Biological Psychology 183 (2023) 108666

Contents lists available at ScienceDirect

Biological Psychology

journal homepage: www.elsevier.com/locate/biopsycho

Examining levels of processing using verbal & pictorial stimuli with the complex trial protocol in a mock theft scenario☆

Michel Funicelli [a,*,1,2], Sarah Salphati [b], Sabina Ungureanu [b], Jean-Roch Laurence [b,3]

[a] Policing at Canterbury Christ Church University, UK
[b] Concordia University, Canada

ARTICLE INFO

Keywords:
Complex Trial Protocol
Concealed Information Test
P300-based CIT
Memory recognition
Levels of processing

ABSTRACT

The Complex Trial Protocol (CTP) is an EEG-based Concealed Information Test (CIT). Depth of processing influences memorability where deeper processing increases recollection. The CTP's performance as a function of shallow versus deep levels of processing has not been explored. Two experiments were conducted, one with verbal stimuli and the other with their pictorial referents. In both experiments, participants were randomly assigned to three groups, Innocent Control (Control) condition, Guilty Immediate Shallow Processing (Shallow) condition, and Guilty Immediate Deep Processing (Deep) condition. Shallow and Deep participants from both experiments underwent the same mock theft scenario and all three groups were later exposed to either a verbal (N = 41) or pictorial (N = 43) stimulus on a computer monitor. In the word study, no differences in CIT effect were found between any of the groups. Areas under the curve (AUCs) did not differ from chance (.624 and .679 for Shallow and Deep groups respectively). In the image study, the CIT effect for the Shallow and Deep groups differed from the Control one. The AUCs (.755 and .943 for the Shallow and Deep groups respectively) differed significantly from each other. Levels of Processing (LOP) did not appear to have any bearing on CTP performance when words were used as probes but did have an effect when images were used. The findings may hint at some of the limitations of the CTP and fail to replicate similar experiments, especially when words are used as probes, from the late Peter Rosenfeld's laboratory.

1. Introduction

Police forces have many reliable investigative tools at their disposal to collect physical traces from a suspect's body and a crime scene from which incriminating or exculpatory evidence can be analysed and documented. Several techniques (e.g., CIT & CQT-based polygraph) have used autonomic nervous system (ANS) responses to provide evidentiary clues in the legal and investigative process. Although psychophysiological instruments measuring central nervous system (CNS) responses, such as the P300 CIT, have been able to produce very good efficiency estimates in the laboratory, the method has not been applied in a real criminal investigation. Whether it could be as efficient in an actual investigation is still an open question?

A potential candidate that has been attracting the attention of social scientists for over half a century is the P300-based Concealed Information Test (CIT). First introduced in 1959 by David Lykken, the CIT was initially designed as a physiological technique to identify individuals in mental possession of crime relevant information by tapping into a person's autonomic nervous system (ANS) and measuring their electrodermal activity (Lykken, 1998). The CIT has since been expanded to collect signals directly from the central nervous system (CNS) (e.g., brainwave activity). One such method is a psychophysiological CIT protocol applicable to EEG instrumentation and better known as the Complex Trial Protocol (CTP) (Rosenfeld et al., 2008).

Regardless of the origin of the signal being detected, ANS or CNS, the fundamentals of a CIT remain the same; an individual is presented with two types of stimuli, criminally pertinent details, called *probes*, known only by the offender and the authorities, and plausible but neutral

☆ The authors wish to express their gratitude for the reviewers' insightful comments.
* Corresponding author.
E-mail address: michel.funicelli@canterbury.ac.uk (M. Funicelli).
[1] ORCID: 0000–0003-2014–520X
[2] lead author
[3] ORCID: 0000–0001-9897–7466

https://doi.org/10.1016/j.biopsycho.2023.108666
Received 13 January 2023; Received in revised form 21 August 2023; Accepted 21 August 2023
Available online 1 September 2023

alternative items, called *irrelevants*. Because the probe is a relevant piece of information to the crime under investigation, and only known by its perpetrator and police, an inference of guilt or innocence could be drawn by the trier of fact from a positive or negative CIT, respectively. The reaction generated from the rare probe presentation versus the more frequent exposure to irrelevants functions as an oddball paradigm (see Polich, 2007 for a complete review of P300 and Lukács et al., 2016 for an explanation of the oddball effect within the CTP). This involuntary cerebral manifestation is considered "a good guilty knowledge index" (Rosenfeld, 2019, p.2). For instance, in a situation where the theft of a woman's purse is committed, the probe item could be the word '*purse*' and the irrelevants could be words like *laptop, wallet, phone*, etc. or their respective pictorial referents. A significant increase of a P300 response to the meaningful word '*purse*', or its image, would lead an examiner to conclude that a person under investigation for the theft of the purse is *'information-present'* from which an inference of guilt could be drawn by judicial authorities. On the other hand, an innocent person would be expected to react similarly to the word/image '*purse*' as s/he would to the words/images *laptop, wallet, phone*, etc. since they are all meaningless to the examinee, and be deemed an *'information-absent'* innocent person following a significant negative CIT result (Labkovsky & Rosenfeld, 2014).

Nevertheless, the CIT does present with one important vulnerability. An ERP-based CIT's rationale is predicated on the notion that a guilty person (i.e., true positive) will manifest a P300 probe-related amplitude to a crime pertinent detail significantly greater than the P300 values for irrelevant-related neutral alternatives, a reaction called the '*CIT effect*' (Olson et al., 2022). On the other hand, an innocent person (i.e., true negative) will exhibit a similar waveform to probe and irrelevant items, signalling a non-recognition of the crime pertinent information. However, if an innocent suspect is exposed to crucial crime items from media leakage or during a poorly conducted police interview, a CIT places the examinee at risk of being classified as a false positive, regardless of the CIT's modality (Meijer et al., 2014; see also Kim et al., 2022, for a promising approach to solving this issue). Conversely, if a guilty individual is summarily in contact with the criminal probe, how likely is that person to be accurately diagnosed?

1.1. Processing levels, memory, and the CTP

A question that has not yet been answered is whether the CTP's performance is influenced by the depth of processing of examinees who have been exposed to a crime detail in a mock crime scenario. The seminal work of Craik and Lockhart (1972) and Craik and Tulving (1975) inform us in the way memory is processed, stored, and retrieved, and thus it is helpful in our appreciation of the functioning of the P300-based CIT. Their investigations revealed two lines of evidence in relation to levels of processing (LOP): (1) stimuli that are sensorially attended only at a low level of analysis, or shallowly encoded, will result in evanescent memory traces, (2) stimuli that received complete attention, and are further enriched by images, or deeply encoded, leave a longer lasting trace. The conclusion of their findings was that deeply encoded information was associated with improved memorability on subsequent retrieval tests. Similar findings were found by Seymour and Fraynt (2009) in a Concealed Knowledge Test where shallow and deep study procedures were employed, and Reaction Time was measured at various time intervals. To all intents and purposes, the CTP's performance in detecting P300 brainwaves of memorised crime related items should follow the same reasoning and fare better the more profound such items are committed to memory.

Rosenfeld and colleagues have published over two dozen articles to date using the CTP (see Rosenfeld et al., 2013 for a review). The CTP's performance has been tested through several research paradigms involving a variety of verbal or pictorial stimuli (i.e., the word or image of the stolen item). In about half of these studies, some form of autobiographical probe was used (e.g., participant's hometown, family name, mother's first name, or a participant's recent life event). The CTP has been shown to perform well with autobiographical data such as given or family names with detection rates in most of these studies, ranging between 90% and 100% (Rosenfeld et al., 2008; Lukács et al., 2016; Rosenfeld et al., 2017; Rosenfeld, Ward, Drapekin, et al., 2017; Funicelli et al., 2021). However, in studies that looked at the word or the image of an alleged stolen item (a ring or a watch), the results are less clear cut. While it can be assumed that autobiographical items have been deeply processed, the same cannot be said of items that were only looked at or handled briefly. In fact, in their 2013 review of the CTP studies, Rosenfeld and colleagues recognized the superiority of autobiographical probes over mock crime probes. Since none of the published studies investigated directly the role played by different levels of cognitive processing during a mock theft scenario, the chief objectives of the present study were to investigate this gap in the literature with the CTP for either words or images that have no autobiographical relevance, and to evaluate and report on the diagnostical performance of the CTP with non-autobiographical words and images.

