International Journal of Psychophysiology 90 (2013) 118–134

Contents lists available at ScienceDirect

International Journal of Psychophysiology

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

Invited Review

Review of recent studies and issues regarding the P300-based complex trial protocol for detection of concealed information

J. Peter Rosenfeld *, Xiaoqing Hu, Elena Labkovsky, John Meixner, Michael R. Winograd

Department of Psychology, Northwestern University, United States

ARTICLE INFO

Article history:
Received 21 January 2013
Received in revised form 23 August 2013
Accepted 28 August 2013
Available online 4 September 2013

Keywords:
P300
Complex trial protocol
Concealed Information Test
Memory detection
Deception
Event-related potentials

ABSTRACT

In this review, the evolution of new P300-based protocols for detection of concealed information is summarized. The P300-based complex trial protocol (CTP) is described as one such countermeasure (CM)-resistant protocol. Recent lapses in diagnostic accuracy (from 90% to 75%) with CTPs applied to mock crime protocols are summarized, as well as recent enhancements to the CTP which have restored accuracy. These enhancements include 1) use of performance feedback during testing, 2) use of other ERP components such as N200 in diagnosis, 3) use of auxiliary tests, including the autobiographical implicit association test, as leading to restored diagnostic accuracy, and 4) a study of the mechanisms underlying CMs. A novel, doubly efficient version of the CTP involving presentation of two probes in one trial is described as a new way to improve accuracy to levels above 90% in mock crime situations. Finally, a thorough analysis of the legal issues surrounding use of the CTP in U.S. is given.

1. Introduction and background

A novel, reliable, and valid system to aid in detecting deception would make a significant contribution to the forensic sciences that serve the criminal justice system. We do not here suggest a variation of the controversial "lie detector" that one sees in popular media. This is the familiar polygraph, which is based on a protocol called the comparison question test (CQT, formerly, the control question test; Reid and Inbau, 1977) in widespread investigative use in the U.S. It is favored by many law enforcement and security agencies because of its relative ease of use and propensity to elicit confessions (including many that are later shown to be false; Furedy and Liss, 1986; Kassin, 2008; Warden, 2012). The CQT is rarely used in U.S. courts since deception researchers and scientists have seriously questioned its validity and reliability for a long time (e.g., see the report of the National Research Council of the National Academy of Sciences; National Research Council, 2003). Also, it tends to elicit an unacceptably high number of false positive outcomes (innocents wrongly diagnosed as guilty). That the CQT lacks general acceptance from the scientific community means it may fail to satisfy the Daubert standard (Gallai, 1999; Saxe and Ben-Shakhar, 1999) for admissibility of expert testimony by all federal courts and a majority of state courts, discussed below (Daubert v. Merrell Dow Pharmaceuticals, 1993). It will likely also fail to satisfy the Frye test, used in a minority of state courts (Frye v. United States, 1923).

Another protocol that may be used to help establish the truth or falsehood of legal testimony is the "Guilty Knowledge Test" (known more recently as the Concealed Information Test or CIT; Lykken, 1959, 1998; Verschuere et al., 2011). This protocol actually does not claim or aim to detect lies; it is instead aimed at detecting whether or not a suspect recognizes information ordinarily known only by guilty perpetrators and, of course, enforcement authorities. For example, in a murder case, the CIT enquires about whether or not a suspect recognizes the murder weapon actually used (details of the CIT below). The point of view in this review is that deception may or may not be inferred only by triers of fact (i.e., juries or judges) if a given suspect is shown, via a CIT, to know something that should be known only by guilty parties. The suspect may have a credible, (dis-)provable explanation for his knowledge (such as press leakage). We think that the veracity of this type of explanation should also be determined by a judge or jury. The CIT is much preferred by the academic deception research community (National Research Council, 2003; Iacono, 2011; Patrick, 2011) and is in regular field and court use in Japan (Osugi, 2011). Because, as will be later explained, the CIT likely does not yet satisfy the Daubert (1993) criteria (Ben-Shakhar and Kremnitzer, 2011), it too is not ready for field use in the U.S., and there are other objections to its use. These, however, are mostly "cultural", meaning that the members of the practicing polygraph community simply do not like giving up the CQT which they are used to and which they were taught in polygraph schools that mostly eschew the CIT, and thus they prefer the CQT (Kraphol, 2011). There are also putative objections concerned with the difficulty of composing a good CIT (Kraphol, 2011; Podlesny, 1993), despite the fact that such problems have been solved in Japan where the CIT is in regular use, including court use (Osugi, 2011).

* Corresponding author at: 2029 Sheridan Rd., Evanston, IL 60208, United States. Tel.: +1 847 491 3629.
E-mail address: jp-rosenfeld@northwestern.edu (J.P. Rosenfeld).

0167-8760/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijpsycho.2013.08.012

Indeed, the scientific research community and now even some in the legal community (Meixner, 2012; Rosenfeld and Greely, 2012) are becoming increasingly persuaded that such objections to a form of the CIT based on the event-related EEG potential (ERP) component, P300, the subject of this review, can be overcome. We suggest here that after the P300 CIT procedures, which have been consistently improved in laboratory conditions over the last decade, are validated in a field population, all the Daubert criteria for admission of this brain-wave-based CIT to U.S. courts will likely have been satisfied (Ben-Shakhar and Kremnitzer, 2011; Meixner, 2012).

In any CIT, many multiple choice questions are presented to the subject about crime details (e.g., "Which was the murder weapon, was it (a) the hunting knife? (b) the baseball bat? (c) the rifle? (d) the pistol? (e) the garrote?" Physiological responses are recorded and acquired as the questions are asked. For each question, there is only one correct crime-relevant answer. When this item — called the probe — is presented to the guilty subject (who outwardly/verbally denies recognizing the probe), the largest physiological response is expected in comparison to responses to the other incorrect answer choices (called irrelevants). For each question, if there are five choices, as above, the probability that a non-knowledgeable (i.e., innocent) subject will respond by chance with the largest physiological response to the probe item (a false positive outcome) is 1/5 = .2. If one can come up with as many as four independent questions about four independent items of information, the chance hit probability for all four is reduced to .2 to the 4th power = .0016. Researchers appreciate this feature of the CIT: the fact that by adding independent questions, the false positive probability can be reduced to whatever specifiable low value is required by a given agency or institution, thereby offering good protection to innocent suspects from mistaken decisions regarding their knowledge of crime-relevant information.

We note that the false probability value, used merely for illustration here, assumes that on each question, the probe elicits the largest response. In our own and others' work discussed below, we will note that for each question, we usually compare the probe with the average of all irrelevants. This will lead to differing values for false positive probabilities than those just presented here, but the general principle holds: the greater the number of independent items, the greater the protection against false positive diagnoses. Indeed, the implication of the above computation is that our and others' criterion for a knowledgeable decision is a hit on all test items. In fact it is not, for such a stringent criterion would also lead to a high miss or false negative rate. In our studies with multiple items, we would require that only about 67% of the total items used result in hits in order for a knowledgeable decision to be made. In this situation, computation of the false positive probability (fp) requires use of the binomial distribution (see http://vassarstats.net/index.html) which is beyond the scope of this review. However, for purposes of example, we note that fp still decreases systematically as items are added to a test as follows: Holding the knowledgeable (guilty) decision requirement that 67% of the items must be hits for these varying numbers of items: 3, 6, 8, 12 (with p[random hit] on any item = 1/5), the respective fp's are .1, .017, .001, and .0006, based on the binomial distribution.

It is also possible to demonstrate with the binomial distribution that sensitivity and specificity also increase as the number of items increases: If we assume the probability that a knowledgeable individual will show maximal response to the critical probe on one item is 0.75 (it is unimportant if this is a precise assumption as it is used for the sake of demonstration and any other number will work similarly), then with 3 items sensitivity is 0.70 and specificity is 0.90 but with 6 items, these two values increase to 0.83 and 0.983, respectively. (Gershon Ben Shakhar reminded us in private communication of this last fact.)

The physiological responses traditionally recorded by both CQT and CIT involve responses of the autonomic nervous system (ANS) such as heart rate, blood pressure, and sweat gland activity (indexed by the skin conductance response or SCR). Among the problems with both the ANS-based CQT and CIT raised by the report of the National Research Council of the National Academy of Sciences (National Research Council, 2003) is the potential susceptibility of all ANS-based methods to countermeasures (CMs). As stated by Honts et al. (1996, p. 84), "Countermeasures are anything that an individual might do in an effort to defeat or distort a polygraph test." The National Research Council report went on to state that "Countermeasures pose a serious threat to the performance of polygraph testing because all the physiological indicators measured by the polygraph can be altered by conscious efforts through cognitive or physical means" (National Research Council, 2003, p. 4). More specifically, CMs are effective against both the autonomic/polygraphic CQT (Honts et al., 2001), as well as against the autonomic/polygraphic CIT (Ben-Shakhar and Dolev, 1996; Elaad and Ben-Shakhar, 1991; Honts et al., 1996).

Deception researchers all hoped and indeed expected that when the P300 Event-Related EEG Potential was introduced as the dependent index of recognition in a CIT (Farwell and Donchin, 1991; Rosenfeld et al., 1991, 1988), the CM issue would be resolved. For example, the eminent inventor of the GKT/CIT (Lykken, 1998, p. 293), suggested about CMs to P300 CITs: "Because such potentials are derived from brain signals that occur only a few hundred ms after the GKT alternatives are presented… it is unlikely that countermeasures could be used successfully to defeat a GKT derived from the recording of cerebral signals." (Ben-Shakhar and Elaad, 2002, expressed a similar view.) All this optimism, as shown below, turned out to be misplaced.

To appreciate this point, one recalls that an event-related potential (ERP) is a series of peaks and troughs in the EEG that are elicited by a discrete stimulus or event. The eliciting event for the P300 ERP component can be any rarely (e.g., probability = p = .1) presented stimulus having special salient meaning for the subject. In CIT applications, the event is typically a meaningful word or picture (a probe) presented rarely among a series of other, frequently occurring, non-meaningful stimuli (irrelevants) from the same item category as the probe. For example, as suggested above, the name (or picture) of an actual murder weapon (e.g., a knife) used in a crime can be presented, as a probe, to a suspect in a series of other possible (but crime-irrelevant) weapons (e.g., pistol, club, tire iron, rifle, rope, ax) in about 10% of the total stimulus presentations. A guilty subject — but not an innocent subject — should recognize only the actual murder weapon, the knife, and his brain will respond by showing the P300 sign of recognition in the knife-evoked, probe wave, but not in the irrelevant waves.

The ongoing scalp-recorded EEG is noisy, and since P300 must ride on it, it is often hard to see in single trial samples. One therefore averages ERP responses to about 30 single presentations of each probe and of each irrelevant (all time-locked to the stimulus event onset), and then uses a statistical procedure to compare averaged probe and irrelevant P300s. The bootstrap method (Efron, 1979) that we use (detailed in Rosenfeld, 2011) gives the confidence (from 0 to 1.0) one has that in a given subject, the average probe P300 is larger than the average irrelevant P300. We have typically required that there must be at least .9 (90%) level of statistical confidence that the average probe P300 is greater than the average of all irrelevant P300s before concluding that a subject recognizes concealed information germane to a crime. The criterion (.9) could be reduced as required by situation specifics, provided an acceptably low false positive rate is obtained.

On the criterion issue, we note further that until a finalized and definitive P300-based test is developed and parametrically optimized for maximum discrimination efficiency and accuracy, as confirmed in representative populations, one cannot arbitrarily set a bootstrap diagnostic criterion at some level for use in all future studies. This is particularly so when an ongoing research program changes one or more experimental parameters and/or specific dependent variables from study to study — which may indeed produce response distributions of differing and asymmetric shapes in these various studies — in the overall aim of sensitivity/specificity maximization (or optimization. We define maximal sensitivity/specificity to be when [correct

detections + correct rejections] / Number of subjects, total, is maximized]. The strict assumptions of signal detection theory (SDT; Green and Swets, 1966; increasingly utilized in diagnostic studies and introduced in a major contribution by Ben-Shakhar and colleagues for deception studies; see Ben Shakhar et al., 1982) often assume that the key dependent variables (probe minus irrelevant P300 amplitude differences (as reflected in percent significant bootstrap iterations in our case)) are normally distributed in both knowledgeable as well as unknowledgeable subjects, with equivalent variance in both subject distributions. However in the real P300 samples we have encountered in several experiments, this assumption is not justified. Usually a non-knowledgeable subject distribution will have higher variance and longer tails than the knowledgeable distribution, in which case, a lowering of a cutoff will greatly improve sensitivity with not much change in specificity (see footnote 2.) Thus, as stated above, a criterion may be reduced as required by situation specifics, provided an institutionally acceptable low false positive rate is maintained, i.e., not much changed.

We hasten to add a self-evident caveat that a given cutoff adjustment is an arbitrary selection of one cut point along the Receiver Operating Curve or ROC (from SDT) so as to illustrate how well a given diagnostic can do in a particular data set collected under specific conditions. (Accuracy, in terms of error rates, is a critical pillar of the Daubert standard for legal admissibility discussed above, as well as in the concluding section below on legal issues.) Clearly, choices of other cut points along the ROC will yield differing sensitivity/specificity values (see Table 5b). In this review (e.g., Table 6) we provide AUCs when possible (ROC figures for each study would take up too much space), accompanied by Grier (1971) A′ values and sensitivity/specificity fractions at specified cutoff values. It is preferable that a criterion-independent index of test discriminability such as the area under a ROC curve or AUC, or better still, the entire ROC plot) be presented in each study. It is probably best to present complete ROC plots with a few illustrative cutoff points extrapolated to the sensitivity and specificity axes. We note that AUC may be computed regardless of the shape of dependent variable distributions. However, simple one-dimensional numerical AUC comparisons may be misleading when ROCs from differing studies have different skews; (as they typically do in our situation; see footnote 2 and van Erkel and Pattynama, 1998). We finally note here that inspection of the ROC shape for a given method and dependent variable set is a good guide to criterion selection (see footnote 2) in conjunction with institutional needs regarding costs of false positives and misses, and of a priori probabilities.

The earliest P300-based CITs (Allen et al., 1992; Farwell and Donchin, 1991; Rosenfeld et al., 1991, 1988) were called "3-stimulus protocols" because in them one presents on every trial either (1) a rare ($p = .1$) probe, (2) a frequent ($p = .8$) irrelevant or (3) a rare ($p = .1$) target stimulus. The target is simply another irrelevant item, but one to which the subject is assigned to make a unique button response different than the same single button pressed either to probe or to other irrelevant items. The idea is to force the subject's attention to whichever unpredictable stimulus is randomly presented on each trial. (Some researchers, e.g., Farwell and Donchin, 1991, use target-evoked P300s also in their diagnostic analyses, although we have criticized this practice in Rosenfeld, 2011. Examples of the P300s evoked by probes in the 3-stimulus protocol are seen in Fig. 4 below, in the SG and CM panels, "Part-2".)

