—such as knowledgeable vs. unknowledgeable—logistic discriminant analysis predicts a normal density function for each classification feature, establishing a linear boundary where they meet. However, unlike linear discriminant analysis, logistic regression instead creates a class boundary by predicting the log-odd function between the classes, which may be preferable since it does not make a-priori assumptions about distributions of predictor variables. Additionally, while Abootalebi (2009) employed discrete wavelet transforms, we did not. Finally, while Abootalebi and colleagues (2009) used genetic algorithms to pick the classifier features, we instead selected a subset of the features they employed. After testing combinations to see which best fit the data, the following were included in our set our features

- **Amplitude** the maximum signal value.

- **Positive area** the sum of the positive signal values.

- **Negative area** the sum of the negative signal values.

- **Total area** the sum of positive and negative signal values.

---

10 For more detailed explanations of the selected features and their formulas, please refer to Abootalebi, Moradi, & Khalilzadeh (2009).
- **Total absolute area** the sum of the positive signal values and the absolute value of the negative signal values.

- **Zero crossing** the number of times the signal value equaled zero, in the peak-to-peak time window.

- **Slope sign alterations** the number of slope sign alterations of a pair of points adjacent on the ERP signal.

- **Raw values** normalized EEG voltage in the second following stimuli presentation, following remove EOG.

Using these features, the statistical classifier was trained on a random subset (SG $n=8$; innocent $n=7$) of the sample, and then applied to the remaining half (SG $n=8$; innocent $n=7$) for diagnosis. After the chosen features were extracted from a given probe EEG single sweep, the algorithm determined whether that sweep more closely resembled a probe or an irrelevant response. For purposes of classification (and in an attempt to reduce variance across Ps), cutoffs were established for each P based on the assumption that probe responses mirror target responses; since target responses should be elevated in all Ps while elevated probe responses should be restricted to the SG group, a knowledgeable diagnosis was made if the classifier identified probe sweeps (in the first part of the trial) as such $\geq$ the % target sweeps classified as a probe (in the second part of the trial). For instance, if a Ps target response was classified as a probe 60% of the time, that P would be determined guilty if probe sweeps were classified as such at a rate $\geq 60\%$.

**Comparing bootstrapping and ML**

Because diagnostic accuracy rates for each method are based on different wave characteristics (e.g., strictly amplitude for bootstrapping and several features for ML) and are
produced using different criteria (ML % target sweeps classified as a probe; bootstrapping number of iterations out of 100 where the probe > Iall), hit rates are not an ideal index for making methodological comparisons. To best compare ML algorithm to the traditional bootstrapping method, areas under the curve (AUCs), which reflect both sensitivity (i.e., hits) and specificity (i.e., correct rejections) at each possible decision criterion, were obtained through ROC analysis.

To determine if one method outperformed the other, we compared the ML and bootstrapping AUCs using Z-tests, as suggested by Hanley and McNeil (1983). Since z-scores are standardized (i.e., they indicate how many standard deviations away from the mean a particular score lies), they are useful for comparing AUCs derived using different sample sizes and classification criteria, as is the case here. We expected that, although both methods would have a high AUC (>.9+), the ML algorithm might improve upon our traditional bootstrapping diagnostic technique, as determined by z-score.

**Results**

*ERP Qualitative Analysis*

After excluding the <10% of Ps who failed to follow instructions (e.g., incorrect recall during more than one pop quiz) or who had excessive artifacts, 30 data sets remained. The mean number of probe trials across all subjects was 28, and all Ps had at least 20 Pr sweeps. Figure 1 shows the grand average ERPs at Pz for the SG and innocent groups, which demonstrate the expected CIT effect elevated P300 responses to the probe when compared to Iall in the SG group, but not the innocent group.
**Figure 1.** Grand average innocent (In; n=14) and guilty (SG; n=16) ERPs.