1.2. Mock crimes stimuli as probes

To date, the CTP has been used in conjunction with other tasks and with a variety of stimuli. Twelve studies have been published where words or images of a stolen item (e.g., ring) were used as probes (See Table 1).

Of these 12 studies, two did not report ROC curve analyses (Rosenfeld et al., 2015; Rosenfeld et al., 2018). When looking at the AUCs reported, the mean AUC for words (10 studies) and for images (3 studies) is 0.88 in each case. So overall, the CTP as a diagnostic test in a group of guilty or innocent participants seems good. Individual

Table 1
Authors (years of publication), protocol type, the type of stimulus used as probes, and AUC when available.

Authors (year)	Protocol Type	Stimulus Type	AUC Results
Winograd and Rosenfeld (2011)	CTP	Word "Ring"	AUC = .93
Hu and Rosenfeld (2012)	CTP + RT-aIAT	Word "exam"	$AUC_{p\text{-}p}$ = .92
Labkovsky and Rosenfeld (2014)	CTP (pictorial) + 3 Stimuli protocol (verbal)	Picture (USB key) & word: (name "Meixner")	AUC = .94 verbal AUC = .89 image
Meixner and Rosenfeld (2014)	CTP	Word (name of friend, city visited, ordinary word)	AUC = 1.0
Winograd and Rosenfeld (2014)	CTP	Word "Ring"	AUC_{gn} = .852 AUC_{gi} = .956
Rosenfeld et al. (2015)	CTP	Picture and word of USB key, ring, keys, iPod, pen, coin	AUCs not reported
Hu et al. (2015)	CTP	Word "Ring"	AUC_{sg}:.84
Sai et al. (2016)	Feedback CIT	Word "Ring"	$AUC_{p\text{-}p}$ = .68 $AUC_{b\text{-}p}$ = .73
Ward and Rosenfeld (2017)	CTP	Word "Ring"	AUCs = .95*
Lu et al., 2017	CTP	Picture of jewelry items	AUC_{sg} = .84 based on simulated innocent group
Rosenfeld et al. (2018)	CTP	Picture or word (not specified) of a watch or bracelet	AUCs not reported
Ward et al. (2020)	CTP	Picture of either a watch, bracelet, or necklace	AUCs = .92*

Note: p-p = Peak to peak, gn = Guilty Naïve, gi = Guilty Informed, sg = Simply Guilty, b-p = Base to Peak. * = These AUCs were supplied by one of the reviewers and based on the Grier (1971) formula.

classification however does not always reflect this overall performance. For example, in Hu and Rosenfeld (2012), even though the reported AUC was.92 which is very good as an index of overall performance, only 66% of guilty participants tested immediately and 75% of those tested at a one-month delay were correctly identified. This discrepancy may be problematic if and when the CTP is used in the legal arena where it would be applied to a single individual.[4]

An objective of this current research was, in essence, to conduct a quasi reproduction of the work accomplished by the late Rosenfeld and his colleagues, with respect to the independent use of words or images in a mock crime scenario with the added LOP variable. As it applies to this study, we can surmise that the Guilty-Informed[5] group in Winograd and Rosenfeld (2014) was exposed to a deeper LOP, through the prior exposure of the stimulus word 'ring' contained in the pre-test instructions, and that the Guilty-Naïve[6] group could constitute the shallow processing condition. In this study the AUC for the guilty-informed group went up to .956 while the naïve group's AUC was .852.

Overall, the question remains as to the efficacy of the CTP in mock crimes experiments when words or images of the object stolen are used as probes. It must be noted here that the use of the CTP in a mock theft scenario has not been replicated at this point by a complete team of independent investigators. Even though two studies suggest that depth of processing may be playing a role in the CTP methodology (Winograd & Rosenfeld, 2014; Deng et al., 2016), depth of processing was not directly addressed in any of the studies outlined above. Furthermore, Gamer and Berti (2012) reported that depth of processing may not play an important role in P300 elicitation when the probes do not have an autobiographical connotation and are learned incidentally during the commission of a mock theft. The role of processing depth is somewhat ambiguous at this point in the combined CIT and CTP literature. If the CTP is as reliable as some of the studies from the Rosenfeld laboratory indicate, then it is important that a replication of a similar mock crime be undertaken. In addition, understanding the minimum LOP required for a successful CTP diagnostic may prove to be essential for police investigators in their decision to test or not the suspect of a crime. For example, if it is determined that only a shallow LOP is necessary to obtain a positive CTP-based CIT finding and that police assess that a suspect engaged in a deep LOP (e.g., a suspect who is arrested several days after having grabbed a stolen item and extensively manipulated it), this technical information would be crucial. On the other hand, if the reverse is true that the CTP is only reliable when testing individuals who have deeply encoded information, this limitation would be valuable for police to know in their testing assessment. For example, a suspect who is arrested minutes after stealing an item and who barely had a chance to manipulate it may not be considered for testing.

With this in mind, given that the basic mock theft CTP scenario has never been replicated independently, the chief objectives of this research were to investigate different levels of processing on the memory encoding of participants in a mock theft scenario and the impact that varying LOP may have on the CTP's diagnostic performance in correctly classifying innocent and guilty participants for both word and image modality. Hence, regardless of their exposure to verbal or pictorial stimuli, we hypothesized that participants engaged in a deep processing task will likely generate significantly higher CIT effects than those involved in the shallow processing and the control group (H1), and those in the shallow processing task will expectedly display significantly higher CIT effect than the Control group (H2).

2. Experiment 1 – Word

2.1. Method

2.1.1. Participants

Following a power analysis with G*Power (Faul et al., 2007) for this current research, it was estimated that a sample group of about 42 (approx. 14 per group) was necessary to enable the detection of an effect size of f(V) = 0.25 or approximately 0.5 Cohen's *d*, at an alpha of.05. The parameters for the power analysis are included in the supplementary material.

A total of 46 (7 males) healthy participants were recruited for this study. The mean age was 21.8 (*SD* = 4.2), ranging from 18 to 41 years old. All were undergraduate students from Concordia University's psychology department and were offered a course credit for their participation. All had normal or corrected-to-normal vision and expressed fluency in English. None reported being color blind, nor suffering or diagnosed with a major psychiatric disorder such as schizophrenia spectrum and other psychotic disorders, bipolar and related disorders, or a medical condition (i.e., epilepsy). The data from four participants was excluded for making too many errors (these are described further below in the procedure section). They were excluded for exceeding a threshold of 20% behavioral errors (including one that made too many cognitive miscues as well as having a disabled right hand). A fifth participant was removed from analysis for technical reasons (electrode at A2 disconnected at approx. trial 190). This left 41 datasets for analysis.

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

2.1.2. Research design

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

2.1.3. Procedure

Volunteer participants were randomly assigned to one of three conditions, innocent control (Control) (n = 14), guilty immediate shallow processing (Shallow) (n = 13), or guilty immediate deep processing (Deep) (n = 14).

Once greeted by a research assistant, individuals in the Shallow group were handed out written instructions on how to perform a mock theft. The mock theft briefing sheet read as follows. "You are to walk over to room PY-051.00. This room is located straight down the hall from the lab. Once in the room locate a Green & Beige North Face backpack. Inside the backpack is an object. Steal the object from the backpack. Leave the backpack there. Hide the item on your person and return to the lab for further instructions." The item they were to 'steal' was a silver watch. They were invited immediately thereafter to enter the laboratory for testing. This shallow condition replicated in part the paradigms in Winograd and Rosenfeld (2011, 2014).