A very serious problem with all such 3-stimulus P300 protocols, despite their initial promise and despite initial high expectations for them, has been their vulnerability to CMs, responses that a subject makes to distort the results of a deception test (Honts et al., 1996). In making the first demonstration of CM effects in P300-based CITs, Rosenfeld et al. (2004) anticipated the ideal CM: Secret conversion by the subject of irrelevant items to P300-generating, covert target items requiring specific covert responses, behavioral or mental. We confidently expected that this strategy would be effective — which it was — because in the ordinary, un-countered 3-stimulus protocol, the subject is instructed to make unique responses to explicitly assigned targets. These are readily executed with the typical result that large target P300s are evoked since these targets are also rare and additionally, meaningful, due to their unique button requirement. (Rareness and meaningfulness are the major antecedents for P300; Johnson, 1993.) We reasoned that if the subject can follow an experimenter's instruction to respond uniquely to an experimenter-chosen irrelevant (an explicit target) then the subject could also covertly define some (or all) irrelevants for himself as implicit targets to which he could make unique responses. These originally irrelevant but now secret targets would also elicit large P300s so that one could no longer depend on the probe P300 amplitude to reliably exceed that of the irrelevant P300. The larger probe P300 is, of course, what ordinarily makes the diagnosis of possession of concealed information.

To explain the effectiveness of CMs in the three stimulus protocol as used by Rosenfeld et al. (2004), Rosenfeld et al. (2008) reasoned that in the 3-stimulus protocol, target and probe stimuli are competing for attention resources, a situation that tends to reduce P300 amplitude (Donchin et al., 1986) and thus weaken the sensitivity of the 3-stimulus protocol. That is, in each trial of the 3-stimulus protocol, a dual task protocol is in effect in that subjects must be prepared to do an explicit target/non-target discrimination task simultaneously with doing an implicit probe recognition task, since on each trial, either a target, an irrelevant or a probe could appear in the same trial-launching time position. Rosenfeld et al. (2008) developed a novel P300 protocol called the Complex Trial Protocol (CTP; pictured below in Fig. 1) which temporally separated the presentation of probe or irrelevant from the presentation of target or non-target.

The trial begins with presentation of either a probe or irrelevant item (Stimulus 1 or S1 for first stimulus), the probe birth date "Aug 4" shown here, to be immediately followed by a button press (R1, the "I saw it" response), signaling the operator that the stimulus, whether probe or irrelevant, was perceived. Then, after a random delay of about 1200 to 1800 ms, the second stimulus (Stimulus 2 or S2), either a target or non-target is presented and the subject makes a second button press (the T/NT response or R2) on either a target or non-target button. (The target shown here is the number string "11111".) As in the 3-stimulus protocol, this S2/R2 sequence is used to maintain attention throughout. However, the main method of forcing attention to S1 on every trial is to warn the subject prior to the test block that, unpredictably, the protocol will be paused every few minutes and the experimenter will ask the subject to recall the just presented S1. More than one error on these five or six unpredictable test trials will result in a report of non-cooperation, a form of test failure in the field. (In the lab, a bonus for error-free runs could be denied as a punishment for such non-cooperation errors.) The target used in our recent complex trial protocol studies is the number string, "11111" (as above) among strings of other, non-target numbers, "22222", "33333" and so on. (Examples of the P300s evoked by probes in the CTP are seen in Fig. 4 below, in the SG and CM panels, "Parts-1".)

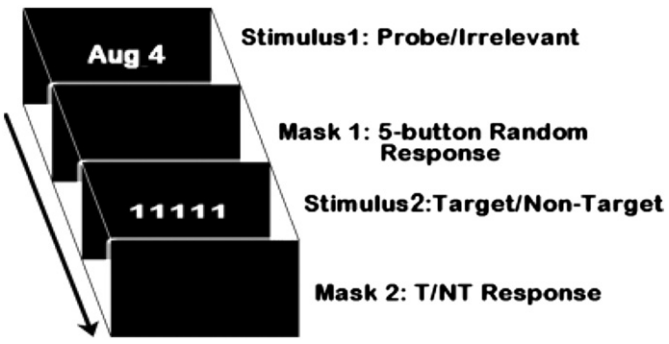

Fig. 1. The complex trial protocol (CTP).

This novel P300 protocol by design has thus far resisted previously effective CMs (Mertens and Allen, 2008; Rosenfeld et al., 2004) in several new studies (see Hu et al., 2012b; Labkovsky and Rosenfeld, 2012b; Rosenfeld and Labkovsky, 2010; Rosenfeld et al., 2008; Winograd and Rosenfeld, 2011). Indeed our novel complex trial protocol (CTP; Rosenfeld et al., 2008) has so far been the only physiologically based CIT reported that is resistant (though not completely immune) to CMs, and additionally provides a simple behavior index, reaction time (RT, the response latency of the "I saw it" response), of the use of CMs by subjects. Thus, the test has so far typically identified recognition of concealed information as well as the attempt by guilty subjects to counter the protocol — which likely constitutes additional evidence of a subject's criminal complicity. Moreover, even in the rare cases we occasionally encounter in which a subject whom we instruct how best to defeat our CTP actually succeeds in not showing the enhanced probe P300 indicator of knowledge recognition, his RT index may still give away his attempt at non-cooperation — useful additional information for enforcement officials.

There have now been a dozen peer-reviewed, published studies of the new P300-based CTP (Hu et al., 2012b; Hu and Rosenfeld, 2012; Labkovsky and Rosenfeld, 2012b; Meixner et al., 2009; Meixner and Rosenfeld, 2010, 2011; Rosenfeld and Labkovsky, 2010; Rosenfeld et al., 2008, 2009; Sokolovsky et al., 2011; Winograd and Rosenfeld, 2011). These studies have established that under certain conditions, the protocol is usually, highly accurate (better than 90% sensitivity and specificity) and CM-resistant (if not immune). We have also reported that in many of these studies, the AUC is typically >.9. In other studies based on an a priori 0.9 decision criterion, the Grier (1971) A′ estimator of AUC (that, based on only one pair of sensitivity/specificity values, is not as reliable as AUC) is also >.9. (See final Table [6] for A′ and AUC data for most of our recent studies reviewed below.) We are not suggesting that these conditions are pre-requisite conditions for usage of the CTP. We are simply noting that these conditions happened to be in effect in the first CTP studies that produced the excellent sensitivity/specificity and discrimination efficiency noted above. The conditions that are this effective usually involve (1) the detecting of self-referring information (as in cases of suspected malingered cognitive deficit in which subjects deny the ability to recall their birth dates, phone numbers, etc.), (2) use of a 5-button response box for the first response (excluding Winograd and Rosenfeld, 2011); the subject randomly selects one of five buttons to press, and (3) usually an asymmetric target probability, meaning a greater probability of a target stimulus following a probe stimulus than following an irrelevant stimulus (described in Rosenfeld et al., 2009). Regarding this latter factor, although we have shown that the asymmetric probability produces results no different than a symmetric target probability with self-referring or autobiographical information (Rosenfeld et al., 2009), we have done more recent studies (not yet published) in which we find that while incidentally acquired information from mock crime scenarios is reasonably well-detected (sensitivity and specificity >80%) with symmetric target probability, it is not as well detected as is autobiographical information (>90%). Moreover, it is clear that one cannot utilize an asymmetric target probability in the field because, given a very low but non-zero probability of a false positive with these methods (Rosenfeld et al., 2009), a defense attorney could argue that the reason a suspect recognizes the probe stimulus in the CTP is that the probe is more often followed by the target stimulus than is any irrelevant stimulus, and that an innocent test suspect could become aware of this during the test and therefore show the P300 recognition sign not because of (non-existent) recognition of a crime-relevant detail, but because the probe has become salient during the CIT as it is more frequently followed by the target than are the irrelevant stimuli.

It was noted that many of our first studies with both the 3SP and the CTP had two apparent problems: 1) autobiographical information is better detected than mock crime information (which would be a serious problem in the field since many more forensic situations involve crime information than involve self-referring information), and 2), many of the previously published studies tested for one probe item per block, and in many of our studies, only one block is used. In the field, it is more persuasive to show that a suspect has knowledge of multiple, versus single crime-relevant items, any single one of which by itself could elicit a false positive.

Regarding 1), first of all, it was pointed out above that the sensitivity difference between autobiographical (90%) and mock crime (80%) scenarios is only 10%. Secondly, in a subsequent section we describe our largely successful attempt to improve CTP sensitivity in mock crime and other scenarios with enhancements to the CTP.

Regarding 2), although we present only one probe per block and protect against false positives in some studies via the use of several blocks, each with a different single probe (e.g., Meixner and Rosenfeld, 2011 used three blocks; see also Rosenfeld, 2011), some workers have used multiple probes within a block of a single session (e.g., Farwell and Donchin, 1991). We have found that this "multiple probe protocol" with its greater task demand produces lower detection rates (Rosenfeld et al., 2004, 2007) than the "single probe protocol". Moreover, in some studies, we have indeed used only a single probe in a single block as that was all that was needed to test some hypothesis, but we do not recommend this practice as a standard one for field use.

The dozen studies noted above have mostly been recently reviewed elsewhere (Rosenfeld, 2011, 2012a; Rosenfeld and Greely, 2012). Very recently, we have undertaken novel studies directed at: 1) Establishing the ecological validity of CIT analogs in the laboratory, 2) Investigating mechanisms by which CMs operate, so as to limit their effect even further, and 3) Finding enhancements to the original CTP aimed at restoring its initial sensitivity with autobiographical information (>0.9) in detection of incidentally acquired information. The present review will therefore be restricted to these three recent lines of investigation, and will conclude with a consideration of legal issues concerning the use of P300-based CITs in U.S. courts.

2. Ecological validity

In the past year, we have conducted studies examining issues that impact the ecological validity of P300-based CITs. Unlike their ANS-based counterparts, which are routinely used in criminal investigations in Japan (Osugi, 2011), the P300 CITs have not yet been used in a similar manner, so it is currently unknown how they might perform in the field.

Our first study on the ecological validity and field relevance of our P300-CIT involved examining the ability of our Complex Trial Protocol (CTP) to detect incidentally acquired information in a mock crime scenario (Winograd and Rosenfeld, 2011). This version of the CTP utilized an asymmetric target probability (a limitation discussed above), where targets more often followed probe stimuli than irrelevant stimuli. For the mock crime, participants were instructed to steal an item that was placed in an envelope in a mailbox in the department office. While specific instructions were given for how to commit the crime, the identity of the item was never revealed prior to the commission of the crime. We did this to ensure that any knowledge the participant had of the stolen item would be solely due to exposure to it through performance of the crime, thus ensuring that it was a purely incidentally acquired episodic memory. Our reasoning behind this choice was that while some crimes may be elaborately planned, in other situations, many details of a crime that might be used in a CIT would be ones that were not well rehearsed, especially in a burglary or theft scenario similar to the one we modeled.

The results from this study were promising. Using an a priori 90% bootstrap criterion (Wasserman and Bockenholt, 1989), we correctly classified 83% (10/12) of guilty participants and 92% (11/12) of innocent participants. The high rates (.87 overall) of both sensitivity and specificity suggested that the CTP is effective at detecting concealed mock crime information in a laboratory scenario. There was one major difference between Winograd and Rosenfeld (2011) and other studies which

used P300-CITs in mock crime scenarios. As previously mentioned, we ensured through our instructions that the only way a participant could gain knowledge of the stolen item was through actually committing the mock crime. In contrast, every other P300-based, mock crime CIT study in the extant literature (see Table 1 below from Winograd and Rosenfeld, submitted for publication) that we found revealed the identity of probe items to participants prior to the execution of the mock crime. In some of these studies, details that would later serve as probes were rehearsed through rote memorization (Farwell and Donchin, 1991; Rosenfeld et al., 2004). In others, details were rehearsed through other means (Abootalebi et al., 2006, 2009; Hu and Rosenfeld, 2012; Mertens and Allen, 2008; Rosenfeld et al., 2007). Only Mertens and Allen (2008) tested for a purely incidentally acquired detail, however it was combined in a block with 11 other probe details which were rehearsed, a procedure that resulted in poor sensitivity. Additionally, this study utilized a virtual reality scenario which also may have contributed to its lower detection rates overall.

Depending on the specific mock crime scenario being studied, exposure to crime details prior to execution of a mock crime and/or a CIT may have effects that harm a study's ecological validity. In a field investigation, the identity of the probe items should never be revealed to a potential suspect (either by investigators or through the media) at any point during the investigation prior to administering a CIT, because simple knowledge of a detail may be sufficient to evoke large P300s to probe items, resulting in a potential false-positive.

In a number of studies using ANS-CITs, the false-positive rates for innocent participants who were exposed to probe details ranged from 25 to 75% (see Bradley et al., 2011, for a review.) These results begged the question of whether the same effect would be true for a P300-based CIT.

To address this question, we (Winograd and Rosenfeld, submitted for publication) employed a simple 2 × 2 between-subjects factorial design with a symmetric conditional target probability. Participants were either naïve (told to steal an "item") or informed (told to steal a "ring") as to the identity of the to-be-stolen item in a mock crime (the same one as in Winograd and Rosenfeld, 2011). For the other manipulation, participants either committed (guilty) or did not commit (innocent) the crime after reading the instructions. By using a fully crossed 2 × 2 design, we were able to determine the effect of prior knowledge on both innocent and guilty participants. A receiver operating characteristic (ROC) analysis (conducted on the number of bootstrap iterations in which probe >Iall) found an area under the curve (AUC) of .956 between the guilty-informed and innocent-naïve groups. The AUC dropped to .852 for the guilty-naïve group. When we compared the guilty-naïve and innocent-informed groups, the AUC was just .519, showing that the informed innocent participants were essentially indistinguishable from those who committed the mock crime. Based upon the ROC analysis, we found that an 80% bootstrap criterion yielded the optimal discrimination between guilty and innocent participants. Correct classification rates based on this cutoff are presented in Table 2.

We note that these detection rates are reported based on using an optimized cutoff calculated from the same data set. However, as noted above in the introduction, methodological changes between studies (such as different target probabilities and subject responses) can make using a single a priori determined cutoff problematic. That said, we interpret the hit rates cautiously.

Sixty-nine percent (9/13) of the participants in the innocent-informed condition were incorrectly classified as "guilty" based on their bootstrap results (compared to a false positive rate of just 2/14 in the innocent-naïve group). Further, ANOVAs revealed that the innocent-informed group was not significantly different from the two guilty groups based on P300 amplitude. Comparing the two guilty conditions, the detection rate for guilty-informed participants (100% — 13/13) was higher than in the guilty-naïve condition (79% — 11/14). This difference neared significance $\chi^2(1,26) = 3.06$, $p = .08$.[1] Overall, these results supported our prediction that prior knowledge of crime details would have an effect on detection rates. Critically, more than half of the innocent participants who knew the identity of the probe item were incorrectly classified as guilty. The results in innocent subjects have major implications for future research and the validity of the CIT, while those in guilty subjects simply support the generality of the effect of prior knowledge.