**ERP Quantitative Analysis**

A 2 (group SG vs. innocent) X 2 (stimulus probe vs. Iall) mixed ANOVA was run on p-p P300 amplitudes to confirm the CIT effect illustrated above. Average probe and Iall amplitudes (in microvolts [µV]) were probe=10.68, Iall=9.92 and probe=14.59, Iall=9.97 for the innocent and SG groups, respectively. The main effect of group did not reach significance, $F(1,28)=1.318$, $p=.261$, possibly due to the interaction. The $\eta^2$ approached medium at .045, but the JZS BF favored the null hypothesis at 2.83. As expected, the main effect of stimulus type was large, $F(1,28)=14.87$, $p=.001$, $\eta^2=.347$, and the JZS BF clearly favored the alternative at 52.821. The interaction was also significant, $F(1,28)=7.64$, $p=.01$, $\eta^2=.214$, with JZS BF favoring the alternative at 4.574. Figure 2 supports that the source of this interaction is the larger probe in the SG group compared to the innocent group, and illustrates the larger probe-Iall difference (i.e., the CIT effect) in the SG group. Post-hoc t-tests provided some support for this, with $t(1,28)=1.919$, $p=.065$, $\eta^2=.116$, and JZS BF suggesting the equal likelihood of the null and alternative hypothesis (1.026, favoring the null) for probes, and $t(1,28)=.033$, $p=.974$, $\eta^2<.001$, with JZS BF favoring the null at 5.141 for Iall.
Figure 2. Computed grand average p-p P300 values for probes and combined irrelevants. Results presented as a function of condition (Guilty [SG] vs. Innocent) and stimulus type.

Bootstrapping

Further evidence supporting that the CIT effect is present and limited to the SG group is offered by results derived using the bootstrapping procedure. A t-test was conducted to assess group differences in the number of bootstrapped iterations where the probe amplitude exceeded that of the irrelevants, and revealed that the probe was larger than the irrelevant more frequently in the SG group (SG mean=83.0; Innocent mean=49.8; t[1,28]=4.475, p<.001, ηp²=.070, JZS BF=240.506, favoring the alternative). A final t-test on the average bootstrapped mean p-p probe-Iall differences (SG mean=4.19; Innocent mean=0.22) showed that this index of the CIT effect was also larger in the SG compared to the innocent group (t[1,28]=3.214, p<.01, ηp²=.060, JZS BF=11.960, favoring the alternative).

Outcomes of both the bootstrapping and ML procedures are displayed in Table 1, by group (SG vs. innocent). Using the criterion of 85 out of 100 iterations where the probe>Iall for a knowledgeable diagnosis, the bootstrapping procedure accurately classified 24 of 30 (80%) individuals (SG 11/16=68.75%; Innocent 13/14=92.86%). Loosening this criterion to 80 of 100
iterations did not change any determinations, and investigation of b-p bootstrapped results showed that measuring P300 b-p caused diagnostic accuracy rates to suffer by >15% (and thus formal b-p analyses were not conducted).

**Table 1.** Values obtained by bootstrapping and machine learning (ML) procedures. Results used for ROC analysis, by group (simple guilty [SG] an Innocent [In]).

<table>
<thead>
<tr>
<th>Method</th>
<th>SG</th>
<th>In</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootstrapping</td>
<td>100, 100, 99, 98, 98, 94, 92, 92, 91, 85, 75, 58, 56, 52, 40</td>
<td>23, 30, 31, 34, 34, 36, 39, 56, 59, 60, 61, 63, 79, 92</td>
</tr>
<tr>
<td>ML</td>
<td>0.65, 0.59, 0.56, 0.55, 0.52, 0.48, 0.38, 0.0</td>
<td>0.1, 0.16, 0.16, 0.17, 0.32, 0.49, 0.58</td>
</tr>
</tbody>
</table>

**ML**

Values produced by the ML method—which represent the percentage of stimuli in the first part of the trial that were labeled as a probe (i.e., another irrelevant item coded as the probe for analysis purposes for innocent Ps), were higher in the SG group compared to the innocent, but not significantly so (SG mean=.466; Innocent mean=.283; \( t[1,13]=1.804, p=.094, \eta^2=.200, JZS BF=1.039, \) favoring the null). Using a P-specific criterion of % target sweeps classified as probes, the ML algorithm produced an overall diagnostic accuracy of 73.33% (SG 5/8=62.5%; Innocent 6/7=85.71%).

**ROC Analysis**

ROC curves derived from the bootstrapping and ML methods are displayed in Figure 3. While results comparing diagnostic methods showed a higher AUC when the bootstrapping approach was used (.872 vs. .712 for ML), standardized z-values revealed that diagnostic accuracies did not differ significantly (\( Z=1.06, p>.2, \) two-tailed).
Discussion

Our goal was to compare the traditionally used bootstrapped amplitude difference method with a novel ML approach designed to diagnose concealed knowledge, using the most empirically validated and countermeasure-resistant version of the P300-based CIT the CTP. While we expected both diagnostic approaches to yield high detection rates for both the SG and innocent groups, it was less clear whether the theoretical benefits of the ML approach, and the promising results of previous research using ML algorithms for P300-based diagnosis of concealed information, would be supported. To investigate this, we applied both methods to a single dataset which included both an innocent and a SG condition, and then compared standardized $z$-scores to determine if one approach outperformed the other.

