Memory Detection

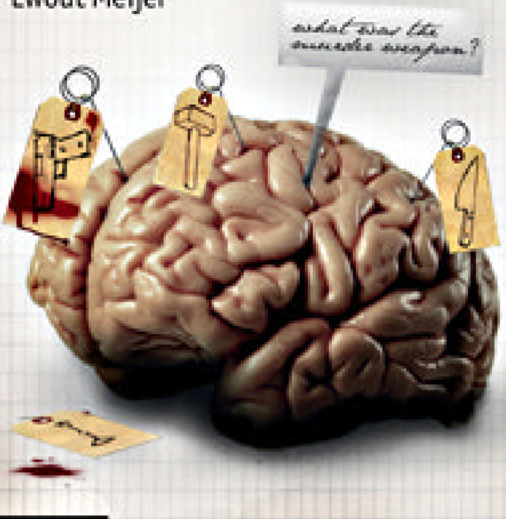
Theory and Application of the Concealed Information Test

EDITED BY

Bruno Verschuere

Gershon Ben-Shakhar

Ewout Meijer



CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521769525

4 P300 in detecting concealed information

J. Peter Rosenfeld

Overview: This chapter reviews the use of the P300 ERP in the detection of concealed information since the first published papers in the late 1980s. First, there is a description of P300 as a cortical signal of the recognition of meaningful information. This attribute was applied directly to concealed information detection in the first P300-based CIT protocol called the "three stimulus protocol." There follows a detailed discussion and review of the methods of analysis used to determine guilt or innocence with the P300, as well as the major papers using and extending the three stimulus protocol in areas beyond those reported in the first publications. This discussion closes with the problematic findings showing that the P300-based, three stimulus protocol is vulnerable to countermeasures. The author's theoretical efforts to understand countermeasure vulnerability with this protocol are then described, followed by an introduction of the theoretically based novel protocol (called the Complex Trial Protocol or CTP) developed to resist countermeasures to P300-based CITs. The use of the CTP in detecting self-referring as well as incidentally acquired information (e.g., in a mock crime scenario) are described, as well as its recent use in detection of details of planned acts of terror prior to actual criminal acts. The use of reaction time as well as a novel ERP component called P900 for detecting countermeasures is also described. The chapter concludes with some caveats about remaining research issues.

The P300 event-related potential

Between an electrode placed on the scalp surface directly over the brain and another electrode connected to an electrically neutral part of the

Acknowledgment: I would like to thank John J. B. Allen and Gershon Ben-Shakhar for good reviews (containing several helpful edits) of this chapter.

head (i.e., remote from brain cells, such as the earlobe), an electrical voltage, varying as a function of time, exists. These voltages comprise the spontaneously ongoing electroencephalogram (EEG), and are commonly known as brain waves. If during the recording of EEG, a discrete stimulus such as a light flash occurs, the EEG will break into a series of larger peaks and valleys lasting up to two seconds after the stimulus onset. These waves, signaling the arrival in the cortex of neural activity elicited by the stimulus, comprise the wave series called the event-related potential or ERP.

The ERP is of a small magnitude compared to the ongoing EEG, so it is often obscured in single trials. Thus, one typically averages the EEG samples of many repeated presentations of either the same stimulus or several stimuli of one particular category (e.g., female names, weapon types, etc.). The resulting averaged stimulus-related activity is revealed as the ERP, while the non-stimulus-related features of the EEG average out, approaching a straight line. The P300 is a special ERP component that results whenever a *meaningful* piece of information is *rarely* presented among a random series of more frequently presented, non-meaningful stimuli often of the same category as the meaningful stimulus. For example, Figure 4.1 shows a set of three pairs of superimposed ERP averages from three scalp sites (called Fz, Cz, and Pz overlaying the midline frontal, central, and parietal areas of the scalp, respectively) of one subject, who was viewing a series of test items on a display (from Rosenfeld *et al.*, 2004). On 17 percent of the trials, a meaningful item (e.g., the subject's birth date) was presented, and on the remaining 83 percent of the randomly occurring trials, other items with no special meaning to the subject (e.g., other dates) were presented. The two superimposed waveforms at each scalp site represent averages of ERPs to (1) meaningful items and to (2) other items. In response to only the meaningful items, a large down-going P300, indicated with thick vertical lines, is seen, which is absent in the superimposed waveforms evoked by non-meaningful stimuli. The wave labeled "EOG" is a simultaneous recording of eye-movement activity. As required for sound EEG recording technique, EOG is flat during the segment of time when P300 occurs, indicating that no artifacts due to eye movements are occurring. Clearly, the *rare, recognized, meaningful* items elicit P300, the other items do not. (Note that electrically positive brain activity is plotted down.) It should be evident that the ability of P300 to signal the involuntary recognition of meaningful information suggests that the wave could be used to signal recognized "guilty knowledge" known only to those familiar with the crime details, such as a guilty perpetrators, accomplices, witnesses, and police investigators.