First, the results show the effect of exposure to crime details on innocent subjects, a serious threat to the validity of a CIT. Simple knowledge of probe items was sufficient to induce a high rate of false positives in innocent participants. Given this finding, details of a crime that will later be used for testing in a CIT would need to be kept confidential, since we demonstrated that simple knowledge of a probe item (based on two brief mentions of the word "ring" in experimental instructions) is sufficient to evoke large P300s, making innocent participants appear to have concealed information. Thus, crime details would need to remain secret, known only to the police, perpetrators, eyewitnesses, and victims of a crime. If information were to be leaked or revealed in the press or by word-of-mouth, a defense attorney could argue that the client knew the specific details of a crime through some legitimate and innocent means. Second, prior knowledge of probe details may bias results towards enhanced sensitivity in guilty participants, an effect that is likely magnified in those studies that utilize a rehearsal or rote memorization procedure. This effect, however, is not as significant or potentially threatening as that found with innocent participants. In the field, there would likely be crimes where specific details are rehearsed and planned (such as targeting a specific residence from which to steal a known valuable item). In situations like this, potential probe details would not be solely episodic and incidentally acquired from the commission of the crime. Researchers employing mock crimes in studies of CITs should take time to carefully develop instructions to model the specific scenario of interest; (e.g. not revealing probe details for a crime in which the stolen item wasn't planned, such as a burglary.

Table 1
P300 mock crime studies.

Authors	Block	Correct detection rates			AUC
		Guilty	Innocent	A′	
Abootalebi et al. (2006)		0.79	0.79	0.87[a]	
Abootalebi et al. (2009)		–	–		0.88[a]
Farwell and Donchin (1991)	Study 1	0.90	.85		0.99[b]
Hu et al. (2013)	High aware	–	–		0.79[c]
	Low aware	–	–		0.55[c]
Hu and Rosenfeld (2012)	Immediate	0.67	1.0		0.89[c]
	1-Month delay	0.75	1.0		0.95[c]
Lui and Rosenfeld (2008)	2 probe	0.87	0.71	0.87[d]	
	3 probe	0.71	0.64	0.76[e]	
Mertens and Allen (2008)		0.47	1.0	0.87[c]	
Rosenfeld et al. (2004)		0.73	0.91	0.89[c]	
Rosenfeld et al. (2007)	1 probe	0.55[f]			
Winograd and Rosenfeld (2011)		0.82	0.92	0.93[c]	

Note: Authors used varying statistical methods of classification. AUC given for papers not reporting separate group hit rates, and for studies including a ROC analysis or classification data for each participant. Reported correct detection rates and calculated A′ (Grier, 1971) values are given for studies that reported only correct classification rates. See original studies for "block" and other details. Hu et al. studies used other conjoint measures besides P300, but only P300 data reported here; see Table 6 for all values in Hu et al. studies.

[a] Wavelet classifier.
[b] Boot-strapped cross-correlation.
[c] Bootstrap amplitude difference.
[d] Bootstrapped spatial–temporal PCA on fronto-central site.
[e] Bootstrapped spatial–temporal PCA on parietal–occipital site.
[f] Bootstrap amplitude difference with multiple blocks (2 of 3 needed for guilty diagnosis).

[1] Campbell (2007) showed that in comparative trial study designs, such as the one employed here, Fischer-Irwin or chi-squared with Yates' correction are too conservative and showed that an N — 1 chi-squared was preferable for detecting differences in outcomes between experimenter-controlled groups when the lowest expected counts are >1.

Table 2
Correct classification rates from Winograd and Rosenfeld (submitted for publication).

Group	N	Correct	Prop
Innocent-naïve	14	12	0.86
Innocent-informed	13	4	0.31
Guilty-naïve	14	11	0.79
Guilty-informed	13	13	1.00

One other threat to the ecological validity of concealed memory detection research is that the memory status examined in the lab scenario often cannot be compared to the memory status examined in the field. In particular, unlike the case in lab studies in which participants are often tested immediately following a mock-crime, suspects in the field may be tested weeks, months, or even years after the crime. It is thus critical for researchers to investigate the influence of time delay on memory detection efficiency. Hu and Rosenfeld (2012) recently explored the potential impact of time delay on crime-related memory and thus detection efficiency in a P300-based CTP. Specifically, participants who enacted the mock crime (here, theft of an exam copy from a professor's mailbox) were asked to come back to the lab for a test about one month after the crime. Another group of comparison participants who enacted the mock crime were tested immediately following the crime. An innocent group of participants was also run to establish false positive rate, and classification efficiency. During the test, one central item (the stolen exam copy) was used as a probe, presented in a series with eight irrelevant stimuli. Results showed that the detection efficiency of the P300-based CTP was not influenced by the 1-month time delay since even after this delay, the classification efficiency measured by receiver operating characteristic (ROC) area analyses (AUC) reached .95 based on the peak–peak P300 amplitude. Similar research using ANS-based CITs and guilty action tests (GATs) have found that the detection of crime details is hindered by a time delay, but that this effect is larger for peripheral than central details (Carmel et al, 2003; Gamer et al., 2010; Nahari and Ben-Shakhar, 2011).

3. Countermeasure (CM) mechanisms

In addition to concerns about ecological validity, another potential threat to P300-based CITs that we have been examining is CM use. As noted, CMs are any method a person can use in an attempt to defeat a CIT. As also noted, in the case of a P300-CIT, the best way to do this is to secretly turn some of the irrelevant stimuli into covert targets. Recently, we have become especially interested in investigating the cognitive mechanisms of effective CMs due to certain unexpected findings by Rosenfeld et al. (2008) and Winograd and Rosenfeld (2011) that probe-irrelevant amplitude differences and detection rates were greater in CM groups than in SG (simply guilty without CMs) groups.

In Rosenfeld et al. (2008) and in Winograd and Rosenfeld (2011), we instructed participants execute a CM to each and every irrelevant stimulus because we believed at that time that this would be the most effective CM strategy for a test subject to use. While this strategy did increase P300 amplitudes to irrelevant stimuli in comparison to amplitudes in the SG group (no CMs), this effect was overcome by an even larger increase in probe P300 amplitude, which was unexpected. By way of explanation, we hypothesized that omitting a response uniquely to one item — the sole un-countered probe item — lends the probe the special salience of what we now call the omit effect, described next:

In Meixner and Rosenfeld (2010), our first study of CM mechanisms, participants executed specific assigned responses on a five-button response box to varying numbers of stimuli. In the guilty no omit condition, a different button response was assigned to each of the four irrelevant stimuli and to the probe. In the innocent omit irrelevant condition, the same procedure was followed, however the fifth stimulus was simply another irrelevant detail, rather than a probe. Finally, participants in a third group, the guilty omit probe condition executed specific and different assigned responses to each of and to only the four irrelevant details, but not to the probe, which was virtually the same as the method of CM use as in Rosenfeld et al. (2008) and Winograd and Rosenfeld (2011).

We called the main finding from this experiment the "omit effect" because when assigned responses are executed to all but one stimulus, this stimulus evokes a large P300, since the single stimulus without an assigned response thereby becomes an oddball, thus making it rare and meaningful, the conditions known to evoke a large P300 (Fabiani et al., 1987). In the innocent omit irrelevant group, the P300 to the omitted (i.e., not responded to) irrelevant was significantly larger than that to the other four irrelevants, and of comparable amplitude to that of the probe stimulus in the guilty no omit condition. So, when guilty omit probe participants executed CMs to each of the irrelevants, but omitted a similar response to the probe, the omit effect added to the standard "oddball" effect (Sutton et al., 1965) due to the probe's meaningfulness, so as to artificially increase the amplitude of the probe P300. This effect reveals that executing covert responses to all irrelevants is — from the perpetrator's point of view — an ineffective method of CM use. Thus in subsequent recent studies, we typically have CM subjects counter only a fraction of the irrelevants presented (e.g., Rosenfeld and Labkovsky, 2010) as a strategy used to maximally challenge our CTP so as to allow our development of the best "counter-countermeasures".

In connection with this strategy, we have also recently learned that the number of countered irrelevants used in a protocol has a significant impact on the effectiveness of CMs during a P300-CIT. In Rosenfeld et al. (2004), participants were instructed to execute a different assigned CM to each countered irrelevant (e.g. press index finger into leg for the first irrelevant, I1, press thumb into leg for the second irrelevant, I2, etc.) The majority of the countermeasures were physical in nature (e.g., toe and finger wiggles versus mental imaging). Since then, we have been using mental CMs (e.g. saying silently to oneself one's own name to I1, father's name to I2, etc.) because these are covert in nature and are not detectable as small physical movements can be (Sokolovsky et al., 2011), and they thus pose a greater challenge to our test. Using these covert mental CMs, we examined the specific impact of the number of countered irrelevants on a P300-CTP. Increasing the number of countered irrelevant stimuli at the same time, of necessity, as increasing the total number of irrelevant stimuli also allowed us to deal with yet another challenge to our ability to detect CM use, namely the simultaneous CM (defined below).

Despite the CTP's initial success in resisting/detecting CMs, the simultaneous CM presented a new type of CM threat: If participants execute a mental CM at the same time as they make the "I saw it" response, the RTs of the "I saw it" response cannot be used as before (e.g., in Rosenfeld et al, 2004; Rosenfeld and Labkovsky, 2010) to detect CM use, although P300 still detects concealed information (Sokolovsky et al., 2011). We note that in previous CTP studies, in which RT was a good index of CM use, participants executed the CM response before the "I saw it" response. To better resist simultaneous CMs, Hu et al., (2012b) increased the number of irrelevant stimuli from four to eight. We reasoned that as the number of irrelevant stimuli increased, guilty participants would find it more difficult to select and execute CMs, and this should be shown via increased RTs to irrelevant stimuli. The abnormally increased RTs might then be used to index even simultaneous CM use.

In this study, five groups were run: (1) a simply guilty (SG) group in which participants were instructed to conceal their hometowns from detection, without any CM use, (2) a 2/8 CM group, in which participants executed CMs to two out of eight irrelevant stimuli, (3) a 4/8 CM group, in which participants executed CMs to four out of eight irrelevant stimuli, and finally, (4) a 6/8 CM group in which participants executed CMs to six out of eight irrelevant stimuli. Mental CMs were used here, i.e. associating each to-be-countered irrelevant with a specific personally significant item such as the participant's first name. A fifth, innocent group was also tested to establish the false positive rate. As

hypothesized, participants especially in the 4/8 and 6/8 CM group (but not in the 2/8 CM group) exhibited significantly longer RTs to irrelevants than to probes, a classic RT pattern indicating CM use, even in this simultaneous CM group. Regarding P300 results, although the amplitude of P300 associated with the probe decreased as countermeasure use increased, the probe-irrelevant difference was still robust: the individual diagnostic accuracies (at bootstrap cutoff = .9) were between 70 and 90% among the three countermeasure groups. Regarding classification efficiency between countermeasure groups and innocent group, the ROC analyses showed that compared to innocent participants, the AUCs associated with 2/8 CMs, 4/8 CMs and 6/8 CMs were 0.93, 0.93 and 0.87, respectively.

Our most recent experiment involving CM mechanisms was designed to determine the specific cognitive mechanisms underlying effective CMs. This idea originally stemmed from a desire to explore a more effective CM that would also be undetectable using a reaction time (RT) analysis. Normally, as noted earlier, CM use can be detected through an increase in RT to the countered irrelevants compared to either the RT to the probe stimulus or to a predetermined baseline (Labkovsky and Rosenfeld 2012b; Rosenfeld et al., 2008). When a participant executes various CMs to a number of irrelevant stimuli, his reaction times tend to increase due to increased task demand (Labkovsky and Rosenfeld 2012b; ; Sokolovsky et al., 2011). We thought that, perhaps, if one were to execute the same CM to all countered irrelevants, rather than executing a different and unique CM to each countered irrelevant, that task demand would be reduced and RT differences between countered and non-countered stimuli would be eliminated.

For this approach to be successful, however, it would mean that the new method (one-for-all) would have to evoke P300s to countered irrelevants comparable in amplitude to those evoked by the old method (one-for-each). Pilot data suggested that the one-for-all method was as effective as the one-for-each method at evoking large P300s to countered irrelevants. Based on this result, we hypothesized that the cognitive mechanism underlying P300s to countered stimuli in a CIT involved a simple recognition process. By assigning a CM (mental, physical, one-for-all or one-for-each) to certain irrelevant stimuli, these irrelevants are each turned into meaningful targets, evoking P300s similar to those evoked by the target stimuli in the 3-stimulus paradigm (Rosenfeld, 2011). We reasoned that it is not the actual execution of a specific CM that is responsible for evoking the P300, but rather the simple recognition and categorization of the stimulus as meaningful — i.e., to be countered — which is critical (Donchin and Coles, 1988). So, we predicted that enhanced P300s to countered irrelevants would be the same no matter whether the all-for-one vs. all-for-each CM method was used.

To test this, we conducted a direct within-subjects comparison of the one-for-all and one-for-each methods of CM use (Winograd and Rosenfeld, 2012a, 2012b). As expected, we found no differences in P300 amplitude (see Fig. 2 below), P300 latency, or bootstrap detection rates, between the one-for-all and one-for-each methods for any stimulus type (Table 3).

Previous research found that more difficult stimulus evaluation processes lead to a reduction in P300 amplitude (Magliero et al., 1984). Additionally, Kutas et al. (1977) determined that P300 latency is affected by stimulus evaluation time. Since the one-for-each method requires more complex stimulus evaluation (identifying not only that a stimulus needs to be countered but also which CM is assigned to it), one would have expected P300s in this method to be smaller and/or more delayed than in the one-for-all method. However, we did not observe either of these effects.

We take these findings as evidence that it is simply recognition of the to-be-countered stimulus as one indeed requiring a CM response that contributes to its salient meaning, based on initial processing of the stimulus and identification of the stimulus itself as being meaningful. This is as opposed to a more complex evaluation process involving first identification, then recall and execution of the specific assigned CM that is responsible for evoking enhanced irrelevant P300s.

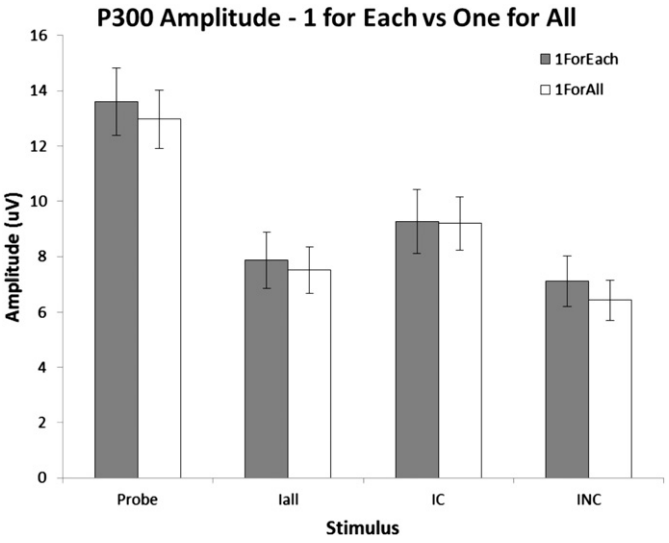

Fig. 2. P300 amplitudes for the one-for-all and one-for-each methods of CM use from Winograd and Rosenfeld (submitted for publication). There were no differences between the two conditions for any stimulus type. IC means countered irrelevant, INC means uncountered irrelevant, Iall means all irrelevants.