Persons assigned to the deep condition were given the same instructions on how to commit the mock theft and to return to the main laboratory room once this task was completed. Prior to the CTP testing, they were asked to read a short text made up of 12 sentences and write-in the missing words (e.g., Inside the backpack was a _____. The make of the _____ is Seiko. The back face of the _____ is Blue. The overall colour of the _____ and bracelet is Silver.). The expected word 'watch' was missing in 10 slots. The sentences and the missing information were

[4] Cut off points could be adjusted in real cases, downwards or upwards, depending on the intent of investigators to, for example, cast a wider net in the search of potential suspects (e.g., terrorist) or to pinpoint a guilty individual, respectively.

[5] Guilty informed participants were instructed to steal the item in question and did in fact take the item.

[6] Guilty naïve individuals were instructed to steal the item in question but did not actually take the item.

designed to force participants to handwrite in the word 'watch' in a set of meaningful sentences (Craik & Tulving, 1975) aligned with the presentation modality in Experiment 1. Repetition learning is a well-known memory enhancing strategy dating back to Ebbinghaus (1964) and many researchers have employed a variety of paradigms with words and images to illustrate that multiple exposures and contextual variability improve memory encoding and recognition (Hintzman & Stern, 1978; McCormick-Huhn et al., 2018; and Chen & Yang, 2020).

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

2.1.4. Trial structure and testing procedure

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

In the CTP, a trial consists in the presentation of one stimulus. A probe or an irrelevant item is always followed by a target or non-target item, and all presentations are separated by a fixation cross. (See Fig. 1). The single button press response from one mouse to the first stimulus is intended to confirm that the participant has implicitly seen the stimulus in question ("I saw it" response), while the conditional button presses from a second mouse in response to the second stimulus is meant to confirm the participant's explicit attention to the stimuli presentation (Rosenfeld et al., 2008).

In this experiment, investigators instructed participants to right-click the mouse on their left-hand side as fast as they could each time they saw a word. They were also told to either right-click or left-click the mouse on their right-hand side immediately when they saw a stimulus made up of a string of numbers. If it was the target item 11111, they had to right-click, and if it was one of any non-target items (i.e., 22222, 33333, 44444, or 55555), they had to left-click the mouse on their right-hand side. Participants who committed more than 20% of button press errors of omission or commission on either stimulus were excluded.[7] These miscues are called behavioral errors.

The CTP investigator usually pauses the experiment periodically to question the participant on the identity of the last stimulus seen. According to Rosenfeld et al. (2008), this step further ensures that the participant maintains his/her attention on the computer screen and the stimuli presentation. We performed these pop quizzes on our participants about every 38–50 trials, (M = 43.0), for a total of nine pauses over 374 trials. Participants were informed prior to testing that they would be questioned periodically on the last stimulus seen, and that more than two slip-ups would lead to their data being discarded. This type of mistake is called a cognitive error. Participants were not informed of a practice session, but investigators edited out the first 20 trials. To summarize, a total of 374 trials were presented but only 354 were kept for analysis. No real practice run exists in preparation for the application of the CTP. However, investigators normally use the first 20 trials to act as a built-in rehearsal.[8] These trials were edited out from the final analysis. The probe was presented 29 times (*8%*), irrelevants 158 times (*45%*), target 39 times (*11%*), and non-targets 128 times (*36%*).

2.1.5. Stimuli

In keeping with Winograd and Rosenfeld (2011), Lukács et al. (2016), and Funicelli et al. (2021), we used one probe and six irrelevant items in our experiment. The probe stimulus was the word "WATCH", and the irrelevants were the words: "CREDIT CARDS", "IPHONE", "SUNGLASSES", "USB KEY", "CAMERA", and "MONEY", in accord with Lykken's (1998) plausibility criterion that all irrelevant items used in a CIT are "equally plausible alternatives" (p. 39) to the probe item.

The stimuli presentation was done through PsyTask and displayed on a 55 cm HP Compact (LA2206x) flat monitor with a 1280 × 1024 resolution in a dimly lit room. The average stimuli size was 6.77 cm × 6.93 cm. At a viewing distance of 63.5 cm (measured from the participant's right eye to the center of the screen) the average stimulus subtended 5.5° x 6.1° of visual angle. The viewing distance from the participant's nasion to the fixation cross at the center of the monitor was 61 cm. All items were presented in Times New Roman, 96 font, Black on a White background surrounded by a wide Black edge.

2.1.6. EEG data acquisition

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

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

2.2. Analysis Methods

2.2.1. P300 amplitude and latency

The peak-to-peak (p-p) method of analyzing P300 amplitude was used as it has been found to be superior to the base-to-peak method (Soskins et al., 2001), and more sensitive in concealed information detection (Rosenfeld, 2011). All three sites were analyzed, but our final analysis rested on data from the Pz site. This site generally yields the largest amplitudes (Johnson, 1993), but Pz is also where P300 values are typically the largest in EEG-based CIT studies conducted with the CTP (Rosenfeld, Ward, Meijer, et al., 2017, Rosenfeld et al., 2018; Ward et al., 2020; Sui et al., 2020).

Grand averages (See Fig. 2) were calculated with all groups and conditions according to Keil et al. (2014) and two search windows were established. The first one was from 408 ms (T1) to 734 ms (T2), and the second was from 734 ms to 1300 ms (T3). Based on the grand mean of the probe curve, T1 (408 ms) was established as the point where the curve began its downward trajectory, T2 (734 ms) was determined to be the point where the curve intersected the X axis past the most downward point, and T3 (1300 ms) was selected arbitrarily as the point where the algorithm stopped searching for any waveform. We used a non-commercially available Matlab compatible software, supplied by Rosenfeld (personal communication, May 2015), to identify the most positive and negative peaks. The algorithm then searched for the mid-point of the most positive 100 ms average segment (also called the P300 latency) in the 408–734 ms look window, and then subtracted the average of the mid-point of the most negative 100 ms segment amplitude found between the 734–1300 ms window (see Rosenfeld et al.,

[7] This threshold is in keeping with the Rosenfeld laboratory.
[8] This preparatory stage is in keeping with the Rosenfeld laboratory.

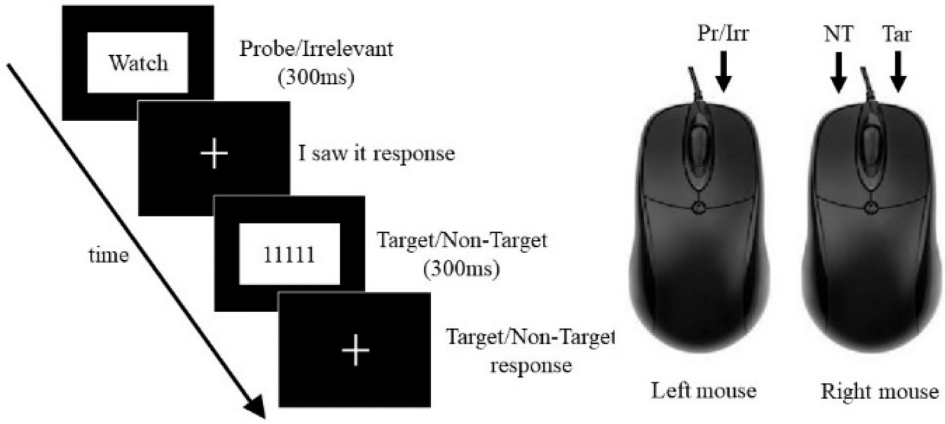

Fig. 1. Complex Trial Protocol. Complex Trial Protocol design. A Probe (Pr) or Irrelevant (Irr) stimulus was shown for 300 ms and the participant was instructed to respond as fast as possible by pressing the right button on the left mouse. This is the implicit "I saw it" response. This item is followed by the presentation of a Target (Tar) or Non-Target (NT) stimulus for 300 ms. The participant pressed the right button of the right mouse in the case of a Target or the left button of the same mouse for a Non-Target. This is the explicit attention-grabbing response.