Visual inspection of the SG and innocent ERPs (see Figure 1) shows an obvious CIT effect in the former but not the later, as expected. This was supported by data suggesting that the elevated P300 response in the SG compared to the innocent condition was specific to the probe, and group differences in the CIT effect, as measured by probe-Iall amplitudes, number of bootstrapped...
iterations where the probe > Iall, and average bootstrapped mean probe-Iall differences. This effect is also suggested by the ML outcome data showing the % probe sweeps classified as a probe, which was marginally larger (p < .1) in the SG group.

However, while these expected group differences in the CIT effect emerged, individual hit rates using the well-vetted bootstrapping method were lower than typically observed. A few limitations are likely to blame. The first—limited exposure to the probe—is a result of probing for episodic information (i.e., jewelry stolen during a mock crime). Because semantic information is well-rehearsed and salient, it often produces P300 responses large enough to obscure differences in accuracies between the bootstrapping and ML approaches. In attempt to prevent these ceiling effects, we used episodic information, which may not have been adequately encoded during the very brief exposure to the stolen item during the mock-crime.

Other factors that may have contributed to suboptimal accuracy rates are stimuli selection and presentation. For example, because CTP stimuli presentation occurs so quickly, guilty Ps may not have recognized the item they stole, reducing its P300. Alternatively, guilty and innocent Ps may only be sure or unsure about what a few items in the stimuli set are, which is problematic for multiple reasons. First, this uncertainty might consume cognitive resources, for example, if a P focuses on the item in an attempt to identify it, or is worried the experimenter will quiz them on an ambiguous item. In guilty Ps, increased cognitive load could decrease the probe’s amplitude (Donchin, Kramer, & Wickins, 1986) and reduce the expected CIT effect (probe-irrelevant difference). When unsure of what items are, Ps may also make unintended associations between them. For example, Ps might think the cufflinks are earrings (which appear as another irrelevant), or a guilty P might think the irrelevant image of a ring is also a bracelet (their probe item). In these cases, irrelevants could generate an oddball response that would harm diagnostic accuracy.
To reduce item confusion in the future, possible solutions include showing all Ps the images they will see on the subsequent CIT (with labels), or using words (e.g., “bracelet”) as stimuli. The false positives seen in the SG group also highlight why it is important to test all stimuli sets on a known innocent group before tests are administered to suspects, so that irrelevant items that produce an elevated response can be removed.

To test the diagnostic efficacy of each approach, our main comparison of interest was between the AUC values derived from each method. While AUCs were rendered by the bootstrapping procedure >15% higher, standardized $z$-scores did not suggest the superiority of one approach over another. However, the findings presented here are admittedly limited by their sample size, and while $z$-scores allow for standardization of unequal sample sizes so AUCs can be compared, our findings would have benefited from more data. As such, we plan compare the ML and bootstrapping methods again, including the data of a recently run study that included an innocent condition (to allow for ROC analysis).

Conclusions and Future Directions

While findings presented here do not support that the ML algorithm should replace the currently used bootstrapping method for diagnosis using the CTP, further refinement of the algorithm and inclusion of additional features might improve its accuracy. For example, since probe P300 latencies and behavioral reaction times, and target/nontarget error rates are often elevated in guilty individuals, algorithms might benefit from incorporating non-neural features. Indeed, combining ERP and behavioral measures has been shown to outperform single outcome diagnosis (Hu & Rosenfeld, 2012), and might offer a promising avenue for future investigation. While the ML algorithm not improve CIT diagnostic accuracy, it may be useful in clinical contexts where P300 is used to make diagnoses based on subjective clinician judgment, which
may benefit more from adopting a ML approach than the already objective bootstrapping method we used for comparison here.
REFERENCES


Hu, X., & Rosenfeld, J. P. (2012). Combining the P300-complex trial-based Concealed
Information Test and the reaction time-based autobiographical Implicit Association Test in concealed memory detection. *Psychophysiology*, 49(8), 1090-1100.