We would finally note that we may not in the future be solely dependent upon RT as a CM indicator. Rosenfeld and Labkovsky (2010) serendipitously found an apparently novel, Cz-maximal ERP component called "P900" which appeared only in blocks in which CMs were attempted. Labkovsky and Rosenfeld (2012b) replicated this finding, and Meixner et al. (2013) explored this component in depth, showing that its evocation does not require a deception situation, nor an oddball paradigm.

Susceptibility to CMs is not a problem limited to P300-based credibility assessment tests. As noted above, participants can be also taught to defeat ANS polygraph versions of the CIT (Honts et al., 1996) and the CQT (Honts et al., 1994), and similar effects are found with fMRI-CITs (Ganis et al., 2011). Determining the cognitive mechanisms involved in CM use during a P300-CIT could lead us to create more effective or simpler CMs, or even to develop protocols that are more resistant to them, which, combined with efforts to improve ecological validity, may lead to a P300-CIT that will be field ready.

4. Enhancing the utility of the P300-CIT

Given that real, crime-relevant details in the field may be just incidentally encoded under real life conditions of time pressure and stress, and given that these details may be not rehearsed after a crime, we have discussed above that such encoding limitations can have unfortunate implications for the CIT's field use. Indeed, two field studies using ANS-based CITs showed that the sensitivity was relatively low: When combining SCR and respiration line length, 75.8% of the guilty examinees were correctly identified when the specificity (correct rejections)

Table 3
Detection rates from Winograd and Rosenfeld (submitted for publication).

1Each	1All	First	Second	1All1st	1All2nd	1Each1st	1Each2nd
14/20	15/20	16/20	13/20	8/10	7/10	8/10	6/10
0.7	0.75	0.8	0.65	0.8	0.7	0.8	0.6

Note: "First" (1st) and "Second" (2nd) refer to the first and second halves of the running blocks. "1All" and "1Each" refer to one-for-all and one-for-each conditions, respectively. CM practice seemed to improve CM performance, but not differentially between conditions.

was 94.1% (Elaad, 2011, p. 173). This relatively low sensitivity may also be due to the fact that too few questions were used in these two field studies. Since our sensitivities in the P300-based test are also less with incidentally acquired than with well-rehearsed information, we are also investigating the extent to which we can increase the test's sensitivity to incidentally acquired information while still maintaining adequate specificity. Recently, we have mainly focused on three strategies to this end: a) use of feedback during the test that directs participants' attention to the probe so as to increase the probe's P300; b) use of an additional, P300-independent ERP component, N200, for conjoint P300/N200-based diagnosis so as to possibly increase the protocol's efficiency and c) combining other, separately administered and independent test data with the data from the CTP.

4.1. Using feedback about guilt and heightened probe awareness to increase the probe P300

Despite the fact that recognition is the most important factor driving the CIT effect, researchers have explored whether other factors, such as forcing verbal deceptive responses in the CIT, may facilitate its detection efficiency. For instance, Elaad and Ben-Shakhar (1989) found that if participants answered the crime-relevant questions deceptively (e.g. "No, I don't know it"), the detection efficiency was higher than if they remained silent (Elaad and Ben-Shakhar, 1989; but see Kugelmass et al., 1967). Regarding these data, Elaad and Ben-Shakhar (1989) hypothesized that deceptive responses may increase the probe's "noteworthiness", which increases participants' physiological activities in responses to the probe. However, in a meta-analysis regarding the CIT's detection efficiency, Ben-Shakhar and Elaad (2003) failed to find a significant improvement associated with deceptive responses as a main effect.

Regarding the P300-CIT, Verschuere et al. (2009a) investigated whether or not heightened awareness of one's deceptive responses could increase the P300's detection efficiency in the 3-stimulus CIT protocol, using a participant's first name as the probe. In that study, two groups were compared: prior to the test, one group of participants was instructed to press a button indicating, "YES, I recognize this name" to target, and to press another button meaning "NO, I don't recognize this name" to probe and irrelevant. Thus, participants were made aware of giving deceptive responses to probes. This group was also told, prior to the P300 test (and was thus made explicitly aware) that they would be lying when pressing a NO button to a probe stimulus. In contrast, the other group of participants was simply instructed to press a button indicating "Target" to target and to press another button meaning "Non-target" to probe and irrelevant, just as in Farwell and Donchin (1991). Thus, these participants were not necessarily aware of giving any deceptive responses to probes. The results based on P300, however, failed to find that increased awareness of deceptive responses improved the detection efficiency: the probe-irrelevant difference in the deception group was not larger than that in the non-deception control group, nor were the individual detection rates different.

It should also be mentioned that the probe used in Verschuere et al. (2009a) is perhaps the most salient (easily recognized and attention commanding) type of personal information: one's first name. It is thus possible that any manipulation that aimed to increase the stimulus salience could not add any additional effect to such an inherently salient stimulus due to the possibility of ceiling effects. Therefore, it is premature to conclude (based on Verschuere et al., 2009a) that enhanced deceptive response awareness could not be made to contribute to detection efficiency.

More recently, Rosenfeld et al. (2012b) re-examined this issue also in the 3-stimulus CIT protocol, with a less salient stimulus; the participants' home town. Most critically, in addition to the instruction participants received about their deceptive responses to the probe prior to the test block (as in Verschuere et al., 2009a), they also received (bogus) feedback regarding their possible deceptive responses during the experiment. Specifically, every few minutes, participants in the deceptive group received the following type of (non-veridical) feedback about deception: "From your brainwaves in the past few minutes, we see you are lying on certain trials". In contrast, participants in the non-deceptive, control group received the following type of deception-unrelated but target/non-target related feedback about possible button-press mistakes: "From your brainwaves in the past few minutes, we see you are making some mistaken button presses on certain trials". We hypothesized that participants' attention would be allocated more to the implicit probe-irrelevant dimension than to the explicit target/non-target dimension in the deceptive group via the deception-relevant feedback. In the control group, however, participants' attention would be allocated more to the explicit target/non-target dimension than to the probe-irrelevant dimension. Results clearly showed that with the deception-relevant feedback, the detection efficiency was improved in the deceptive group relative to the control group: 1) At the group level, the probe-irrelevant P300 differences were significantly larger in the deceptive group than in the control group (also, Cohen's $d = 1.93$ vs. 0.73). 2) At the individual level and based on a bootstrap criterion $= .9$, 100% of participants in the deceptive group were correctly identified whereas only 50% of the participants in the non-deceptive control group were correctly identified (Fisher Exact Test $p < .05$).

We note, that this 50% detection rate is actually less than what we typically achieved in our earlier 3-stimulus protocol studies with no feedback. We suspect that the explanation for this resides in our control procedure being distracting in terms of directing subjects' attention away from the critical probe-irrelevant dimension and towards the target-irrelevant dimension. It will be seen in the next paragraph that in a recent, similar feedback study, but based on the CTP (versus 3-stimulus protocol), a similar control group with low sensitivity might also have suffered a related distraction effect. Further research is needed to clarify this.

Since deceptive responses and feedback were helpful in the 3-stimulus P300-CIT, we hypothesized that participants' enhanced awareness of the probe occurrence would also be effective in the countermeasure-resistant CTP. This was studied in Hu et al. (2013), in which the bogus feedback manipulation was used in a CTP in detecting incidentally acquired, mock-crime information. Unlike the older 3-stimulus protocol in which participants explicitly discriminate target and non-target, and implicitly discriminate probe and irrelevant, it is recalled that in the CTP participants simply make button presses indicative of having seen the stimulus regardless of whether a probe or irrelevant was presented. Thus no explicit stimulus discrimination is involved. The feedback in the CTP therefore could not be about deceptive responses.

In Hu et al. (2013), four groups of participants were run: 1) high awareness/guilty; 2) high awareness/innocent; 3) low awareness/guilty and 4) low awareness/innocent. Guilty participants were instructed to enact a mock crime: to steal an item from a professor's mailbox. As in Winograd and Rosenfeld (2011), participants acquired the crime-relevant information (the stolen item was a ring) only by seeing the ring during the mock crime. During the CTP, we used periodic feedback in the high awareness group designed to direct the participant's attention to the probe, e.g. "based on your brainwaves, it seems that there is a certain stimulus that is important to you". In contrast, participants in the low awareness, control group received feedback about general task performance, e.g. "based on your brainwaves, it seems that you are following the task instructions well". Via this manipulation, we hypothesized that participants in the experimental group would be made more aware of the probe occurrence (i.e. high-awareness group) than would the control group (i.e. low-awareness group).

Results again showed that in the high-awareness feedback group, guilty participants showed a larger probe-irrelevant P300 difference than guilty participants in the low-awareness feedback group (Cohen's $d = 1.53$ vs. 0.40). In the high awareness condition, the P300 could

effectively differentiate guilty from innocent participants (AUC = .79, $p < .01$). However, the AUC in the low awareness condition was not significantly different from chance level (AUC = .55, $p > .6$). Moreover, this feedback manipulation did not influence the innocent group, since innocent participants did not recognize any of the stimuli, so that the feedback could not direct attention to any specific stimulus.

4.2. Using the additional ERP component, N200

In addition to the parietally distributed P300, we also found (in Hu et al., 2013) that the frontal-centrally distributed N200 was increased in response to crime-relevant information only in the above described high-awareness guilty group, but neither in the low-awareness guilty group nor in the innocent groups. We hypothesized that this N200 reflected psychological processes other than stimulus recognition in the CIT (which was reflected via the parietally-distributed P300; Donchin and Coles, 1988), based on three lines of evidence: First, recognition itself seemed not sufficient to elicit N200, as we did not find an increased N200 to probes among low-awareness guilty participants. Second, there was no significant correlation between the N200 and the P300, which suggested that the psychological processes reflected by these two components were independent. Third, this frontal-central N200 has previously been found in cognitive control tasks involving conflict monitoring such as the Go/No go task or deception tasks (for a review see Folstein and Van Petten, 2008; Gamer and Berti, 2010; Hu et al., 2011; Wu et al., 2009). As the feedback heightened guilty participants' awareness of the crime-relevant information, it may be hypothesized that they might more likely be engaged in monitoring their responses and performance.

The N200–P300 findings had further implications: First, from a theoretical perspective, the results provided empirical evidence suggesting that despite our and others' previous emphasis that recognition of rarely presented, salient/meaningful items is the key psychological substrate of P300 generation, recognition may not be the only mechanism underlying memory detection. Performance monitoring may also play a role, depending on the specific task demand and context (e.g. receiving high-awareness feedback). Second, from an applied view, as the N200 and P300 may reflect non-overlapping psychological processes in the CIT, combining these components may further improve memory detection efficiency. Indeed, in Hu et al. (2013), combining N200 and P300 in the ROC analyses improved the AUC index (from signal detection theory) of detection efficiency to .91, higher than the AUCs associated with either single indicator (.72–.79).

4.3. Combining different tests

Although the P300-based CIT has received much study, other tests are also available for memory detection (Meijer et al., 2007; Nahari and Ben-Shakhar, 2011; Verschuere et al., 2011). As a different test may be differentially sensitive to other (non-overlapping) psychological processes underlying memory concealment, this suggests the possibility that combining different tests may improve the sensitivity of memory detection. One such attempt has been described in Hu and Rosenfeld (2012). Specifically, during the memory detection procedures following the mock crime, participants underwent two tests: the CTP (a P300-based CIT) and a RT-based autobiographical implicit association test (aIAT; Sartori et al., 2008).

The aIAT is based on pairing responses — i.e., assigning various pairs of responses to single buttons in various trial blocks — which will differ in compatibility depending upon whether or not the subject is guilty or innocent of a specific act, such stealing an exam copy. This act is the mock crime, versus the innocent (control) act of, for example, reading an article. Incompatible responses will have greater reaction times (RTs) and error rates than compatible ones. Thus differing pairs of the following four types of sentences may be assigned to single button responses: 1) generally true sentences (e.g. I am in front of a computer), which is true for all (guilty and innocent) participants; 2) generally false sentences (e.g. I am playing football), which is false for all participants; 3) crime-relevant sentences (e.g. I stole an exam copy), which is true for guilty but false for innocent participants; and 4) innocent act-relevant sentences (e.g. I read an article), which is true for innocent but false for guilty participants.

Based on the shared button-press responses to different types of sentences, the aIAT contains two critical blocks for diagnoses. In one block, participants are instructed to press one button for both generally true and crime-relevant sentences (compatible — both true responses for guilty subjects), but to press another button for both generally false and innocent act-relevant sentences (likewise compatible — both false — for guilty subjects). In the other block, participants are instructed to press one button for both generally true (hereafter simply "true") and innocent act-relevant sentences (incompatible for guilty subjects), and to press another button for both generally false (hereafter simply "false") and crime-relevant sentences (again incompatible for guilty subjects). Since this latter response pairing manipulation results in an incompatibility of responses for guilty subjects, they are expected to show longer RTs and more errors during this incompatible block than during the compatible block, as the crime is the truth for them. For innocent participants, the reverse pattern of behavior is expected, since for them, the innocent act is the truth.

Since the aIAT effect is based on stimulus–response compatibility whereas the P300-based CTP relies on recognition of the crime-relevant detail, Hu and Rosenfeld (2012) hypothesized that these two tests may capture independent psychological processes underlying memory concealment. Thus, we expected that combined tests would outperform either test alone. Indeed, we did not find significant correlations between the results of the CTP and the RT-aIAT. This non-correlation result was similarly reflected via individual diagnoses: there were only a small number of guilty participants that were detected in both the P300-based CTP and the RT-aIAT: thus combining both tests yielded the highest classification efficiency (AUC = .98). Moreover, there may be an additional advantage in using the aIAT combined with the CTP, in that the aIAT by itself has been recently shown to be vulnerable to CMs:

Despite the aIAT's initial success in detecting autobiographical memory (Sartori et al., 2008), it has recently been found that guilty participants can easily control their performance in the aIAT to obtain an innocent result (i.e. false negative, see Hu et al., 2012c; see also Verschuere et al., 2009b). Specifically, upon finishing a baseline aIAT, one group of participants was told about the rationale of the test, and then instructed to speed up their RTs in the incongruent response block so as to distort the test. It was found that participants were able to do this in the subsequent aIAT. In another group in which participants received not only instruction but also training on this "speed up" strategy, they were even more successful in beating the test (see also Hu et al., 2012a, in which training eliminated the behavior differences between honest and deceptive responses). In contrast, participants who simply repeated the aIAT twice (serving as a repetition control group) and participants who merely practiced the incongruent blocks without an intention to speed up (serving as a practice control group) failed to beat the test. Moreover, this speed up strategy was not detectable via previously developed algorithms that were used to detect faking in the aIAT (Agosta et al., 2011). This study clearly demonstrated that an intention to speed up was critical for this faking strategy. Indeed, merely being instructed to speed up was sufficient for participants to voluntarily change their aIAT performance. Given that there are ample IATs available on-line, and that it is very easy for any interested people to understand the rationale of the test, future studies should investigate either how to detect aIAT faking strategies or to develop a more faking-resistant protocol. Research should also investigate whether or not the aIAT, when combined with the CTP remains vulnerable to CMs.