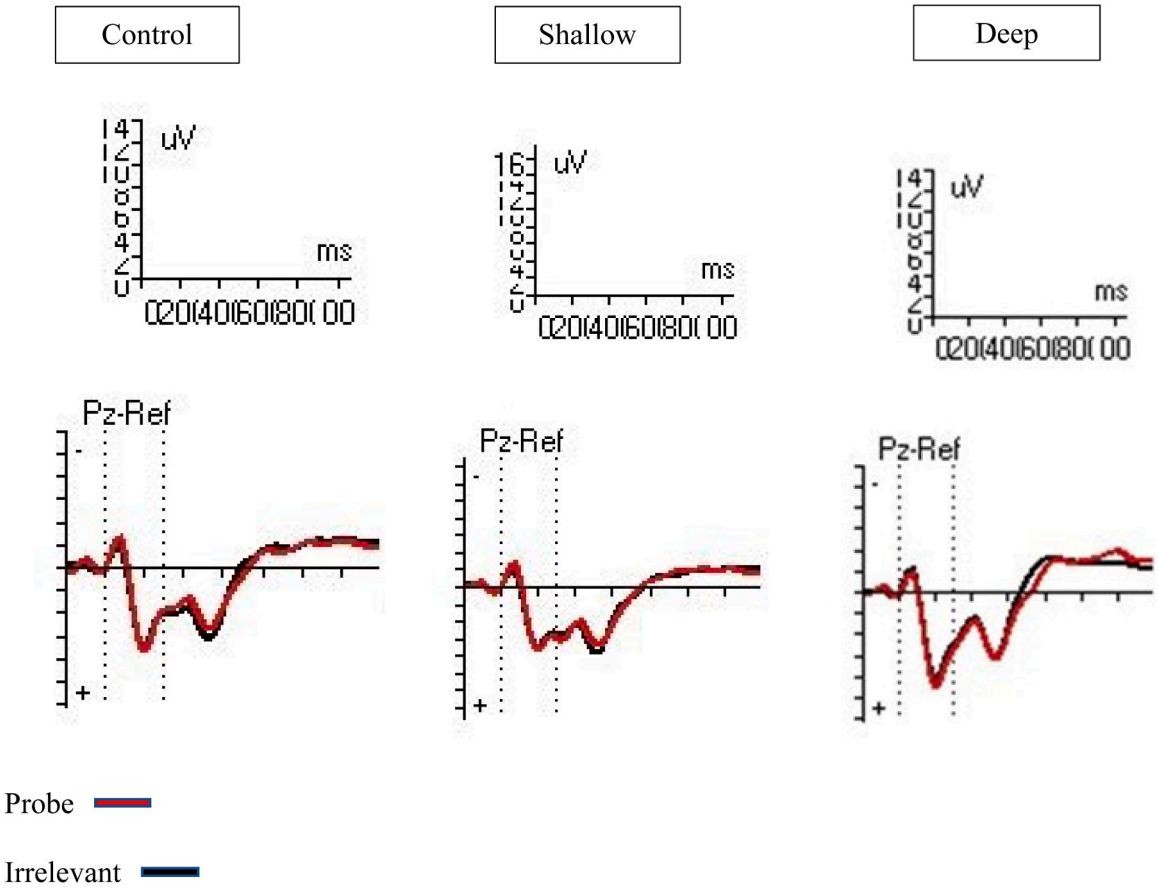

Fig. 2. Grand Averages – Word – Control, Shallow, and Deep Conditions at Pz.

2018, suppl. Mat. for a brief description of what the algorithm searches for). The subsequent value after subtraction was defined as the P300 p-p amplitude.

2.2.2. *Group statistical analyses*

We performed a repeated measures analysis of variance (ANOVA) for group analysis to identify the site where the largest amplitude was found between probe and irrelevants (Fz, Cz or Pz) and used a Bonferroni correction as appropriate for post-tests. From there on, we focused on the CIT effect, the difference between probe and irrelevant in Study 1 and 2. Estimates of effect size are reported as partial eta-squared and/or Cohen's (1969) *d*. We also report JZS Bayes factor values (BFs, scaled $r = 0.707$; Rouder et al., 2009); as obtained from the JASP Team, (2023). The BFs likelihood ratios are stated as supporting either the null hypothesis (no difference) when $BF \leq 1$, or the alternative hypothesis when $BFs \geq 1$ and should be read in accordance with the rough heuristic guide supplied in Schönbrodt and Wagenmakers (2018). For example, if we state that the $BF_{10} = 5$, it means that the data is approximately 5 times more likely under the alternative hypothesis than under the null hypothesis, and this is moderate evidence in favor of the alternative hypothesis (van Doorn et al., 2020).

2.2.3. *Individual diagnostics*

Bootstrapping is the statistical measure of choice for individual

diagnostics in EEG-based CIT research (Rosenfeld, 2011; Rosenfeld & Donchin, 2015). Instead of repeatedly submitting an individual to the same test, this technique permits the random resampling with replacement (n-1) of an EEG single sweep data distribution. In other words, bootstrapping resamples a single dataset to create many simulated samples. An average amplitude can then be calculated by bootstrapping a set of P300 probe waveforms for each participant. The same procedure is then applied to a corresponding set of waveforms for irrelevant items. The irrelevant P300 amplitude mean is subtracted from the probe mean and subjected to multitude iterations. We elected on 100 iterations as recommended by Rosenfeld, Ward, Meijer, et al. (2017) for both experiments.

We used here, and for the second experiment, three dependent variables for each electrode site. First was the p-p P300 amplitude in microvolts for probe (labelled as PrDx)[9] and irrelevant (labelled as IallDx)[10] items of participants in each condition. The second dependent measure, also called the CIT effect, was the difference in means between the iterated bootstrapped average of p-p P300s for probe and its equivalent for irrelevant items. The third was the number of iterations, out of 100, where probe p-p P300 bootstrapped iterations exceeded irrelevant p-p P300 bootstrapped iterations items (labelled as BSITERS) (Rosenfeld et al., 2018, suppl. Mat.; Davydova et al., 2020), at a confidence value over the 0.9 criterion. This criterion level has been found to be effective as a diagnostic cut-off at the individual level (Meixner & Rosenfeld, 2010). In sum, if at least 90 out of 100 bootstrapped probe p-p P300 iterations were greater than 0.9 of the irrelevants p-p P300 iterations, a participant was classified as *knowledgeable* (see Rosenfeld & Donchin, 2015 for a complete explanation on bootstrapping).

2.2.4. Receiver Operating Characteristic (ROC) Analysis

We used a ROC analysis to verify to a greater degree the CTP's effectiveness in discriminating knowledgeable participants from non-knowledgeable ones. A ROC curve is "the function that relates the proportion of correctly recognized target items (i.e., the *hit rate*) to the proportion of incorrectly recognized lure items (i.e., the *false rate*) across variations in response criterion (i.e., the propensity to make a positive recognition response" (Yonelinas & Parks, 2007, p. 800). The overall accuracy of a test is represented by the Area Under the Curve (AUC) from the resulting values of sensitivity (true positives/true positives + false negatives) and specificity (true negatives/true negatives + false positives) (Lalkhen & McCluskey, 2008). Classifiers that give a score of 1.00, shown as a curve closest to the top-left hand corner of the graph, indicates a perfect diagnostic performance. A result of 0.50, or the closer the curve comes to the 45-degree diagonal line, demonstrates that the test is accurate at chance level. For both experiments, we used MedCalc, (2023) to generate ROC curves.

2.3. Results

2.3.1. Between-Groups Comparisons

2.3.1.1. P300 p-p amplitudes. The data was verified for normality, skewness, and kurtosis. Apart from one outlier, all other participants were within + 3 / − 3 Z score. The original value for the amplitude level at FzIallDx[11] for participant 121 was 11.93 μV. The solution for dealing with the outlier was to seek the next highest valid value, in this case 8.18 μV, add 1.00 to it, and replace it with the new value of 9.18 μV.