Meixner, J. B., & Rosenfeld, J. P. (2011). A mock terrorism application of the P300-based


the IEEE 84 (4), 626–638.
Chapter 1

Appendix A Mock-Crime Instructions

You have been assigned to the guilty condition. As part of your participation, we ask that you complete the lab crime task as follows: Your assignment is to steal an item from a faculty member’s office and then take it to a mailbox located in the basement of Swift. With this key, you will unlock Office 205. The door is located in the main room of this lab where the couch is. Before entering, knock to make sure no one is in the office. If no one answers, use the key to enter. Once inside, you will see a desk. In the top drawer of the desk, you will find a yellow envelope containing the item you will steal. Take the item out of the envelope, examine it, and then hide it on your person, leaving the envelope in the drawer. Next, you will go hide the item in a mailbox located in the basement of Swift Hall. To get there, exit the lab, turn right, and go up the stairs at the end of the hallway to the 3rd floor of this building (Cresap). When you get to the top, turn left and enter Swift. Keep walking until you go through one closed door into the Swift stairwell. Go down the large staircase on your left to basement. You will see many mailboxes in front of you. Find the mailbox with the name “WARD” on it. Place the stolen item in the mailbox and return to the lab. Make sure to return the key to your experimenter. Under no circumstances should reveal the details of the theft to anyone, including your experimenter. From this moment on, try to make sure that nobody notices your actions. If someone notices you, please carry on and act naturally. Now, go steal the item from the office and hide it in the mailbox. Let your experimenter know if you have any questions.

Appendix B Interference/Control Instructions

**Interference Instructions**

“You will be asked to complete a brainwave-based test designed to identify the item you stole. However, I’m going to read to you a description of an event similar to the one you took part in. The purpose of this scenario is to distract you from the crime you have committed. If you listen carefully and remember all the details of the scenario, it might help you beat the brainwave-based test you will take. If you are successful, you will receive a $5 bonus and the end of the test. Notice that in the description I’ll read, there are details that are different from the real details you’ve seen. After I’ll read you the scenario, you’ll be tested briefly about its details.”

**Interference Manipulation (read twice)**

Now, please sit still, close your eyes, and imagine the next scenario: Now that you are relaxed, pretend leaving the chair and walking towards OFFICE 204. After knocking to make sure OFFICE 204 is empty, you enter using the key in your hand. You open the door and see a small room containing a DRESSER. You open the BOTTOM drawer of the DRESSER. Now you look for an envelope. In the BOTTOM drawer, you find a BLUE envelope. In the BLUE envelope, you find CUFFLINKS/TIARA/BROOCH (randomized). You examine the CUFFLINKS/TIARA/BROOCH and hide them on yourself. Before you leave OFFICE 204, you make sure again that you have the CUFFLINKS/TIARA/BROOCH that were in the BLUE envelope you found in the BOTTOM drawer of the DRESSER. Next, you hurry to leave the room for the mailbox. You walk to the basement, put the items in the mailbox, and return to the lab.”

**Interference Memory Check**

Please fill in the details you have memorized

I arrived to Office and opened the door with a key. In the room there was a . In its drawer, I found a colored envelope. I opened it and found a . I took it and hid it in the mailbox, and then returned to the lab.

**Control Instructions**

“Now you will be asked to complete a word find for a few minutes. Please fill it as quickly as possible. When I ask you to stop, please stop even if you aren’t finished. The goal of the puzzle is to distract you from the crime you have committed. If you concentrate on the puzzle, it might help you beat the brainwave-based test you will take, which is designed to identify the item you stole. If you are successful, you will receive a $5 bonus and the end of the test.”

Appendix C CTP Instructions

Your experimenter does not know the details of the task you have performed, so please do not reveal them to the experimenter. During their visit, some participants were asked to steal an item from an office. We want to see if our test can identify whether you stole an item or not. Your goal is to appear innocent on the test. If you succeed
in appearing innocent, you will receive a $5 bonus. In this experiment, sensors will be placed on your scalp (3), face (2), and behind your ears (2) so that we can record brainwaves. Harmless conductive paste will be applied to the electrodes and is easily removed. A sink, shampoo, and towels are available if you want to clean up after the study. This experiment lasts about 30 minutes. During the brainwave test, you will see images on the blank screen in front of you. If you need glasses to see clearly, you MUST be sure to wear them. Do NOT wear contacts. If you are chewing gum, please remove it now.