5. Another enhancement: the dual probe complex trial protocol

In the original CTP (Rosenfeld et al., 2008), as described above in Fig. 1, the second part of the complex trial involves target presentation. After the probe or irrelevant is presented, and following the "I saw it" response, a target ("11111") or a non-target ("22222", "33333", etc.) is presented and the subject presses a target or non-target button. Note that the targets and non-targets have nothing to do with the crime, and are thus from an independent stimulus category that is unrelated to the information being probed in the first part of the trial. In contrast, in our most recent, novel dual probe CTP (or DPCTP; Labkovsky and Rosenfeld, 2012a), the first part of the trial is as before, but in the second part of the trial, the targets and non-targets are drawn from another category of crime-relevant details that are independent of the details used in the first part of the trial: Thus these non-targets are either probes or irrelevants, and the target is one, so-designated irrelevant. In other words, the older 3-stimulus protocol or "3SP" (with (1) targets, (2) non-target probes and (3) non-target irrelevants) replaces the set of number strings of the original CTP.

For example, suppose a crime under investigation involves the theft of a diamond ring from a safe inside a hall closet. In the DPCTP, the probe in the first part of the trial would be "diamond ring" and irrelevants would include "ruby bracelet", "sapphire necklace", and so on, just as in the original CTP. In the second part of the trial, however, there are no number strings such as 11111, 22222, and so on. Instead, some of the non-targets would include the correct relevant location, "closet safe" — a probe — and the other non-targets would include other irrelevants such as "kitchen drawer", "cabinet shelf" and so on, while the designated target might be the irrelevant, "bathroom cabinet."

The potentially great advantage of this new DPCTP is that it provides recognition information about two probe items in the same time (i.e., in the same trial block) as the original CTP provided recognition information about just one item. Each item probed in a trial block is like an independent question in the CIT, so that doubling the probes reduces the probability of a false positive identification. Since subjects become fatigued over test time, this efficiency of the DPCTP is to be appreciated. Although the protocol has yet to be fully tested (regarding sensitivity, specificity, and CM resistance) in a P300-CIT dealing with a mock crime, we have begun such a study, to be described shortly.

We first validated the DPCTP with autobiographical information and the resulting accuracies are shown in Table 4 below. In the first (CTP) part of a trial there were four different Irrelevant stimuli and one Probe. All stimuli in the first part were dates. The probe was a subject's birth date. In the second (3SP) part, the stimuli were city names. There were three city names that were Irrelevants (not especially meaningful for the subject), one probe, and one target. The probe was subject's hometown name, and the target was an irrelevant city with an assigned (unique response) significance.

The three groups tested in Table 4 were SG = simply guilty, no CMs; CM = guilty with CMs; and IN = Innocent, no CMs. Fractions represent proportions of correct diagnoses, based on bootstrap tests (with cutoff criteria set to .9; see Rosenfeld, 2011). PART 1 means the first (CTP) part of the trial, PART 2 means the second (3SP) part of the trial, and the third column gives accuracy where guilt/recognition is diagnosed if either Part 1 or 2 or both yield a diagnosis of recognition; note all accuracies and Grier (1971) A′ values >.90. CMs were applied to both parts of the trials in the CM group subjects. These CMs were mental: Upon seeing one of the two, to-be-countered irrelevants, the subjects imagined a specific person.

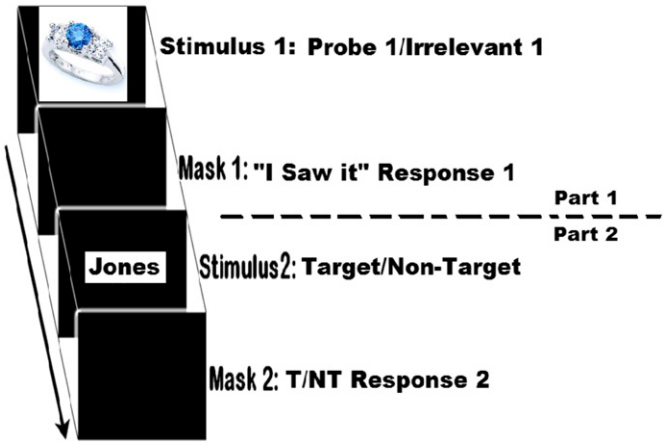

Fig. 3. The dual probe CTP with a mock crime.

As noted above, the DPCTP protocol has yet to be tested (regarding sensitivity, specificity, and CM resistance) in a P300-CIT dealing with a mock crime, however we are partly done with such a study (reported in part in Labkovsky and Rosenfeld, 2012a) to be now summarized: SG and CM subjects were sent to a departmental mailbox with the name "Meixner" on it and told to open an envelope in the box and remove what was inside the envelope and hide the object on their persons, replace the envelope, then return to the lab for the CIT. IN subjects were sent to the door of the department mailbox room and then (as instructed) returned to the CIT test room. The object in the box was originally chosen to be a ring, whose identity was to be learned incidentally by performance of the mock crime. SG, CM, and IN groups were then to be tested with a DPCTP in which the stimulus in the first CTP part was either a picture of the probe (such as the ring probe in Fig. 3 below) or of an irrelevant item of jewelry, or the verbal probe ("Meixner") or irrelevant name (e.g., "Jones" as in the figure below). IN subjects were tested with the same mock crime related items that SG and CM subjects were tested with, but the IN subjects had not seen actually them. For counterbalance purposes, in the second (3SP) part of the trial (lasting about 4.5 s), probes and irrelevants were also either pictured items or names, and the target was one designated irrelevant item. An exemplary event flow chart of a trial in such a DPCTP with the pictured item in the first part and verbal name in the second part of the trial is shown in Fig. 3 below. A dashed line separates the two parts.

We note that for the experiment we now report on, the probe used was a USB drive (not a ring), and the eight irrelevants (including the target for Part 2) were other similar items, including pen, notebook, i-pad, cell phone, watch, CD, computer mouse, and DVD player. The CMs used here were mental as in the autobiographical DPCTP described above. However three irrelevant items were countered with three distinct mental images as the CM responses.

The results (in terms of diagnostic accuracy and AUC) that we have collected so far with this protocol are shown in Fig. 4 and Table 5a below. We note that for these results, the CTP was always the first part of the trial and the 3SP was the second part of the trial. For all

Table 4
Results of first autobiographical DPCTP; A′ = Grier (1971) A′ scores.

Group	Accuracy: PART I at .9 confidence level	Accuracy: PART 2 at .9 confidence level	Accuracy: Either-or at .9 confidence level
SG	12/13 (.92 sensitivity; A′ = .98)	12/13 (.92 sensitivity; A′ = .98)	13/13 (1.0 sensitivity; A′ = .98)
IN	11/12 (.92 specificity)	12/12 (1.0 specificity)	11/12 (.92 specificity)
CM	10/11 (.91 sensitivity; A′ = .98)	10/11 (.91 sensitivity; A′ = .98)	11/11 (1.0 sensitivity; A′ = .98)

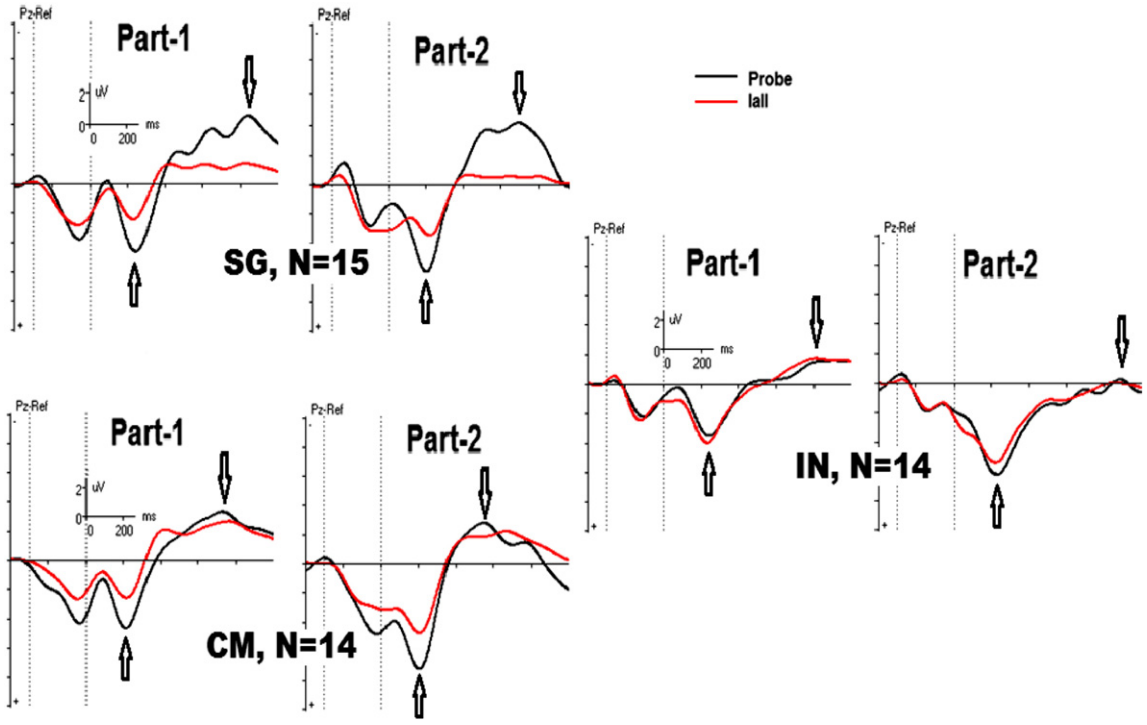

Fig. 4. Grand average ERPs in the three groups from first and second parts of first DPCTP study based on a mock crime. Positive is down. Vertical dotted lines show onset and offset (300 ms later) of stimuli. Arrows show P300s (first up-arrows) and subsequent negative peaks (down arrows) that indicate locations our more accurate peak to peak P300 measurement in deception studies (Rosenfeld, 2011).

groups, pictorial versus verbal presentation in Parts 1 and 2 was counterbalanced across subjects, with pictorial representation coming in Part 1 for about half the subjects in a group, and verbal presentation in Part 1 for the other half.

The Pz grand averages in Fig. 4 show probe P300s in the SG and CM groups that appear clearly larger than superimposed irrelevant P300s ("Iall" means average of all irrelevants), which is not at all the case in the IN group. More quantitatively (Table 5a), these are promising sensitivity and test efficiency (AUC) results based on the "either-or" criterion, and the 7% false positive rate is usually acceptable in many applications, despite the relatively easy "either/or" guilt criterion, and is also comparable to the rates we saw in the original CTPs, which were always <10%. We plan to complete this study in the future with another set of CM, SG, and IN subjects receiving the 3SP in Part 1 and the CTP in Part 2 (the reverse of what was done here).

It is also noted in Table 5a that although accuracies (sensitivities/specificities) increase from either the first or second parts of the trial to the "Either-or" column, the AUCs do not always change appreciably. We note that for all AUCs, the continuous dependent variable is number of iterations in which probe P300 > Iall P300.[2] In the cases of Parts 1 and 2, we use iterations associated with these conditions for each subject. For the "either-or" value, we use whichever iteration number is larger, from Part 1 or Part 2. However this number is still a choice of one number (continuously varying 0–100) from a single subject. For accuracies, however, the "either-or" column makes a dichotomous decision (e.g., 0 or 1 in a range varying from 0 — not knowledgeable to 1 — knowledgeable) that is based on the larger of the two single column numbers (from 0 to 1). AUCs will thus not necessarily vary with accuracy, particularly if the ROCs for the individual trial parts differ from the "either-or" ROC, which they did in the presently described study. It is finally interesting to note that in the CM group, sensitivity is .21 larger with the CTP than with the 3SP, as we suggested (in Rosenfeld et al., 2008) would obtain. The AUC is also correspondingly larger (.22 larger) in Part 1 (CTP) than in Part 2 (3SP).

Table 6 summarizes results (in terms of test discriminability and accuracy) from most recent studies described above in which data were available.

6. Legal issues surrounding the P300-based CIT

While we are nearing 25 years of applied and theoretical research regarding the P300-based CIT, there has been relatively little discussion of the legal relevance of the test,[3] and what little discussion there has been has occurred almost exclusively in the legal literature as opposed to the psychological literature, with little interaction between the two. In light of the applied nature of CIT research, we consider it rather important that the psychological community conducting basic research be aware of the legal hurdles that may prevent the P300-based CIT from being admitted in an American court, given the current state of the field. Our hope is that an increased awareness of these issues will encourage scholars in the field to tailor their research agenda to these issues with the unified goal of allowing the CIT to become a useful tool in the real world. In the remainder of this review, we will briefly discuss the current issues that will need to be addressed before the P300-based CIT would likely be ruled admissible in either federal or state courts. Though the chapter to this point has focused on our Complex Trial

[2] The distributions of these knowledgeable and un-knowledgeable bootstrapped iteration variables are always different for us (in at least six studies); the un-knowledgeable distribution is somewhat normal appearing, but with a long right tail (from the few high values that result in false positives), and the knowledgeable distribution which has many values from 90 to 100 (100 = the maximum) and thus does not look normal, but ogive-like with a long left tail (from the few low values to result in misses). Thus, the resulting ROC is asymmetric and very atypical, such that higher chosen beta or cutoff values (.8 to .9) typically result in higher overall accuracies, as in Table 5b for these DPCTP data.

[3] To demonstrate this, we conducted a search on Westlaw, the most popular legal scholarship database index. On October 28, 2012, Only 56 legal articles contained either the phrase "concealed information test" or the phrase "guilty knowledge test," and only 27 of those articles has more than one usage of either phrase. In contrast, a Google Scholar search for the same terms yielded 224 articles in the psychological literature that use one of the phrases.

Table 5a
Results of DPCTP test on a mock crime; AUC = area under the ROC curve.

Group	Accuracy: PART I at .9 confidence level	Accuracy: PART 2 at .9 confidence level	Accuracy: Either-or at .9 confidence level
SG	AUC = .94 (11/15 = .73 sensitivity)	AUC = .89 (14/15 = .93 sensitivity)	AUC = .95 (15/15 = 1.0 sensitivity)
IN	14/14 (1.0 specificity)	13/14 (.93 specificity)	13/14 (.93 specificity)
CM	AUC = .91 (10/14 = .71 sensitivity)	AUC = .69 (7/14 = .5 sensitivity)	AUC = .89 (13/14 = .93 sensitivity)

Table 5b
Mock crime DPCTP: Overall detection accuracies [(correct detections + correct rejections) / total N] as a function of various cutoff criteria .6 to .9 in SG and CM groups, for Parts 1 & 2 (P1 and P2) of trial or for the Larger of these. Accuracy is seen to mostly decline as criterion declines from .9 to .6.

Criterion	SG			CM		
	P1	P2	Larger	P1	P2	Larger
.9	.86	.93	.97	.86	.71	.93
.8	.90	.90	.93	.86	.68	.89
.7	.90	.83	.83	.82	.68	.79
.6	.79	.76	.83	.75	.64	.64

Protocol (CTP) version of the CIT, we here focus on the more broad question of the legal challenges that all versions of the P300-based CIT might face, both the CIT and its CTP exemplar are likely to face similar issues regarding admissibility.