The first step of our data analysis was to conduct a mixed repeated measure ANOVA with JASP (for Bayes factors; version 0.17.2.1) to determine the site that presented the best results. We found a main effect of sites (F (2, 76) = 83.19, $p < .001$, $\eta p^2 = .686$; $B_{10} = 1.05 \times 10^{+17}$). As expected, pair wise comparisons revealed that the Pz site produced mean amplitude values ($M = 10.36$ μV, $SE = 0.592$) significantly higher than Fz ($M = 4.61$ μV, $SE = 0.297$, $p < .001$; $BF_{10} = 7.83 \times 10^{+17}$) and Cz ($M = 8.32$ μV, $SE = 0.491$, $p < .001$; $BF_{10} = 1.49 \times 10^{+6}$). Accordingly, we continued our data analysis based on the Pz site for this experiment (study 1) as well as for study 2. We focused our analysis on the CIT effect, the difference in amplitude between probe-irrelevant stimuli.[12] A one-way ANOVA on the CIT effect revealed no effect of Group ($F(2,38) = 1.59$ $p = .22$;; $\eta p^2 = .08$; $BF_{10} = 0.52$). There was no difference between any of the groups (control-shallow t (25) = −1.45, p = .47, $BF_{10} = 0.73$; control-deep t (26) = −1.62 p = .34, $BF_{10} = 1.18$; shallow-deep t (25) = −0.14, $p > .05$, $BF_{10} = 0.36$). It should be noticed here that all the Bayes factors point to either an anecdotal support for the null hypothesis (BF between.3 and 1), or for the alternate hypothesis (BF between 1 and 3). We will address this point later in the discussion.

2.3.2. Individual Classification

2.3.2.1. ROC Curves. Using the bootstrapped scores from the BSITERS variable, none of the results produced significant findings in detection efficiency of Control versus Shallow group (AUC =.624, SE =.112, $p = .270$, 95% CI:.418–.801) or the Control versus the Deep condition group (AUC =.679, SE =.103, $p = .08$, 95% CI:.476–.841). There was no difference between the shallow and deep AUC curves ($z = -.52$ $p > .05$). Figs. 3 and 4 feature the ROC findings of the Control group compared to the Shallow group and that of the Control group relative to the Deep group respectively.

2.3.2.2. Bootstrapping. As expected from the results of the AUC curves, the bootstrapping analysis produced the following outcomes. We identified all our participants in the Control group (14/14, 100%), meaning that we had no false alarm (mistakenly identifying an innocent person as guilty). However, we only identified accurately 2/13 (15%) of our Shallow subjects and only 1/14 (7%) of our Deep individuals which translates into a false negative or miss rate (a guilty person wrongly classified as innocent) of 89%.

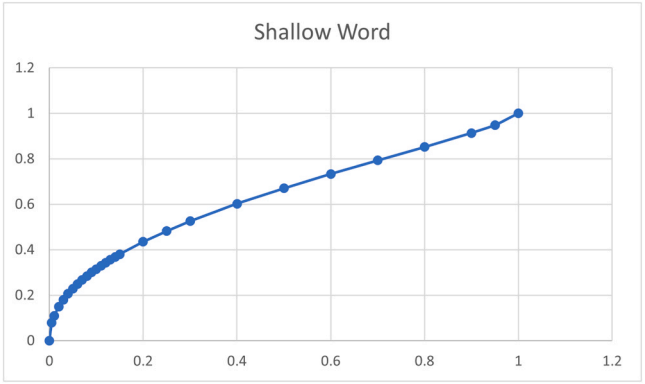

Fig. 3. ROC Curve Word – Control vs Shallow. The curved line is the distribution of the shallow condition sensitivity scores of the CTP compared to the Control group.

[9] PrDx represents the type of stimuli, in this case Probe, and Dx signifies the Peak-to-Peak amplitude difference.

[10] IallDx represents the type of stimuli, in this case Irrelevant all items, and Dx signifies the Peak-to-Peak amplitude difference.

[11] FzIallDx represents the location of electrode (frontal cortex at the hemispheric midline), the type of stimuli (Iall = Irrelevant all), and Dx signifies the Peak-to-Peak amplitude difference for the Iall items.

[12] A 2 (probe-irrelevant) x 3 (control-shallow-deep) repeated measure ANOVA was also conducted. The main results can be found in the Supplementary Material.

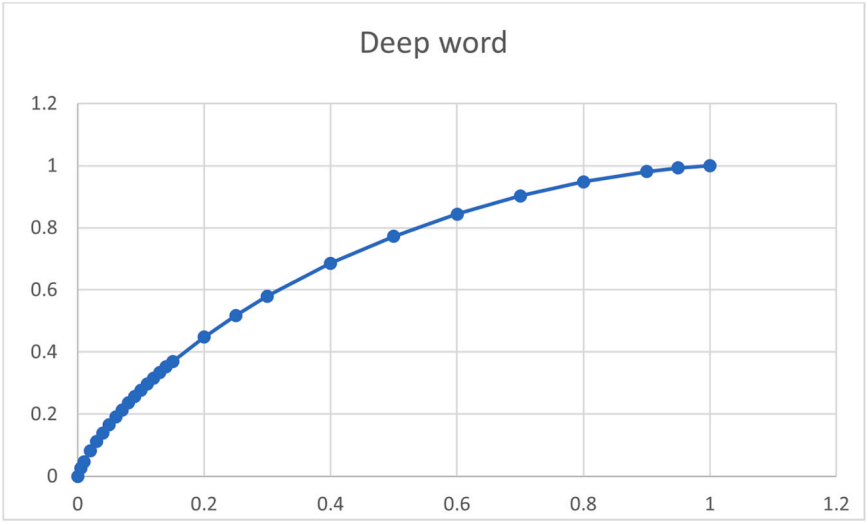

Fig. 4. ROC Curve Word – Control vs Deep. The curved line is the distribution of the deep condition sensitivity scores of the CTP compared to the Control group.

2.3.2.3. Post Experiment Questionnaire. Upon completion of the test, participants were asked to complete a post experiment questionnaire. Of relevance here was the question about whether they had recognised any objects from inside the backpack during the test. In the Shallow group, 12 out of 13 individuals reported having recognised the word 'watch' while 11 out of 13 persons in the Deep group made such a recognition with one participant who did not respond.

2.4. Discussion

We had expected the P300 amplitude levels of the Deep processing group to be significantly higher than those from the Shallow and Control group (H1) and the latter meaningfully higher than the innocent Control group (H2) as reflected in the CIT effect. The hypotheses were not confirmed. The Bayes factors may tell a different story. Their magnitude at best pointed to anecdotal evidence for the null or the alternate hypothesis. In other words, for the CIT effect, the current findings are inconclusive. For the LOP effect, the Bayes factor showed anecdotal evidence for the alternate hypothesis. It cannot be ruled out that LOP manipulation had no effect.

The AUC curves were very low compared to the mean AUC reported in previous experiments and this outcome was not expected. Not surprisingly, the classification rates mirrored these low AUCs values.

Our procedures mimicked Winograd and Rosenfeld (2011, 2014) except on one point. Their mock theft scenario consisted of asking their participants (in the guilty conditions) to bring a manila envelope to the office of the Psychology department and enquire with the secretaries as to the location of Dr. Rosenfeld's mailbox. Having located the mailbox, they were further instructed to look for a matching manila envelope labeled in Dr. Rosenfeld's name, to surreptitiously steal an item from inside that envelope, and to return to the lab with the stolen item. The participants were also informed of the secretaries' (also lab confederates) naïveté about the study and to do their best not to get caught (to increase participants arousal and realism during the performance of the theft). They were to have the secretaries contact the lab in the event they were discovered. In contrast to the above studies, but similar to Ward and Rosenfeld (2017), we asked our guilty participants to simply walk over to a nearby room outside the laboratory, locate a backpack once inside the room, find the only object inside the backpack, steal it, leave the backpack there, hide the object on their person, and to return to the lab for further instructions. Our scenario did not comprise this 'arousal' component so it may not have elicited much reaction among our guilty participants which could explain, at least in part, our results. However, Ward and Rosenfeld (2017) did not use this arousal component and still got successful hit rates of 87% and 93% in their suppression guilty group and simply guilty group respectively (see Table 1 for AUCs). We will explore this possible explanation and other issues in the general discussion.