Each trial will have 2 stimuli. THE FIRST STIMULUS in each trial will be a picture of an item. Many items will be presented on many trials. One of the items was stolen from the faculty member. However, no matter which stimulus is presented, using your left index finger, you must press the right button on the left mouse to indicate that you have seen the image. Remember, no matter which stimulus is presented, you should press the button as soon as possible. THE SECOND STIMULUS, which appears later in the trial, will be a string of numbers (11111, 22222, 33333, 44444, or 55555). The 11111 string is your target. If the second stimulus is the target string (11111), use the middle finger of your right hand to press the YES button on the right-hand mouse. If the second stimulus is 22222, 33333, 44444, or 55555, use the index finger of your right hand to press the NO button on the right-hand mouse.

PAY ATTENTION, ESPECIALLY TO THE FIRST STIMULUS! At a few random points during testing, the experimenter will pause the stimulus presentation and ask you to identify the first stimulus in that trial. More than one incorrect identification 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 stimuli appear. It is important that you do not blink, move your eyes, or activate any facial/head muscles while a stimulus is on the screen. When you need to blink, try only to do so between trials when a crosshair (+) appears on the screen. If you accidentally blink or twitch during a trial, don’t worry about it – just prepare yourself for the next trial. Please let your experimenter know if you have any questions.

Appendix D Control and Interference Post-Experiment Questionnaires

Control Questionnaire

In the brainwave test, you were asked to distract yourself from processing crime-relevant information by focusing on details that were different than the lab crime you completed. Sometimes people fail to follow these instructions because they suspect there is something else going on in the study that they are not told about, because they misunderstood the instructions, or for other reasons. It's very important for us to know if this happened in the brainwave test so that we can understand the results of our study. Therefore, we need you to answer the questions below. There is no penalty or cost for any answer, so please be as honest as possible, and note that your responses are completely anonymous and will not affect your participation in any way.

1. Please rate the extent to which you followed the instruction to focus on the word find we provided you with

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No focus</td>
<td>Decent focus</td>
<td>Excellent focus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Please rate the extent to which you focused on the word find during the brainwave test (if at all)

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No focus</td>
<td>Moderate focus</td>
<td>Extreme focus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Please rate your confidence level in beating the test

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No confidence</td>
<td>Not sure</td>
<td>Extremely confident</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. During the brainwave test, when presented with the real item you took from the office and hid in the mailbox, how much did you think about the lab crime you completed?

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Sometimes</td>
<td>Almost every time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. How motivated were you to beat the test?

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Somewhat motivated</td>
<td>Extremely motivated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Besides what we asked you to do, do you do anything else to try to beat the brainwave-based test you just took? (Please circle) YES NO
6b. If yes, what did you do?

****Please only complete the questions below if you had 2 study appointments.

7. How much do you think you forgot the details of the crime between your first and second appointment?

0  1  2  3  4  5  6
No forgetting  Moderate forgetting  No remembering

8. What office did you enter in order to steam an item?

9. Where was the envelope located?

10. Was it in the bottom or top drawer?

11. What color was the envelope you took the item out of?

12. What was the item you took and hid in the mailbox?

Interference Questionnaire

In the brainwave test, you were asked to distract yourself from processing crime-relevant information by focusing on details that were different than the lab crime you completed. Sometimes people fail to follow these instructions because they suspect there is something else going on in the study that they are not told about, because they misunderstood the instructions, or for other reasons. It's very important for us to know if this happened in the brainwave test so that we can understand the results of our study. Therefore, we need you to answer the questions below. There is no penalty or cost for any answer, so please be as honest as possible, and note that your responses are completely anonymous and will not affect your participation in any way.

1. Please rate the extent to which you followed the instruction to focus on the alternative scenario we provided you with while it was being read to you by the experimenter

0  1  2  3  4  5  6
No focus  Decent focus  Excellent focus

2. Please rate the extent to which you visualized the alternative scenario we provided you with during the brainwave test

0  1  2  3  4  5  6
No visualization  Moderate visualization  Extreme visualization

3. Please rate your confidence level in beating the test

0  1  2  3  4  5  6
No confidence  Not sure  Extremely confident

4. Please rate how difficult you found it to distinguish between the details of the real lab crime you completed and the alternative details we provided you with

0  1  2  3  4  5  6
Extremely easy  Moderately easy  Extremely difficult

5. During the brainwave test, when presented with the real item you took from the office and hid in the mailbox, how much did you think about the real lab crime you completed?

0  1  2  3  4  5  6
None  Sometimes  Almost every time

6. How motivated were you to beat the test?

0  1  2  3  4  5  6
None  Somewhat motivated  Extremely motivated

7. Besides what we asked you to do, do you do anything else to try to beat the brainwave-based test you just took? (Please circle)

YES  NO

7b. If yes, what did you do?