Any scientific expert testimony offered by a party in either a civil or criminal case must be evaluated by the judge as to its reliability. If the judge finds the evidence to be reliable, he may admit the evidence,[4] and the expert may testify before the jury. How the judge makes this reliability decision depends on jurisdiction: in all federal courts and in the majority of state courts,[5] the judge assesses reliability under the Daubert standard, derived from the United States Supreme Court's opinion in Daubert v. Merrell Dow Pharmaceuticals (1993).[6]

Under that (Daubert) test, the judge must evaluate four factors to determine reliability: (1) "whether [the theory or technique] can be (and has been) tested," (2) "whether the theory or technique has been subjected to peer review or publication," (3) "the known or potential rate of error," and (4) the "general acceptance" of the technique. In Daubert, the Supreme Court specifically noted that they were providing a nonexclusive set of factors, and that some of the factors may not be applicable in all cases, so the Daubert standard provides the trial judge great leeway in making a decision as to reliability and thus, admissibility.

In a minority of state courts, a different (and much simpler) standard is followed, derived from the nearly 100-year-old case Frye v. United States (1923), which itself evaluated the admissibility of ANS polygraph-based lie detection evidence. Under the Frye standard, any scientific evidence that is to be admissible "must be sufficiently established to have gained general acceptance in the particular field in which it belongs" (Frye, 1923). This standard likely sounds familiar: when the Supreme Court decided Daubert, it elected to retain this general acceptance test as one of the four Daubert factors described just above. However, in those state courts that have not adopted Daubert as their standard for admissibility of scientific evidence, the general acceptance inquiry is the sole determinant of reliability. While this standard is less open-ended and thus may restrict judges more, there is still significant interpretation involved in determining how to define the "particular field" to which a science belongs. It is not clear, for example, whether the relevant field for the P300-based CIT would be only deception researchers, or perhaps only psychophysiologists, or something as broad as all psychologists.

This inquiry into reliability is likely to be the primary determinant of the admissibility of the P300-based CIT, and we will assess the test under Daubert and Frye below. Before that, however, it is worth briefly mentioning a second, less discussed legal standard that may also limit admissibility of the P300-based CIT: When confronted with lie detection evidence, many courts have ruled that any testimony that solely assesses the credibility of other witnesses competes with the role of the jury, which has been given the role of determining the credibility of the witnesses. For example, based on this standard, a court might rule inadmissible a polygraph expert whose testimony essentially amounts to a statement that one of the other witnesses was or was not lying. Thus, even if a lie-detection tool achieved 100% accuracy when used in the hands of an expert, it would likely be precluded from use because it would "invade the … province of the jury" (State v. Porter, 1997, p. 769) and "[b]y its very nature … diminish the jury's role in making credibility determinations." (United States v. Scheffer, 2003, p. 313). This view was espoused in the United States Supreme Court's most recent polygraph decision, United States v. Scheffer (2003), in which the Court barred a defendant's attempt to admit ANS polygraph-based control question test evidence indicating that he was truthful. The principal opinion, written by Justice Clarence Thomas, focused on "[p]reserving the court members' core function of making credibility determinations in criminal trials," (p. 312–313) stating plainly that "the jury is the lie detector." (p. 313). Recently, a New York state court tasked with deciding whether to admit fMRI-based lie detection evidence used this rationale in ruling the evidence inadmissible, stating that "credibility is a matter solely for the jury" (Wilson v. Corestaff Services. L.P., 2010, p. 642). Similar rationales have been used to limit the testimony of expert witnesses calling eyewitness testimony into question based on psychological theories such as weapon focus,[7] as these experts' primary purpose is to call the credibility of the eyewitness into question (e.g., Criglow v. State, 1931; People v. Collier, 1952; United States v. Amaral, 1973; United States v. Lumpkin, 1999). Though a full analysis of whether this is a proscriptively good standard is beyond the scope of this review, we note that the standard has come under some criticism in the legal literature due to empirical evidence that lay people are typically poor at making credibility judgments based on demeanor (e.g. Bond and DePaulo, 2006; Fisher, 1997; Meixner, 2012).

However, regardless of whether the standard is a good one, the important thing for CIT researchers to note is that the way the CIT is represented to judges will determine whether it is ruled inadmissible

[4] The judge might also exclude the evidence based on a number of other rationales. For example, under Federal Rule of Evidence 403, if the probative value of the evidence is substantially outweighed by a danger or unfair prejudice, the evidence may be excluded.

[5] As of 2011, 32 states have adopted the Daubert standard (Ambrogi, 2011).

[6] An important distinction to make here is the distinction between validity (a principle's ability to show what it purports to show) and reliability (an application's ability to produce consistent results). The Daubert court noted this distinction in a footnote, and argued that while the terms have differences, they are "different from each other by no more than a hen's kick," and thus the court stated that their focus was on "evidentiary reliability — that is, trustworthiness" (Daubert, p. 590 fn. 9). While it is not exactly clear how evidentiary reliability compares to the more common definitions of validity and reliability, it appears to involve some amalgamation of the two but remains closer to the former, as the court stated, "In a case involving scientific evidence, evidentiary reliability will be based upon scientific validity" (Daubert, p. 590 fn. 9). Thus, when we refer to reliability here in terms of a court making an admissibility decision, we are speaking of evidentiary reliability.

[7] Weapon focus refers to the well-studied phenomenon that a witness to a violent crime tends to divert his or her attention to the weapon that the perpetrator is holding, leaving less attention for other details of the crime and thereby reducing the accuracy of the eyewitness testimony (Loftus et al., 1987).

Table 6
Summary of recent papers; see legend for abbreviations.

Study	CONDIT	AUC	A′	HITS	CUT	FP	INFO	MISC
Labkovsky and Rosenfeld (2011)	IN				.9	.08	BD	
	SG	.97	.98	1.0	.9			
	1/4CM	.88	.96	.92	.9			
	2/4CM	.92	.98	1.0	.9			
	3/4CM	.92	.98	1.0	.9			
	4/4	.94	.96	.92	.9			
Meixner and Rosenfeld (2011)	IN				.5–.9	0	MOCKr	3 block
	SG	1.0	1.0	1.0	.9			P v Iall
	SG	1.0	1.0	1.0	.5			P v Imx
	SG	.98	.96	.83	.75			P v Ibx
Winograd and Rosenfeld (2011)	IN				.9	.08	MOCKi	Asym
	SG	–	.93	.83	.9			
	4/4CM	–	.98	1.0	.9			
Sokolovsky et al. (2011)	IN				.9	.08	BD	
	serCM	–	.93	.83	.9			
	simCM	–	.94	.85	.9			
Hu et al. (2012b)	IN				.9	.08	HT	
	SG	.99	.98	1.0	.9			
	2/8CM	.93	.96	.92	.9			
	4/8CM	.93	.93	.83	.9			
	6/8CM	.87	.91	.7	.9			RTscreened
Hu and Rosenfeld (2012)	IN				.85	0	MOCKr	
P3	SGimmed	.89	.92	.67	.85			
	SGdelay	.95	.94	.75	.85			
P3	SGimmed	.97	.98	1.0	−.48*			P3 + IAT
	SGdelay	.99	.98	1.0	−.43*			P3 + IAT
Hu et al. (2013)	INhiA				.7	.07	MOCKi	
	SGhiA	.79	86	.67	.7		P3	
	SGhiA	.91	.93	.87	−.03*		P3 + N2	
Winograd and Rosenfeld	IN				.8	.14	MOCKi	Not
(submitted for publication)	SG	.85	.89	.79	.8			Not
	SG	.96	.96	1.0	.8			Inform
Labkovsky and Rosenfeld (in prep.)	IN				.9	.07	MOCKi	DP
	SG	.95	.98	1.0	.9	I or II criterion		
	3/8CM	.89	.95	.93	.9	I or II criterion		

Data for selected conditions (CONDIT or blocks) of various recent studies. Selections are based on most sensitive conditions anticipated for possible field use. Original papers and other text herein give all other values. AUCs (areas under the ROC curve) where data were available and Grier (1971) A′ estimates of test discriminability between various knowledgeable and non-knowledgeable groups. IN means the latter. In these IN rows, the false positive proportion (FP) values used for A′ calculations for the given study are shown. They are based usually on an a priori criterion (CUT) of 90% (90 or more bootstrap iterations out of at least 100 in which Probe P300 > Irrelevant P300), except in cases where optimal criteria are given based on ROC data. These CUT choices are justified in original papers and elsewhere herein. Asterisked (*) colored CUT values are not bootstrap scores, but combined z-score values running from −3 to +3. INFO refers to the type of information probed: BD = birth date, HT = home town, MOCKr = mock crime with pre-learned crime details, MOCKi = mock crime based solely on incidentally acquired information during crime act. MISC gives other miscellaneous information. Other abbreviations: HITS = correct detection proportion, SG = simply knowledgeable (guilty) group with no CMs, x/yCM is a CM group in which subject counters x of y total irrelevants, serCM means serial CMs, see text, versus simCM meaning simultaneous CMs. SGimmed refers to an SG group tested immediately after mock crime, SGdelay refers to an SG group tested a month later. IN and SG hiA mean groups in high attention conditions. For Meixner and Rosenfeld (2011), 3 block means 3 blocks of testing with 3 different INFO categories used. P v Iall means the bootstrap test compared probe and all irrelevant average P300s. P v Imx means the bootstrap test compared probe and maximum irrelevant P300. P v Ibx (blind Imax) means the bootstrap tested maximum and next largest P300 averages. All other values in Table 6 are based on one block, only P vs Iall P300 (P3) are compared unless otherwise noted: RTscreened means that for this 6/8CM block, irrelevant P300 averages associated with very high RTs were not used in the Iall average. P3 + IAT is a z-score based metric combining results of these 2 tests. P3 + N2 is a z-score based metric based on both Pz P300 and Fz N200 components. Inform and not refer to whether or not subjects knew probes prior to mock crime. DP is the dual probe CTP (DPCTP) and the criteria are based on the either-or (I or II) criterion (see text).

as impinging on the role of the jury. While the CQT and its variants are truly credibility assessment tests in that they make claims about whether the tested individual was truthful or deceptive, the CIT makes no such claim. Instead, the CIT provides substantive evidence of whether an individual recognizes information that is relevant to the legal question at hand, and leaves the credibility assessment itself to the jury. While this information may undermine the credibility of a witness indirectly, it is no different than any other piece of evidence offered at trial, such as the presence of a fingerprint that may undermine the defendant's denial of guilt. In our recommendations at the end of this review, we will revisit this concept in exploring ways that P300-based CIT researchers can help make this distinction clear.

To date, there are only a few instances of courts discussing the P300-based CIT (or the ANS polygraph-based CIT, for that matter), and there has never been a complete Daubert analysis regarding admissibility of the P300-based CIT, so any discussion of how the test would fare under the four Daubert factors is speculation. However, we can cull some information from related cases in which courts have ruled on fMRI-based lie detection tools, and from the two cases that have ruled on the admissibility of the Brain Fingerprinting test (Farwell and Smith, 2001), a variant of the P300-based CIT.

Perhaps the most relevant test case for admissibility under Daubert was United States v. Semrau (2010), a federal case involving the admissibility of an fMRI-based lie detection test conducted by the Cephos Corporation. While the court ruled the evidence inadmissible on several grounds, the chief concern that the court discussed was regarding the rate of error factor, specifically with regard to real-world error rates: "[T]here are no known error rates for fMRI-based lie detection outside the laboratory setting, i.e. in the 'real-world' or 'real-life' setting." (Semrau, 2010, p. 11). While the court in Semrau was talking specifically about fMRI-based research more analogous to the CQT than to the CIT, the P300-based CIT suffers from the same problem: there is no extant published field study examining the accuracy of the test on real criminal details, so courts are likely to take the stance that there are no reliable error rates that can be assessed. Interestingly, accuracy rates in certain situations in the field may actually exceed those in the laboratory because criminals who planned crimes might be intimately familiar with the details of the crime, leading to better recognition of those

details and thereby larger P300 responses as compared to lab tested participants who are only briefly exposed to the crime-related details (e.g. Meixner and Rosenfeld, 2011). However, this remains conjecture; it should not be relied upon until it has been empirically tested. Similarly, motivation to avoid detection has been shown in lab contexts to lead to increased detection efficiency (Elaad and Ben-Shakhar, 1989; Ben-Shakhar and Elaad, 2003), and such motivation is likely to be significantly greater in the field, where the results of the test could have heavy consequences for the suspect.

However, field tests that have been conducted to date provide some reason for concern that the CIT may be less effective in the field than it is in the lab. Two prominent field tests of the autonomic nervous system-, ANS-based CIT, conducted by Elaad and colleagues, found excellent accuracy rates in classifying innocent individuals (between 95 and 98%), but accuracy rates for guilty individuals were much worse, around 75% and only 50% for a single ANS measure (Elaad, 1990; Elaad et al., 1992). Additionally, one examination of FBI case records indicates that the CIT may only be useful in a small subset of cases (Podlesny, 1993).[8] Similar field testing for the P300-based CIT is critically important, as it is unknown whether the P300 variant would be similarly insensitive in more realistic conditions. Until such testing is done, courts will very likely remain unconvinced by lab analogs.

How the P300-based CIT would fare under the general acceptance prong of the Daubert test (or as the sole criterion in a Frye inquiry) is less clear, and there is little legal precedent to turn to. In Semaru, the court was skeptical that fMRI-based lie detection was generally accepted by the scientific community, primarily based on a rapidly increasing literature from the legal community expressing concern regarding fMRI-based lie detection (Alexander, 2006; Greely and Illes, 2007; Moriarty, 2009; Semrau, 2010). However, this will likely have little bearing on the general acceptance test applied to the P300-based CIT, as the theoretical underpinning and history of research differs greatly between the fMRI-based lie detection applications and the CIT. Notably, Iacono and Lykken (1997) surveyed members of the Society for Psychophysiological Research, asking them whether the CQT and CIT were based on scientifically sound psychological principles or theory. Only 36% of the respondents stated that the CQT was based on scientifically sound principles, but 77% agreed that the CIT was scientifically sound. Among members of the American Psychological Association, the results were very similar, with 30% and 72% agreeing that the CQT and CIT, respectively, were scientifically sound (also from Iacono and Lykken, 1997). These results imply that there may be general acceptance of the CIT in the psychophysiology and psychology communities, though the survey results speak to the ANS variant of the CIT, not the P300-based version.