3. Experiment 2 – Image

3.1. Method

Experiment 2 with pictorial stimuli was conducted in the same manner as experiment 1. We address below only the methodological sub-sections, or parts thereof, that differed from those applicable to Experiment 1.

3.1.1. Participants

A total of 51 (4 males) participants were recruited for this study. The mean age was 22.0 ($SD = 5.7$), ranging from 18 to 52 years old. The data from six participants was excluded for making too many errors either behavioral or cognitive, another was removed on suspicion of drug use, and a participant was not tested because of a hair extension that made it impossible to apply electrodes. This left 43 datasets for analysis.

3.1.2. Procedure

Upon reading and signing a written consent and completing a demographic data sheet, volunteer participants were then randomly assigned to one of three groups, innocent control ($n = 14$), guilty shallow processing condition (Shallow) ($n = 14$), or guilty deep processing condition (Deep) ($n = 15$).

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

3.1.3. Stimuli

The stimuli, probe and irrelevants alike, used in experiment 2 were the pictorial equivalents of the stimuli used in experiment 1 (Fig. 5) and presented in the same fashion as in Fig. 1.

3.1.4. Search windows

The two search windows for this study were established to be from 394 to 720 ms, and from 720 to 1300 ms based on the probe curve. The grand averages for each respective group are found at Fig. 6.

3.2. Results

3.2.1. Between-Groups Comparisons

3.2.1.1. P300 p-p amplitudes. The data was verified for normality, skewness, and kurtosis. Except for one outlier, all other participants were within + 3 / − 3 Z score. The original value for the amplitude level at FzPrDx for participant 10 was 14.35 μV. The solution for dealing with the outlier was to seek the next highest valid value, in this case 9.54 μV, add 1.00 to it, and replace it with the new value of 10.54 μV.

Given our initial finding in experiment 1 of a main effect of site, we focused our analysis of the CIT effect at the Pz site. A 1 (CIT score) x 3 (groups) ANOVA revealed a significant main effect (F (2, 40) = 6.92, p = .003, ηp^2 = .257; BF_{10} = 10.60) of the CIT effect. Relative to the Control group (M_{diff} = 0.208 μV, SE = 0.355), the CIT effect of both Shallow group (M_{diff} = 3.566 μV, SE = 1.167) and Deep group (M_{diff} = 4.708 μV, SE = 0.930) were significantly larger (t(26) = −2.753, p = 0.011, BF_{10} = 4.88) and (t(27) = −4.397, p < 0.001, BF_{10} = 145.14) respectively. The CIT effect between the two guilty groups did not differ (t (27) = −.91, p > .05, BF_{10} = .44).

3.2.2. Individual Classification

3.2.2.1. ROC Curves. As in the Word experiment the bootstrapped scores from the BSITERS variable were used to compute the ROC analysis. AUC findings showed that the CTP has good to very good detection efficiency for the control and shallow conditions *(AUC =.755, SE =.102, p < .05, 95% CI:.557–.897)* (Fig. 7) and for the Control and Deep groups *(AUC =.943, SE =.056, p < .001, 95% CI:.789–.995)* (Fig. 8) significantly above chance level. The difference between the AUC curves for the Shallow and Deep groups, relative to the Control group, were significant (z = −2.88, p = .003).

3.2.2.2. Bootstrapping. We were able to accurately classify 100% (14/14) of Control participants, 43% (6/14) of the Shallow ones, and 60% (9/15) of the Deep ones. As in Experiment 1, there were no false positives, but our miss rate (false negatives) was 48%.

3.2.2.3. Post Experiment Questionnaire. Post-experimentally all participants reported having recognized the watch during testing.

3.3. Discussion

Dealing specifically with the pictorial modality in experiment 2, H1 was partially supported while H2 was supported. Both guilty groups had higher CIT effects and AUC results than their innocent counterpart, and they significantly differed from each other, but only in terms of AUC values. Our findings pointed to a significant effect of LOP in relation to our Control group. As was the case in Experiment 1, the Bayes factor of the difference between the two Guilty groups pointed only to anecdotal evidence for the null hypothesis. The concerns expressed in the discussion section of experiment 1 are also applicable to experiment 2. We address the overall issues in the next section.

4. General Discussion

When we set up to evaluate the CTP performance in a mock theft scenario, we expected that the CTP would perform as in previously published studies with an AUC around .80 to .90. Our experiments are the first independent quasi replication of the CTP using such a paradigm. In experiment 1, the AUCs were not different from chance level and the classification rates were poor. Looking at the Bayes factors, however, the evidence in favor of the null hypothesis was merely anecdotal, and in the case of the Deep group in favor of the alternate hypothesis. As far as words being used as probes, the findings cast doubts as to the CTP's efficiency when non-autobiographical verbal stimuli are used, but we cannot conclude firmly that the method does not work. As for the level of processing, here the Bayes factor indicated that the results were slightly in favor of the level of processing having an effect. So again, we cannot conclude that manipulating levels of processing either did or did not

Fig. 5. Pictorial Stimuli.

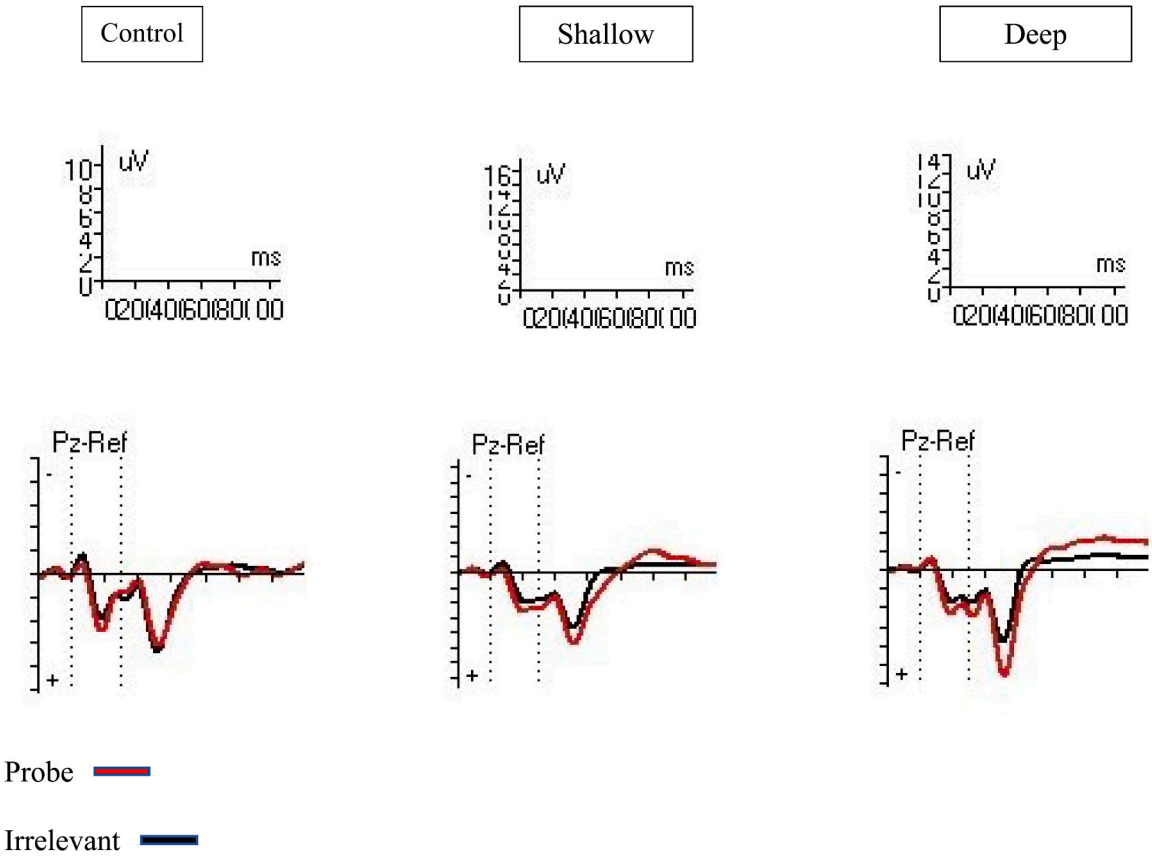

Fig. 6. Grand Averages – Picture – Control, Shallow, and Deep Conditions at Pz.