****Please only complete the questions below if you had 2 study appointments.

8. How much do you think you forgot the details of the crime between your first and second appointment?

0  1  2  3  4  5  6
No forgetting  Moderate forgetting  No remembering

9. What office did you enter in order to steam an item?

10. Where was the envelope located?

11. Was it in the bottom or top drawer?

12. What color was the envelope you took the item out of?
13. What was the item you took and hid in the mailbox?

Chapter 2

Appendix E: Mock-Crime/Control Instructions (e.g., ring, bracelet, and earrings were randomly assigned as the probe).

Mock-Crime
As part of your participation, we are asking you to commit a crime. Specifically, your assignment is to hide a stolen ring. After doing so, you will undergo a brainwave-based test. Your researcher will remain unaware of who stole the ring and its hiding spot, so the goal of the test is to determine who stole the ring and where it is hidden. Although you can take these instructions with you, please read them through completely before you complete your crime, and let your experimenter know if you have any questions.

Steal the ring
- Go to the desk by the windows.
- On the left side of the desk, there are two drawers.
- Open the top left-hand drawer, where you will find the ring you will steal.
- Take a moment to examine the ring very carefully, making sure to pay attention to all its details.
- Conceal the ring on yourself.

Hide the ring
- Walk toward the front door (as if you are going to exit).
- On the way, you will see a tall black cabinet.
- Open both doors of the cabinet.
- Inside, you will see several containers.
- Select and pick up ONE container (minimally touching any of the others).
- Take a moment to examine it very carefully, making sure to pay attention to all of its details.
- Open the container.
- Inside, you will see several smaller containers.
- Select and pick up ONE container (minimally touching any of the others).
- Take a moment to examine it very carefully, making sure to pay attention to all of its details.
- Open the container.
- Take the stolen ring, place it inside the smaller container, and close it.
- Place the smaller container (now holding the ring) back inside the larger container.
- Close the larger container.
- To clarify, the ring should be inside the smaller container, which is inside a larger container.

- Next, return to the testing room for the brainwave-based test.

Control
Enter Office 203 (office numbers are located on doors in the main room where the couch is). As soon as you enter you will see drawers on your left. Open the top left drawer, where you will find a blank piece of paper and a pen. Sign your name on the piece of paper to document your participation in the experiment today. Next, find your experimenter to let them know you are ready for the next phase of the study. Before you being, let your experimenter know if you have any questions.

Appendix F CTP Instructions
Before we start

This part of the study will last about 45 minutes, so please let your experimenter know now if you need to use the restroom, if you are wearing contacts, or if you need glasses to see clearly and do not have them. Please remove any gum/candy and place your cell phone (on silent or off) in the room with the experimenter or with your belongings. In this experiment, harmless sensors and gel will be placed on your scalp (3), face (2), and behind your ears (2) to record your brainwaves. Materials are available if you wish to clean yourself up after the experiment.

Purpose

An item has been stolen from the lab and hidden somewhere. The test you are about to take is very accurate and is designed to identify where this stolen item is hidden. If you are innocent of this crime, your brainwaves should show us that you lack knowledge about the crime’s details, and it is your goal to be determined innocent by the test. However, if you are guilty of committing the crime, we expect that your brainwaves will show us which item you stole and where you hid it. If you are guilty, it is your goal to beat the test by concealing this information.

Test Instructions

Two alternating types of stimuli will appear on the screen in front of you pictures and numbers. Any time you see a picture, no matter which picture is presented, press the R button on the L mouse as soon as possible (using your L index finger), indicating you have seen it. After each picture (and a brief “+” in the middle of the screen), a string of numbers (11111, 22222, 33333, 44444, or 55555) will appear. “11111” is your target, so when you see it, use the middle finger of your R hand to press the “YES” button on the R-hand mouse. When you see any other number strings, use the index finger of your right hand to press the “NO” button on the R-hand mouse.

At a few random times, the experimenter will pause the test and ask you to verbally identify the last picture you saw. If you respond incorrectly more than once, your data will be unusable, so pay attention to the images on the screen. Please try to remain still throughout testing and keep your eyes focused on the middle of the screen. When you need to blink, try to do so when the “+” is on the screen. If you make a mistake, don’t worry about it and just prepare for the next trial.

You will complete three versions of this test, each lasting under 15 minutes. You will get a short break between each test, but if you get too tired in the middle of a test, let your experimenter know so the test can be paused. Before we begin, please let your experimenter know if you have any questions.