Only two P300-based CITs have been examined for general acceptance, and both involved the Brain Fingerprinting test, which limits the extent to which these cases are likely to predict future P300 CIT outcomes as this test includes a proprietary and non-peer reviewed (thus controversial) methodology different from other P300-based CITs (termed "MERMER" by Farwell). In Harrington v. State (2001), a criminal appeal in which a defendant attempted to admit the Brain Fingerprinting test results, an Iowa district court ruled that while the P300 component itself is well accepted among psychophysiologists, the MERMER effect, unique to Brain Fingerprinting, was not generally accepted and this fact argued against admissibility. Similarly, in Slaughter v. State (2005), an Oklahoma appeals court found no evidence that Brain Fingerprinting is generally accepted in the psychological community, a sentiment that is echoed in both the psychological and legal literature (e.g. Guadet, 2011; Meijer et al., 2012; Meixner, 2012; Rosenfeld, 2005; Sip et al., 2007; Bizzi et al., 2009). Of course, more mainstream and peer-reviewed P300-based CITs (see Rosenfeld, 2011) would likely fare better under the standard, given the Iacono and Lykken (1997) survey results discussed above.

The final two Daubert factors, testability and peer review, will likely cause fewer problems for the CIT. Like any other diagnostic test, the P300-based CIT's accuracy can be determined by conducting the test on an individual for whom ground truth is known. Though one might argue that laboratory analogs of the CIT mean that it has not yet been tested in realistic scenarios, the field testing that would need to be done to resolve the error rate issues discussed above would also solve any testability problem. Likewise, the peer review factor, which was designed to ensure that other experts in the field have scrutinized a line of research to identify potential confounds and methodological issues (Daubert, 1993), would likely cut in favor of admissibility, given the dozens of P300-based CIT articles published in peer reviewed journals.

Our chief purpose in examining the legal implications of the P300-based CIT is to increase awareness in the psychological community of the legal challenges that the CIT will likely face when offered as evidence in American courts. Given the great potential of the CIT as an accurate alternative to the CQT, which has little support in the academic community (National Research Council, 2003) and other more traditional tests that purport to actually diagnose lies, research aimed toward increasing the admissibility of the test is critical if we hope to allow the test to have real impact. Based on this aim, we provide the following four recommendations to those conducting basic research on the CIT:

- Most importantly, researchers should be looking for opportunities to do field testing of the P300-based CIT. This is obviously very difficult to arrange given the general lack of interest American law enforcement has typically shown regarding the CIT (Kraphol, 2011). Even with police cooperation, a CIT would ideally have to be given prior to interrogation of a suspect where critical details of a crime would very likely be revealed to the suspect, confounding any subsequent CIT testing for those details (although Osugi, 2011, reminds us that this is routinely done in Japan). Likewise, such a study would necessarily involve suspects under full custodial arrest, making institutional review board approval daunting as such participants are considered vulnerable subjects. Perhaps the most difficult problem to solve would be establishing the proper criterion of ground truth. In the lab, participants can be randomly assigned to a guilty or innocent condition, but in the field it is rare that experimenters could be certain of a participant's guilt, absent incontrovertible evidence against him. Additionally, in field scenarios, investigators are likely to be exposed to information about the suspect and the crime prior to administration of the CIT, which may bias results (Ginton et al., 1982). Despite these difficulties, we feel that this should be every lab's top priority: a successful field test demonstrating accuracy rates similar to those found in the lab would strongly increase the likelihood of admissibility of the test under Daubert. Though difficult, a real-world test would be among the most important P300-based CIT studies in the field's history.
- Until such field testing is possible, researchers should be focused on strengthening the realism of laboratory studies to mimic the field as closely as possible. We have focused much of our recent work on improving ecological validity of mock crime CIT testing and found that small exposure to the probe item can cause innocent participants to evoke large P300 waves. Such work can inform future mock crime experiments so that the results are more likely to mimic what would happen in the field. We encourage others to carefully consider the ecological validity of mock crime testing so that these experiments will be as useful as possible to courts later considering the admissibility of the test.
- CIT researchers should collaborate to proactively encourage general acceptance of the P300-based CIT, and the CIT more generally, in the broader psychological community. While this can be done through the more traditional means of publication in widely circulated

[8] However, this issue was evidently overcome in Japan, where the polygraph-based CIT is used regularly in police investigations (Osugi, 2011).

journals like Psychophysiology, the International Journal of Psychophysiology, and the Journal of Experimental Psychology: Applied, more targeted attempts may increase the likelihood that the test is ruled admissible under Daubert or Frye. Researchers should attempt to present at conferences where the CIT may not typically be featured or where the audience is especially broad (e.g., Society for Neuroscience, Association for Psychological Science, or American Polygraph Association) and publish in journals that increase knowledge and understanding of the CIT among psychologists in other areas. Specific research that can demonstrate general acceptance among psychologists or psychophysiologists would prove particularly useful to a court attempting to assess general acceptance; thus, a study similar to the Iacono and Lykken (1997) survey but specifically examining P300-based methods would be simple and highly useful.

- In publishing CIT experiments, researchers should make extremely clear that the nature of the CIT is memory detection; more similar to a DNA test than to a CQT. If courts continue to consider credibility assessment evidence inadmissible based the "jury as the lie detector" standard, the CIT's admissibility rests on being able to show the judge that the CIT does not assess whether a witness is telling the truth, but only what he recognizes. In the past, loaded terms like "lie detection" have been featured in the titles of CIT experiments (e.g. Farwell and Donchin, 1991; and works from our own lab cited in Rosenfeld, 2002). This type of language is likely to mislead courts in the future, and the field would do well to draw a clear distinction between lie detection and memory detection, as Verschuere et al. (2011) do in their book titled, "Memory Detection: Theory and Application of the Concealed Information Test."

We hope this will spark an increased discussion of these legal issues in the psychological community conducing basic CIT research. The field has made great strides in recent years, and we hope that the critical final steps can soon be taken so as to allow the CIT to be used in the real world where it can aid justice.

Acknowledgment

We are most grateful to Gershon Ben Shakhar (who does not always agree with our methods of reporting of diagnostic data) for a valuable review and ongoing thought provoking correspondence. We also appreciated two other, most helpful anonymous reviews. We are also deeply indebted to our Northwestern colleague Satoru Suzuki for ongoing valuable consultation about Signal Detection.

References

Abootalebi, V., Moradi, M.H., Khalilzadeh, M.A., 2006. A comparison of methods for ERP assessment in a P300-based GKT. Int. J. Psychophysiol. 62 (2), 309–320. http://dx.doi.org/10.1016/j.ijpsycho.2006.05.009.

Abootalebi, V., Moradi, M.H., Khalilzadeh, M.A., 2009. A new approach for EEG feature extraction in P300-based lie detection. Comput. Methods Programs Biomed. 94 (1), 48–57. http://dx.doi.org/10.1016/j.cmpb.2008.10.001.

Agosta, S., Ghirardi, V., Zogmaister, C., Castiello, U., Sartori, G., 2011. Detecting fakers of the autobiographical IAT. Appl. Cogn. Psychol. 25 (2), 299–306. http://dx.doi.org/10.1002/acp.1691.

Alexander, A., 2006. Functional magnetic resonance imaging lie detection: is a "brainstorm" heading for the "gatekeeper"? Houston J. Health Law Policy 7, 1–56.

Allen, J.J., Iacono, W.G., Danielson, K.D., 1992. The identification of concealed memories using the event-related potential and implicit behavioral measures: a methodology for prediction in the face of individual differences. Psychophysiology 29 (5), 504–522. http://dx.doi.org/10.1111/j.1469-8986.1992.tb02024.x.

Ambrogi, R., 2011. Two more states adopt Daubert, bringing total to 32. Retrieved from http://www.ims-expertservices.com/blog/2011/two-more-states-adopt-daubert-bringing-total-to-32/.

Ben Shakhar, G., Lieblich, I., Bar-Hillel, M., 1982. An evaluation of polygraphers' judgments: a review from a decision theoretic perspective. J. Appl. Psychol. 67 (6), 701–713. http://dx.doi.org/10.1037/0021-9010.67.6.701 (Dec).

Ben-Shakhar, G., Dolev, K., 1996. Psychophysiological detection through the guilty knowledge technique: effects of mental countermeasures. J. Appl. Psychol. 81 (3), 273–281. http://dx.doi.org/10.1037/0021-9010.81.3.273.

Ben-Shakhar, G., Elaad, E., 2002. The Guilty Knowledge Test (GKT) as an application of psychophysiology: future prospects and obstacles. In: Kleiner, M. (Ed.), Handbook of Polygraph Testing. Academic Press, San Diego, California, pp. 87–102.

Ben-Shakhar, G., Elaad, E., 2003. The validity of psychophysiological detection of information with the Guilty Knowledge Test: a meta-analytic review. J. Appl. Psychol. 88 (1), 131–151. http://dx.doi.org/10.1037/0021-9010.88.1.131.

Ben-Shakhar, G., Kremnitzer, M., 2011. The Concealed Information Test in the courtroom: legal aspects. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, New York, New York, pp. 276–292.

Bizzi, E., Hyman, S.E., Raichle, M.E., Kanwisher, N., Phelps, E.A., Morse, S.J., Sinnot-Armstrong, W., Rackoff, J.S., Greely, H.T., 2009. Using Imaging to Identify Deceit. American Academy of Arts & Sciences, Cambridge, Massachusetts.

Bond Jr., C.F., DePaulo, B.M., 2006. Accuracy of deception judgments. Pers. Soc. Psychol. Rev. 10 (3), 214–234. http://dx.doi.org/10.1207/s15327957pspr1003_2.

Bradley, M.T., Barefoot, C.A., Arsenault, A.M., 2011. Leakage of information to innocent suspects. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory detection: Theory and application of the Concealed Information Test. Cambridge University Press, pp. 187–199. http://dx.doi.org/10.1017/CBO9780511975196.011.

Campbell, I., 2007. Chi–squared and Fisher–Irwin tests of two–by–two tables with small sample recommendations. Stat. Med. 26 (19), 3661–3675. http://dx.doi.org/10.1002/sim.2832.

Carmel, D., Dayan, E., Naveh, A., Raveh, O., Ben-Shakhar, G., 2003. Estimating the validity of the guilty knowledge test from simulated experiments: the external validity of mock crime studies. J. Exp. Psychol. Appl. 9, 261–269.

Council, National Research, 2003. The Polygraph and Lie Detection. Joseph Henry Press (0309084369).

Criglow v. State, 1931. 36 S.W.2d 400 (Ark. 1931).

Daubert v. Merrell Dow Pharmaceuticals, 1993. 509 U.S. 579.

Donchin, E., Coles, M.G., 1988. Is the P300 component a manifestation of context updating? Behav. Brain Sci. 11 (3), 355–372. http://dx.doi.org/10.1017/S0140525X00058027.

Donchin, E., Kramer, A.F., Wickens, C., 1986. Applications of brain event-related potentials to problems in engineering psychology. In: Coles, M.G.H., Donchin, E., Porges, S.W. (Eds.), Psychophysiology: Systems, Processes, and Applications. The Guilford Press, New York, pp. 702–718.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7 (1), 1–26.

Elaad, E., 1990. Detection of guilty knowledge in real-life criminal investigations. J. Appl. Psychol. 75 (5), 521. http://dx.doi.org/10.1037/0021-9010.75.5.521.

Elaad, E., 2011. Validity of the Concealed Information Test in realistic contexts. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, New York, New York, pp. 171–186.

Elaad, E., Ben-Shakhar, G., 1989. Effects of motivation and verbal-response type on psychophysiological detection of information. Psychophysiology 26 (4), 442–451. http://dx.doi.org/10.1111/j.1469-8986.1989.tb01950.x.

Elaad, E., Ben-Shakhar, G., 1991. Effects of mental countermeasures on psychophysiological detection in the guilty knowledge test. Int. J. Psychophysiol. 11 (2), 99–108. http://dx.doi.org/10.1016/0167-8760(91)90001-E.

Elaad, E., Ginton, A., Jungman, N., 1992. Detection measures in real-life criminal guilty knowledge tests. J. Appl. Psychol. 77 (5), 757–767. http://dx.doi.org/10.1037/0021-9010.77.5.757.

Fabiani, M., Gratton, G., Karis, D., Donchin, E., 1987. The definition, identification, and reliability of measurement of the P300 component of the event-related brain potential. In: Ackles, P.K., Jennings, J.R., Coles, M.G.H. (Eds.), Advances in Psychophysiology, vol. 2. JAI Press, Greenwich, CT, pp. 1–78.

Farwell, L.A., Donchin, E., 1991. The truth will out: interrogative polygraphy ("lie detection") with event-related brain potentials. Psychophysiology 28 (5), 531–547. http://dx.doi.org/10.1111/j.1469-8986.1991.tb01990.x.

Farwell, L.A., Smith, S.S., 2001. Using brain MERMER testing to detect knowledge despite efforts to conceal. [Clinical Trial]. J. Forensic Sci. 46 (1), 135–143.

Fisher, G., 1997. Jury's rise as lie detector. The Yale Law J. 107, 575–713.

Folstein, J.R., Van Petten, C., 2008. Influence of cognitive control and mismatch on the N2 component of the ERP: a review. Psychophysiology 45 (1), 152–170. http://dx.doi.org/10.1111/j.1469-8986.2007.00602.x.

Frye v. United States, 1923. 293 F. 1013, 1014 (D.C. Cir.1923).

Furedy, J.J., Liss, J., 1986. Countering confession induced by the polygraph: of confessionals and psychological rubber hoses. Crim. Law Q. 29, 91.

Gallai, D., 1999. Polygraph evidence in federal courts: should it be admissible. Am. Crim. Law Rev. 36, 87.

Gamer, M., Berti, S., 2010. Task relevance and recognition of concealed information have different influences on electrodermal activity and event-related brain potentials. Psychophysiology 47 (2), 355–364. http://dx.doi.org/10.1111/j.1469-8986.2009.00933.x.

Gamer, M., Kosiol, D., Vossel, G., 2010. Strength of memory encoding affects physiological responses in the Guilty Action Test. Biol. Psychol. 83, 101–107.

Ganis, G., Rosenfeld, J.P., Meixner, J., Kievit, R.A., Schendan, H.E., 2011. Lying in the scanner: covert countermeasures disrupt deception detection by functional magnetic resonance imaging. NeuroImage 55 (1), 312–319. http://dx.doi.org/10.1016/j.neuroimage.2010.11.025.

Ginton, A., Daie, N., Elaad, E., Ben-Shakhar, G., 1982. A method for evaluating the use of the polygraph in a real-life situation. J. Appl. Psychol. 67 (2), 131–137.

Greely, H.T., Illes, J., 2007. Neuroscience-based lie detection: the urgent need for regulation. Am. J. Law Med. 33, 377–431.

Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley, New York0-471-32420-5.