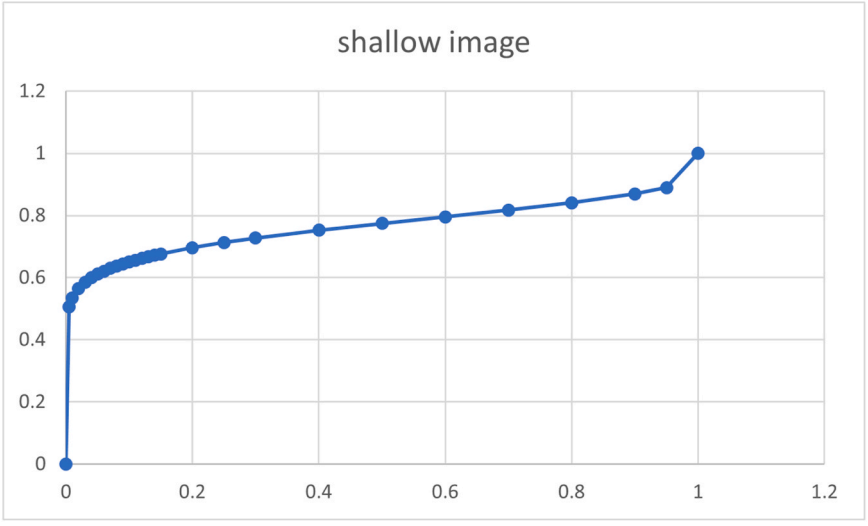

Fig. 7. ROC Curve Image – Control vs Shallow Conditions. The curved line is the distribution of the shallow condition sensitivity scores of the CTP compared to the Control group.

have an effect.

In Experiment 2, the CIT effect was clearly replicated. Both guilty groups showed higher CIT effect than the control group. Although the two guilty groups did not differ statistically, again the resulting Bayes factor did not favor the null or the alternate hypothesis. The AUC findings in the pictorial experiment proved to be more consistent. The Control/Shallow and Control/Deep ROC curves were significantly different, and statistically different from each other, but we obtained a 94% diagnostic rate in the deep condition and a respectable 76% rate in the shallow condition.

As mentioned earlier however, the AUC score does not always reflect on the individual classification score. Would the results have been different if we had used a .80 criterion for classification? Classification scores are dependent quite arbitrarily on the criterion chosen for classification. In the word condition, the classification would not have changed much. In the image shallow condition, however, eight participants would have been found guilty rather than six (57%); in the deep condition, 13 out of 15 (87%) would have been classified as guilty while

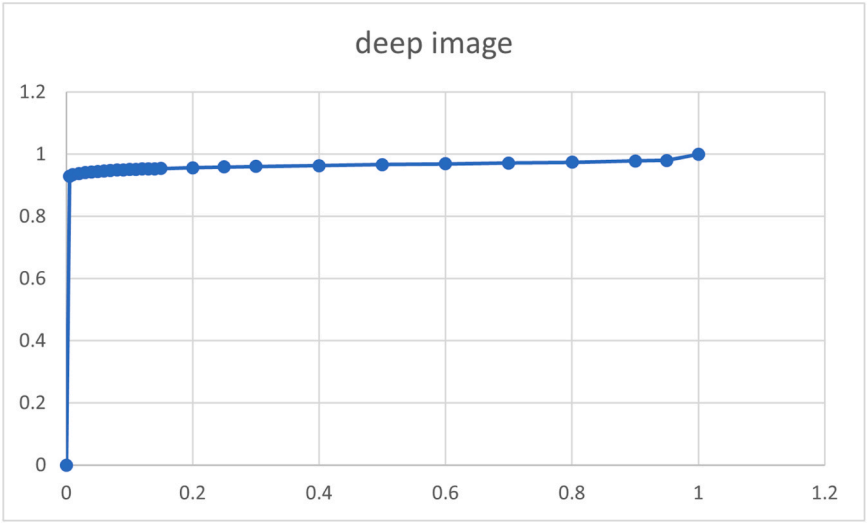

Fig. 8. ROC Curve Image – Control vs Deep Conditions. The curved line is the distribution of the deep condition sensitivity scores of the CTP compared to the Control group.

all controls would have remained correctly identified. Given the arbitrary nature of the cut-off points, the AUC is usually considered a better index of the goodness of a test than individual classification. What should be the ideal cut-off point is dependent on the question asked from the test like in any diagnostic test and the consequences of a positive or negative result. Additional independent experimentations are needed to evaluate the performance of the CTP in mock crimes scenarios, especially with verbal stimuli.

We can still look at the performance of the CTP with words or images as probes and evaluate the role of depth of processing. In the shallow conditions, there were no statistical difference in the proportion of guilty participants correctly identified (images .43 vs words .15) ($z = -1.5938$ $p > .05$, $BF_{10} = .53$). Here, it seems that the use of words or images did not have any influence on classification rate, but the Bayes factor did not confirm the null findings. These two groups correspond best to the methodology usually utilised by Rosenfeld and colleagues. Despite the overall poor classification rates, the individual detection rates were independently better in the deep pictorial modality 60% (9/15) than in the deep word condition 7% (1/14 ($z = 3.002$, $p < .01$. $BF_{10} = 21.86$). Given the results in the shallow condition, one may suspect that the LOP manipulation may work better with images as stimuli. However, we need to caution here that the research design in both experiments does not permit a word vs image comparison as the LOP procedure was slightly different for both groups.

The low detection rates in our Guilty groups, both Shallow and Deep, in the word condition, require a closer examination. The experimental condition described above in Winograd and Rosenfeld (2011) resembled our shallow processing condition. We expected our deep processing condition including a missing word text exercise where participants were asked to fill 10 blank spots spread across 12 sentences with the word 'WATCH', to generate a quasi-priming effect. The word 'stole' in one of the sentences was the only direct association of criminality to the experimental task. The other sentences were either descriptive in nature (i.e., The _____ is a man's _____. The make of the _____ is Seiko. The back face of the _____ is Blue) or meant to enhance its semantic significance (i.e., A _____ is a mechanical instrument designed to tell time. This particular _____ indicates the time as well as the date and day of the week). The results show that in this type of scenario and procedures, deep processing did not improve hit rates significantly. In the image condition, a similar exercise, each adapted to their respective modality presentation, did not increase the CIT effect to differentiate the shallow from the deep condition, but it did increase the AUC of the Deep group significantly when compared to the AUC of the Shallow group as well as improve classification rates. So, the use of an image and deeper processing seems to lead to an improved performance of the CTP.

It appears reasonable to assume that an increased level of arousal during the commission of the 'theft' could have generated higher P300s. Greater amygdala activation has been found to correlate positively with memory performance (Canli et al., 2000). The minimal level of arousal experienced by our guilty subjects during the mock crime could have had the cascaded effect of poor encoding. As in a domino effect, insufficient encoding, especially for our Shallow processing group, may have decreased recollective retrieval which may account for the lackluster detection efficiency. Price et al. (2009) found some support for increased remembrances with increased arousal. Peth et al. (2012) manipulated the level of stress during mock-crime execution and concluded that "emotional arousal might facilitate the detection of concealed information sometime after the crime" (p. 381). This line of inquiry is supported by evidence from Kennedy et al. (2014) where levels of emotionality have a bearing on P300 amplitudes to the extent that pictures produce higher magnitudes whether the visual stimulus is positive or negative. Furthermore, Klein Selle et al. (2017) suggested that emotional arousal may enhance detection efficiency with the SCR measure. As noted before, however, some of the published studies by Rosenfeld and colleagues did not have an arousal component and still got acceptable classification rates and AUCs.