Appendix G Stimulus sets.

Block 1

![Stimulus images]
Sometimes people fail to follow instructions during a study because they suspect there is something else going on in the study that they are not told about, because they misunderstood the instructions, or for other reasons. It's very important for us to know if this happened today so we can understand the results of our study. So, please answer the questions below as truthfully as possible. There is no penalty or cost for any answer, and your responses are completely anonymous and will not affect your participation in any way.

***The following questions refer to the 1st brainwave-based test containing the following items*** (See Appendix G for images)

1. Aside from the jewelry item you stole, did any of the other jewelry items you saw during the brainwave-based test stand out to you among the others?
   
   YES       NO
1b. If yes, which jewelry item, and why do you think it stood out to you?

2. How likely do you think it is that we could detect that you stole the jewelry item?

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely unlikely</td>
<td>Likely</td>
<td>Very likely</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Did you do anything to try to beat the 1st brainwave-based test you took? (If no, skip to Q4)

YES NO

3b. If yes, what did you do?

______________________________________________________________________________

3c. If yes, please rate your confidence level in beating the test

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No confidence</td>
<td>Not sure</td>
<td>Extremely confident</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

***The following questions refer to the 2nd brainwave-based test containing the following items*** (See Appendix C for images)

4. Please circle the above container that you selected.

4b. Aside from the container you selected, did any of the other containers stand out to you among the others?

YES NO

4c. If yes, which container, and why do you think it stood out to you?

______________________________________________________________________________

5. How likely do you think it is that we could detect the container you chose?

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extremely unlikely</td>
<td>Likely</td>
<td>Very likely</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Did you do anything to try to beat the 2nd brainwave-based test you took? (If no, skip to Q7)

YES NO

6b. If yes, what did you do?

______________________________________________________________________________

6c. If yes, please rate your confidence level in beating the test

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No confidence</td>
<td>Not sure</td>
<td>Extremely confident</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

***The following questions refer to the 3rd brainwave-based test containing the following items*** (See Appendix C for images)

7. Please circle the above item that you selected.

7b. Aside from the item you selected, did any of the other items stand out to you among the others?

YES NO

7c. If yes, which item, and why do you think it stood out to you?
8. How likely do you think it is that we could detect the item you chose?

0 1 2 3 4 5 6
Extremely unlikely Likely Very likely

9. Did you do anything to try to beat the 2nd brainwave-based test you took? (If no, skip to Q10)

YES NO

9b. If yes, what did you do?

9c. If yes, please rate your confidence level in beating the test

0 1 2 3 4 5 6
No confidence Not sure Extremely

THANK YOU FOR YOUR PARTICIPATION!

10. If you would like to tell your experimenter anything else about your experience, please do so below

________________________________________________________________________

Innocent Questionnaire

Sometimes people fail to follow instructions during a study because they suspect there is something else going on in the study that they are not told about, because they misunderstood the instructions, or for other reasons. It’s very important for us to know if this happened today so we can understand the results of our study. So, please answer the questions below as truthfully as possible. There is no penalty or cost for any answer, and your responses are completely anonymous and will not affect your participation in any way.

***The following questions refer to the 1st brainwave-based test containing the following items *** (See Appendix C for images)

1. Did any of the jewelry items you saw during the brainwave based test stand out to you among the others?

YES NO

1b. If yes, which jewelry item, and why do you think it stood out to you?

________________________________________________________________________

2. How likely do you think it is that we could detect that you did not steal anything?

0 1 2 3 4 5 6
Extremely unlikely Likely Very likely

3. Did you do anything to try to beat the 1st brainwave-based test you took? (If no, skip to Q4)

YES NO

3b. If yes, what did you do?

3c. If yes, please rate your confidence level in beating the test

0 1 2 3 4 5 6
No confidence Not sure Extremely
***The following questions refer to the 2nd brainwave-based test containing the following items*** (See Appendix C for images)

4. Did any of the above containers stand out to you among the others?
   
   YES  
   
   NO

4b. If yes, which container, and why do you think it stood out to you?

______________________________________________________________________________

5. Did you do anything to try to beat the 2nd brainwave-based test you took? (If no, skip to Q6)

   YES  
   
   NO

5b. If yes, what did you do?

______________________________________________________________________________

5c. If yes, please rate your confidence level in beating the test

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No confidence</td>
<td>Not sure</td>
<td>Extremely</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

***The following questions refer to the 3rd brainwave-based test containing the following items*** (See Appendix C for images)

6. Did any of the other items stand out to you among the others?

   YES  
   
   NO

6b. If yes, which item, and why do you think it stood out to you?