Grier, J.B., 1971. Non-parametric indexes for sensitivity and bias: computing formulas. Psychol. Bull. 75, 424–429.
Guadet, L.M., 2011. Brain fingerpriting, scientific evidence, and Daubert: a cautionary lesson from India. Jurimetrics 51 (3), 293–319.
Honts, C.R., Raskin, D.C., Kircher, J.C., 1994. Mental and physical countermeasures reduce the accuracy of polygraph tests. J. Appl. Psychol. 79 (2), 252–259. http://dx.doi.org/10.1037/0021-9010.79.2.252.
Honts, C.R., Devitt, M.K., Winbush, M., Kircher, J.C., 1996. Mental and physical countermeasures reduce the accuracy of the concealed knowledge test. Psychophysiology 33 (1), 84–92. http://dx.doi.org/10.1111/j.1469-8986.1996.tb02111.x.
Honts, C.R., Amato, S.L., Gordon, A.K., 2001. Effects of spontaneous countermeasures used against the comparison question test. Polygraph 30 (1), 1–9.
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. http://dx.doi.org/10.1111/j.1469-8986.2012.01389.x.
Hu, X., Wu, H., Fu, G., 2011. Temporal course of executive control when lying about self- and other-referential information: an ERP study. Brain Res. 1369, 149–157. http://dx.doi.org/10.1016/j.brainres.2010.10.106.
Hu, X., Chen, H., Fu, G., 2012a. A repeated lie becomes a truth? The effect of intentional control and training on deception. Front. Psychol. 3. http://dx.doi.org/10.3389/fpsyg.2012.00488.
Hu, X., Hegeman, D., Landry, E., Rosenfeld, J.P., 2012b. Increasing the number of irrelevant stimuli increases ability to detect countermeasures to the P300-based Complex Trial Protocol for concealed information detection. Psychophysiology 49 (1), 85–95. http://dx.doi.org/10.1111/j.1469-8986.2011.01286.x.
Hu, X., Rosenfeld, J.P., Bodenhausen, G.V., 2012c. Combating automatic autobiographical associations: the effect of instruction and training in strategically concealing information in the autobiographical implicit association test. Psychol. Sci. 23 (10), 1079–1085. http://dx.doi.org/10.1177/0956797612443834.
Hu, X., Pornpattananangkul, N., Rosenfeld, J.P., 2013. N200 and P300 as orthogonal and integrable indicators of distinct awareness and recognition processes in memory detection. Psychophysiology 50 (5). http://dx.doi.org/10.1111/psyp.12018.
Iacono, W.G., 2011. Encouraging the use of the Guilty Knowledge Test (GKT): what the GKT has to offer law enforcement. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, New York, New York, pp. 12–26.
Iacono, W.G., Lykken, D.T., 1997. The validity of the lie detector: two surveys of scientific opinion. J. Appl. Psychol. 82 (3), 426–433. http://dx.doi.org/10.1037/0021-9010.82.3.426.
Johnson, R., 1993. On the neural generators of the P300 component of the event-related potential. Psychophysiology 30 (1), 90–97.
Kassin, S.M., 2008. False confessions causes, consequences, and implications for reform. Curr. Dir. Psychol. Sci. 17 (4), 249–253. http://dx.doi.org/10.1111/j.1467-8721.2008.00584.x.
Kraphol, D.J., 2011. Limitations of the Concealed Information Test in criminal cases. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, New York, NY, pp. 151–170.
Kugelmass, S., Lieblich, I., Bergman, Z., 1967. The role of "lying" in psychophysiological detection. Psychophysiology 3, 312–315. http://dx.doi.org/10.1111/j.1469-8986.1967.tb02711.x.
Kutas, M., McCarthy, G., Donchin, E., 1977. Augmenting mental chronometry: the P300 as a measure of stimulus evaluation time. Science 197 (4305), 792–795. http://dx.doi.org/10.1126/science.887923.
Labkovsky, E., Rosenfeld, J.P., 2012a. Detection of concealed information in a mock crime scenario with a novel Dual Probe Complex Trial Protocol (DPCTP) and pictorial stimuli. [Poster Abstract] Psychophysiology 49 (S1), S94.
Labkovsky, E., Rosenfeld, J.P., 2012b. The P300-based, complex trial protocol for concealed information detection resists any number of sequential countermeasures against up to five irrelevant stimuli. Appl. Psychophysiol. Biofeedback 37 (1), 1–10. http://dx.doi.org/10.1007/s10484-011-9171-0.
Loftus, E.F., Loftus, G.R., Messo, J., 1987. Some facts about" weapon focus.". Law Hum. Behav. 11 (1), 55. http://dx.doi.org/10.1007/BF01044839.
Lui, M., Rosenfeld, J.P., 2008. Detection of deception about multiple, concealed, mock crime items, based on a spatial–temporal analysis of ERP amplitude and scalp distribution. Psychophysiology 45 (5), 721–730.
Lykken, D.T., 1959. The GSR in the detection of guilt. J. Appl. Psychol. 43 (6), 385–388. http://dx.doi.org/10.1037/h0046060.
Lykken, D.T., 1998. A Tremor in the Blood: Uses and Abuses of the Lie Detector. Plenum Trade, New York, New York.
Magliero, A., Bashore, T.R., Coles, M.G., Donchin, E., 1984. On the dependence of P300 latency on stimulus evaluation processes. Psychophysiology 21 (2), 171–186. http://dx.doi.org/10.1111/j.1469-8986.1984.tb00201.x.
Meijer, E.H., Smulders, F.T., Johnston, J.E., Merckelbach, H.L., 2007. Combining skin conductance and forced choice in the detection of concealed information. Psychophysiology 44 (5), 814–822. http://dx.doi.org/10.1111/j.1469-8986.2007.00543.x.
Meijer, E.H., Ben-Shakhar, G., Verschuere, B., Donchin, E., 2012. A comment on Farwell (2012): brain fingerprinting: a comprehensive tutorial review of detection of concealed information with event-related brain potentials. Cogn. Neurodyn. 1–4. http://dx.doi.org/10.1007/s11571-012-9217-x.
Meixner, J.B., 2012. Liar, liar, jury's the trier? The future of neuroscience-based credibility assessment and the court. Northwest. Univ. Law Rev. 106 (3), 1451–1488.
Meixner, J.B., Rosenfeld, J.P., 2010. Countermeasure mechanisms in a P300-based concealed information test. Psychophysiology 47 (1), 57–65. http://dx.doi.org/10.1111/j.1469-8986.2009.00883.x.
Meixner, J.B., Rosenfeld, J.P., 2011. A mock terrorism application of the P300-based concealed information test. Psychophysiology. http://dx.doi.org/10.1111/j.1469-8986.2010.01050.x.
Meixner, J.B., Haynes, A., Winograd, M.R., Brown, J., Rosenfeld, J.P., 2009. Assigned versus random, countermeasure-like responses in the P300 based complex trial protocol for detection of deception: task demand effects. Appl. Psychophysiol. Biofeedback 34 (3), 209–220. http://dx.doi.org/10.1007/s10484-009-9091-4.
Meixner, J.B., Labkovsky, E., Rosenfeld, J.P., Winograd, M.W., Sokolovsky, A., Weishaar, J., Ullman, T., 2013. P900: a putative novel ERP component that indexes countermeasure use in the P300-based concealed information test. Appl. Psychophysiol. Biofeedback 38, 121–132.
Mertens, R., Allen, J.J., 2008. The role of psychophysiology in forensic assessments: deception detection, ERPs, and virtual reality mock crime scenarios. Psychophysiology 45 (2), 286–298. http://dx.doi.org/10.1111/j.1469-8986.2007.00615.x.
Moriarty, J., 2009. Visions of deception: neuroimaging and the search for evidential truth. Akron Law Rev. 42, 739.
Nahari, G., Ben-Shakhar, G., 2011. Psychophysiological and behavioral measures for detecting concealed information: the role of memory for crime details. Psychophysiology 48 (6), 733–744. http://dx.doi.org/10.1111/j.1469-8986.2010.01148.x.
Osugi, A., 2011. Daily application of the Concealed Information Test: Japan. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, New York, NY.
Patrick, C.J., 2011. Science on the rise: birth and development of the Concealed Information Test. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, New York, New York, pp. 3–11.
Podlesny, J.A., 1993. Is the guilty knowledge polygraph technique applicable in criminal investigations? A review of FBI case records. Crime Lab. Dig. 20 (3), 57–61.
People v. Collier, 1952. 249 P.2d 72 (Cal Dist. Ct. App. 1952).
Reid, J.E., Inbau, F.E., 1977. Truth and Deception: The Polygraph ("Lie Detection") Technique. Williams & Wilkins, Baltimore.
Rosenfeld, J.P., 2002. Event-related potentials in the detection of deception, malingering, and false memories. In: Kleiner, M. (Ed.), Handbook of Polygraph Testing. Academic Press, New York, pp. 265–286.
Rosenfeld, J.P., 2005. Brain fingerprinting: a critical analysis. Sci. Rev. Mental Health Pract. 4, 20–37.
Rosenfeld, J.P., 2011. P300 in detecting concealed information. In: Verschuere, B., Ben-Shakhar, G., Meijer, E. (Eds.), Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, New York, NY, pp. 63–89.
Rosenfeld, J.P., Greely, H.T., 2012. Deception, detection of, P300 event-related potential (ERP). Wiley Encyclopedia of Forensic Science.John Wiley & Sons, Ltd.
Rosenfeld, J.P., Labkovsky, E., 2010. New P300-based protocol to detect concealed information: resistance to mental countermeasures against only half the irrelevant stimuli and a possible ERP indicator of countermeasures. Psychophysiology 47 (6), 1002–1010. http://dx.doi.org/10.1111/j.1469-8986.2010.01024.x.
Rosenfeld, J.P., Cantwell, B., Nasman, V.T., Wojdac, V., Ivanov, S., Mazzeri, L., 1988. A modified, event-related potential-based guilty knowledge test. Int. J. Neurosci. 42 (1–2), 157–161. http://dx.doi.org/10.3109/00207458808985770.
Rosenfeld, J.P., Angell, A., Johnson, M., Qian, J.H., 1991. An ERP-based, control-question lie detector analog: algorithms for discriminating effects within individuals' average waveforms. Psychophysiology 28 (3), 319–335. http://dx.doi.org/10.1111/j.1469-8986.1991.tb02202.x.
Rosenfeld, J.P., Soskins, M., Bosh, G., Ryan, A., 2004. Simple, effective countermeasures to P300-based tests of detection of concealed information. Psychophysiology 41 (2), 205–219. http://dx.doi.org/10.1111/j.1469-8986.2004.00158.x.
Rosenfeld, J.P., Shue, E., Singer, E., 2007. Single versus multiple probe blocks of P300-based concealed information tests for self-referring versus incidentally obtained information. Biol. Psychol. 74 (3), 396–404. http://dx.doi.org/10.1016/j.biopsycho.2006.10.002.
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. http://dx.doi.org/10.1111/j.1469-8986.2008.00708.x.
Rosenfeld, J.P., Tang, M., Meixner, J., Winograd, M., Labkovsky, E., 2009. The effects of asymmetric vs. symmetric probability of targets following probe and irrelevant stimuli in the complex trial protocol for detection of concealed information with P300. Physiol. Behav. 98 (1–2), 10–16. http://dx.doi.org/10.1016/j.physbeh.2009.03.030.
Rosenfeld, J.P., Ben-Shakhar, G., Ganis, G., 2012a. Detection of concealed stored memories with psychophysiological and neuroimaging methods. In: Nadel, L., Sinnott-Armstrong, W.P. (Eds.), Memory and Law. Oxford University Press, New York, pp. 263–306.
Rosenfeld, J.P., Hu, X., Pederson, K., 2012b. Deception awareness improves P300-based deception detection in concealed information tests. Int. J. Psychophysiol. http://dx.doi.org/10.1016/j.ijpsycho.2012.06.007.
Sartori, G., Agosta, S., Zogmaister, C., Castiello, U., 2008. How to accurately detect autobiographical events. Psychol. Sci. 19 (8), 772–780. http://dx.doi.org/10.1111/j.1467-9280.2008.02156.x.
Saxe, L., Ben-Shakhar, G., 1999. Admissibility of polygraph tests: the application of scientific standards post-Daubert. Psychol. Publ. Pol. Law 5, 203–1173.
Sip, K.E., Roepstorff, A., McGregor, W., Firth, C.D., 2007. Detecting deception: the scope and limits. Trends Cogn. Sci. 12 (2), 48–53. http://dx.doi.org/10.1016/j.tics.2007.11.008.
Slaughter v. State, 2005. 105 P.3d 832 (Okla. Crim. App. 2005).
Sokolovsky, A., Rothenberg, J., Labkovsky, E., Meixner, J., Rosenfeld, J.P., 2011. A novel countermeasure against the reaction time index of countermeasure use in the P300-based complex trial protocol for detection of concealed information. Int. J. Psychophysiol. 81 (1), 60–63. http://dx.doi.org/10.1016/j.ijpsycho.2011.03.008.
State v. Porter, 1997. 698 A.2d 739 (Conn. 1997).
Sutton, S., Bararen, M., Zubin, J., John, E.R., 1965. Evoked-potential correlates of stimulus uncertainty. Science 150, 1187–1188. http://dx.doi.org/10.1126/science.150.3700.1187.
United States v. Amaral, 1973. 488 F.2d 1148 (9th Cir. 1973).
United States v. Lumpkin, 1999. 192 F.3d 280 (2d Cir. 1999).

United States v. Scheffer, 1998. 523 U.S. 303 (1998).
Van Erkel, Pattynama, 1998. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. Eur. J. Radiol. 27, 88–94.
Verschuere, B., Rosenfeld, J.P., Winograd, M.R., Labkovsky, E., Wiersema, R., 2009a. The role of deception in P300 memory detection. Leg. Criminol. Psychol. 14 (2), 253–262. http://dx.doi.org/10.1348/135532508X384184.
Verschuere, B., Prati, V., De Houwer, J., 2009b. Cheating the lie detector: faking in the autobiographical IAT. Psychol. Sci. 20, 410–413. http://dx.doi.org/10.1111/j.1467-9280.2009.02308.x.
Verschuere, B., Ben-Shakhar, G., Meijer, E., 2011. Memory Detection: Theory and Application of the Concealed Information Test. Cambridge University Press, New York, New York.
Warden, R., 2012. Developments in criminal justice: whither false confessions? Chicago Bar Assoc. Rec. 26 (2), 28–33.
Wasserman, S., Bockenholt, U., 1989. Bootstrapping: applications to psychophysiology. Psychophysiology 26 (2), 208–221. http://dx.doi.org/10.1111/j.1469-8986.1989.tb03159.x.
Wilson v. Corestaff Services L.P., 2010. 900 N.Y.S.2d 639 (Sup. Ct. 2010).
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. http://dx.doi.org/10.1111/j.1469-8986.2010.01054.x.
Winograd, M.R., Rosenfeld, J.P., 2012a. Countermeasure mechanisms in the complex trial protocol. Int. J. Psychophysiol. 85, 305. http://dx.doi.org/10.1016/j.ijpsycho.2012.06.046 (Proceedings of the 16th World Congress of Psychophysiology of the International Organization of Psychophysiology (IOP) Pisa, Italy September 13–17, 2012).
Winograd, M.R., Rosenfeld, J.P., 2012b. Ecological validity and countermeasures mechanisms in P300 concealed information tests. Psychophysiology 49, S94 (abstract).
Winograd, M.R., Rosenfeld, J.P., 2013. The impact of prior knowledge from participant instructions in a mock crime P300 concealed information test (submitted for publication).
Wu, H., Hu, X., Fu, G., 2009. Does willingness affect the N2–P3 effect of deceptive and honest responses? Neurosci. Lett. 467 (2), 63–66. http://dx.doi.org/10.1016/j.neulet.2009.10.002.