Is the meaningfulness or salience of our probe stimulus for our guilty participants of concern here? Put differently, the question to be answered is whether a minimally meaningful stimulus detectable by the CTP. The literature underpinning the orienting reflex, to explain an elevated response when one is exposed to a novel stimulus that carries a special significance, is robust (Verschuere & Ben-Shakhar, 2011; Klein Selle et al., 2018). Our post-experimental inquiry indicates that 100% of the participants in the Exp 2 stated that they recognised the image of the watch during testing, while nearly all participants in Exp 1 (92% and 85% in the shallow and deep conditions respectively) responded similarly in relation to the word "watch". Yet, our findings are at odds with those post test results. Neither salience nor meaningfulness was independently manipulated here, leaving this possibility as an outstanding issue for further research or raising the likelihood of a limitation about Rosenfeld's CTP.

Modality congruency may have played a role in our results. Van der Cruyssen et al. (2021) found that detection performance improved when items were encoded and tested in the same modality. In our case guilty participants were asked to steal a watch in both experiments but were tested in a congruent modality in the picture experiment and tested in an

incongruent modality in the verbal experiment.

We should add here another potential limitation of the current experiments. It is possible that the number of participants was sufficient to find a strong CIT effect, but not sufficient to find a more subtle effect like the level of processing. This is exemplified by a number of non-significant results that were not clearly supported by the Bayes factors. In most of these then the verdict is out for the moment.

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

As it stands, the jury, on the efficacy of the CTP to reliably distinguished between innocent and guilty participants when non-autobiographical stimuli (in the present case, words) are used, is still out. In the case of images, however, the AUCs demonstrated that the test has good discriminability and as is the case of any test, the use of different cut-off points influences the classification rate. Manipulating arousal, salience, depth of processing, stimuli significance, the use of images, and testing modality may thus be ways to improve the CTP performance and these variables should be explored further. Future researchers should be mindful of these as they attempt to conduct investigations under more ecologically valid conditions either in laboratory or in controlled field-like settings.

Declaration of Competing Interest

This research did not involve any grant. There are no conflicts of interest to declare.

Data Availability

data avail through osf.io see hyperlink in manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.biopsycho.2023.108666.

References

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

Chen, H., & Yang, J. (2020). Multiple exposures enhance both item memory and contextual memory over time. *Frontiers in Psychology, 11*(565169), 1–14.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

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

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

Davydova, E., Rosenfeld, J. P., & Labkovsky, E. (2020). Necessity of the target discrimination in the P300–based complex trial protocol test for concealed information. *Psychophysiology, 57*(5), Article e13548.

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

Ebbinghaus, H. (1964). Memory (translated by HA Ruger and CE Bussenius) Dover. *New York (original work published 1885)*.

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

Funicelli, M., White, L., Ungureanu, S., & Laurence, J. R. (2021). An independent validation of the EEG-based complex trial protocol with autobiographical data and corroboration of its resistance to a cognitively charged countermeasure. *Applied Psychophysiology and Biofeedback*, 1–13.

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

Hintzman, D. L., & Stern, L. D. (1978). Contextual variability and memory for frequency. *Journal of Experimental psychology: Human Learning and Memory, 4*(5), 539.

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

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.

JASP Team (2023). JASP (Version 0.17.2.1) [Computer software]. Retrieved from https://jasp-stats.org/://jasp-stats.org/.

Johnson, R. A. Y., Jr (1993). On the neural generators of the P300 component of the event-related potential. *Psychophysiology, 30*(1), 90–97.

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

Kennedy, L., Dorrance, S., Stoneham, T., Bryant, M., Boyd, K., Flippen, K., & Nichols, D. F. (2014). Event-related potentials in humans for emotional words versus pictures. *Impulse: The Premier Undergraduate Neuroscience Journal*, 1–16.

Kim, S. C., Kim, H., Lee, K. E., Song, I., Chang, E. H., Kim, S., & Kim, H. T. (2022). Retroactive memory interference reduces false positive outcomes of informed innocents in the P300-based concealed information test. *International Journal of Psychophysiology, 173*, 9–19.

Klein Selle, N., Verschuere, B., Kindt, M., Meijer, E., Nahari, T., & Ben-Shakhar, G. (2017). Memory detection: The effects of emotional stimuli. *Biological Psychology, 129*, 25–35.

Klein Selle, N., Verschuere, B., & Ben-Shakhar, G. (2018). Concealed information test: Theoretical background. *Detecting concealed information and deception* (pp. 35–57). Academic Press,.

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

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia Critical care & pain, 8*(6), 221–223.

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

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

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

MedCalc® Statistical Software version 22.009 (MedCalc Software Ltd, Ostend, Belgium; https://www.medcalc.org; 2023).

McCormick-Huhn, J. M., Bowman, C. R., & Dennis, N. A. (2018). Repeated study of items with and without repeated context: aging effects on memory discriminability. *Memory, 26*(5), 603–609.

Meijer, E. H., Selle, N. K., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology, 51*(9), 879–904.

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

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

Olson, J. M., Rosenfeld, J. P., Ward, A. C., Sitar, E. J., Gandhi, A., Hernandez, J., & Fanesi, B. (2022). The effects of practicing a novel countermeasure on both the semantic and episodic memory-based complex trial protocols. *International Journal of Psychophysiology, 173*, 82–92.

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

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Price, H. L., Lee, Z., & Read, J. D. (2009). Memory for committing a crime: Effects of arousal, proximity, and gender. *The American Journal of Psychology, 122*(1), 75–88.

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

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

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

Rosenfeld, J. P., Hu, X., Labkovsky, E., Meixner, J., & Winograd, M. R. (2013). Review of recent studies and issues regarding the P300-based complex trial protocol for detection of concealed information. *International Journal of Psychophysiology, 90*(2), 118–134.

Rosenfeld, J. P., Labkovsky, E., Davydova, E., Ward, A., & Rosenfeld, L. (2017). Financial incentive does not affect P300 (in response to certain episodic and semantic probe stimuli) in the complex trial protocol (CTP) version of the concealed information test (CIT) in detection of malingering. *Psychophysiology, 54*(5), 764–772. https://doi.org/10.1111/psyp.12835

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

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

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

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

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Sai, L., Lin, X., Rosenfeld, J. P., Sang, B., Hu, X., & Fu, G. (2016). Novel, ERP-based, concealed information detection: Combining recognition-based and feedback-evoked ERPs. *Biological Psychology, 114*, 13–22.

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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

Seymour, T. L., & Fraynt, B. R. (2009). Time and encoding effects in the concealed knowledge test. *Applied Psychophysiology and Biofeedback, 34*, 177–187.

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

Sui, T., Sitar, E., Rosenfeld, J. P., Labkovsky, E., Ward, A., & Davydova, E. (2020). The enhancing effect of incongruent verbal priming stimuli on the CIT effect with pictorial probes in the P300-based complex trial protocol. *International Journal of Psychophysiology, 148*, 59–66.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., & Wagenmakers, E. J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 1–14.

Van der Cruyssen, I., Regnath, F., Ben-Shakhar, G., Pertzov, Y., & Verschuere, B. (2021). Is a picture worth a thousand words? Congruency between encoding and testing improves detection of concealed memories. *Journal of Applied Research in Memory and Cognition, 10*(4), 667–676.

Verschuere, B., & Ben-Shakhar, G. (2011). Theory of the Concealed Information Test. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the Concealed Information Test*. Cambridge University Press.

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

Ward, A. C., Rosenfeld, J. P., Sitar, E. J., & Wasserman, J. D. (2020). The effect of retroactive memory interference on the P300-based Complex Trial Protocol (CTP). *International Journal of Psychophysiology, 147*, 213–223.

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

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

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