______________________________________________________________________________

7. Did you do anything to try to beat the 2nd brainwave-based test you took? (If no, skip to Q8)

   YES  
   
   NO

7b. If yes, what did you do?

______________________________________________________________________________

7c. If yes, please rate your confidence level in beating the test

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>No confidence</td>
<td>Not sure</td>
<td>Extremely</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

THANK YOU FOR YOUR PARTICIPATION!

8. If you would like to tell your experimenter anything else about your experience, please do so below ____________________________________________________________

Chapter 3

Appendix I

Pre-Task Instructions “The project in which you are about to participate focuses on lie-detection tests. This is of major interest in the field of applied psychophysiology. Criminals who are questioned about a certain crime or crime scene show brain wave responses when they see test item stimuli which they recognize from the crime.” TO SG GROUP ONLY “In the next time period you will commit a laboratory crime, and afterwards, you will take our brainwave test.”

Mock-Crime Instructions “Go into room 203E, the last door on your left nearest the window as you enter the lab. In room 203E, as you enter, there will be a set of 8 drawers on your left. In the topmost drawer on the left, you will find a padded envelope with a valuable item inside. Open the envelope, take it out, hold it in your hand and rotate it as you examine it from all angles, then put it in your pocket and exit the room without letting the
experimenter know which item it is. Return to the room where you signed the consent form. Any questions? Now, exit the room and perform the robbery.’”

**Innocent Instructions** “Go into room 203E, the last door on your left nearest the window as you enter the lab. In room 203E, as you enter, there will be a set of 8 drawers on your left. In the topmost drawer on the left, you will find a padded envelope. Open the envelope, look inside, and then return to the room where you signed the consent form. Any questions? Now, exit the room and complete your task.”

**Appendix J CTP Instructions**

**Pretest Instructions**

In the present scenario, you are suspected of having committed a crime. In order to find out if you are guilty or innocent of the crime, you will take a brain wave lie-detection test during which we will measure your brain wave responses. For this reason I have attached the electrodes to your head. In the test, several items will be presented on this screen to you, including the image corresponding to the item you are suspected of taking. The test is based on the theory that your brain wave responses get bigger when you recognize an item related to the crime.

**Test Instructions**

These tests will be administered on the computer and I will monitor them myself. In this experiment, electrodes will be placed on your head and behind your ears. A harmless conductive paste is used to apply the electrodes; it can be easily cleaned off. A sink, water, and towels are available in the lab if you wish to clean yourself immediately afterwards.

Do NOT wear contact lenses. If you need glasses, you MUST be sure to wear them, not contact lenses. You will be seated in a comfortable chair while a series of stimuli are presented on a screen in front of you. The stimuli will be presented in pairs. Each pair includes an item – that would be the first stimulus in a pair and it will be followed by a string of numbers 11111, or, 22222, or, 33333, or, 44444, or, 55555) – that would be the second stimulus in the same pair. There will be 210 pairs presented in a run.

**THE FIRST STIMULUS.** When you see the first stimulus in a pair, you have to immediately press the RIGHT button using your index finger on the LEFT-HAND mouse to indicate that you have seen the item. Some images might be familiar or relevant to you, but others will have no relevance to you. No matter if an item is relevant or irrelevant to you, you should respond to all of them exactly the same way (with a right button press on the left mouse). Remember, no matter which stimulus is presented, you always want to press the button as soon as possible.

**THE SECOND STIMULUS,** which you see a bit later in the trial, will be a string of numbers and it will appear soon after the first stimulus. The 11111 string is your target. When you see the target string 11111, respond immediately by pressing the RIGHT (“YES”) button on the RIGHT-HAND mouse. If the second stimulus is 22222, 33333, 44444, or 55555, to press the LEFT (“NO”) button on the RIGHT-HAND mouse.

**PAY ATTENTION TO ALL STIMULI!** At a few random times during testing, the experimenter will pause the stimulus presentation and ask you to identify the last stimulus you saw on the screen. More than one incorrect identification will result in unusable data and early termination of your participation.

It is crucial that you control your blinking, trying to reduce it as much as you can. If needed, you can blink when you see numbers (not when you see an item).

Remember during the course of a trial, keep your eyes focused on the center of the screen where the stimuli appear. Please try to sit still, relax all your facial and body muscles and try to minimize blinking.

Thank you!