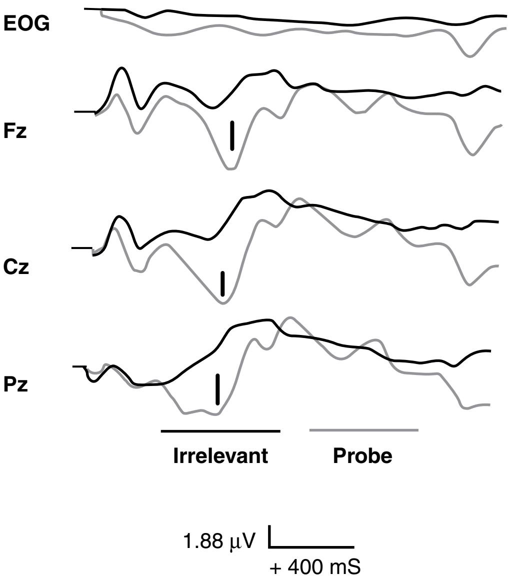


Figure 4.1 Three ERPs and EOG, based on Rosenfeld *et al.* (2004), from the scalp sites Fz (frontal), Cz (central), and Pz (parietal). The sweeps are 2,048 ms long. P300 peaks are down-going and indicated with thick vertical lines. They are in response to personally meaningful items (gray lines). They are superimposed on responses to personally non-meaningful items (black lines). Given that the sweeps are about 2 s long, the P300s begin around 400 ms and end around 900 ms. Positive is plotted down.

History of P300 used as a concealed information detector

Fabiani *et al.* (1983) showed that if a list of words, consisting of rare, previously learned (i.e., *meaningful*) and frequent novel words were presented one at a time to a subject, the familiar, previously learned words but not the others elicited a P300. Rosenfeld *et al.* (1987) recognized that the Fabiani *et al.* (1983) study suggested that P300 could be used to detect concealed guilty knowledge. Therefore, P300 could index recognition of familiar items even if subjects denied recognizing them. From this fact, one could infer deception. The P300 would not represent a lie per se but only a recognition of a familiar item of information, the verbal denial of which would then imply deception.

Soon after seeing Fabiani *et al.* (1983), we executed a study (Rosenfeld *et al.*, 1988) in which subjects pretended to steal one of ten items from a box. Later, the items' names were repeatedly presented to the subject, one at a time on a display. Based on visual inspection of the P300s, we found that the items the subjects pretended to steal (the *probes*), but not the other, *irrelevant* items, evoked P300 in nine of ten cases. In that study, there was also one special, unpredictably presented stimulus item, the *target*, to which the subjects were required to respond by saying "yes," so as to assure us they were paying attention to the screen at all times and would thus not miss probe presentations. They said "no" to all the other items, signaling non-recognition, and thus lying on trials containing the pretended stolen items. The special target items also evoked P300, as one might expect, since they too were rare and meaningful (task-relevant). This paradigm had many features of the guilty knowledge test (GKT) paradigm (developed by Lykken in 1959; see Lykken, 1998), except that P300s rather than autonomic variables were used as the indices of recognition.

Donchin and Farwell also saw the potential for detecting concealed information with P300 as a recognition index in the later 1980s, and they presented a preliminary report of their work (in poster format) at the 1986 Society for Psychophysiological Research (SPR) meeting (Farwell and Donchin, 1986), just after our 1987 paper was submitted. This conference abstract summarized Experiment 2 of the paper later published as Farwell and Donchin (1991). This study reported two experiments, the first of which was a full-length study using twenty paid volunteers in a mock crime scenario. The second experiment contained only four subjects admittedly guilty of minor transgressions. In both experiments, subjects saw three kinds of stimuli, quite comparable to those used in our Rosenfeld *et al.* (1988) study, noted above: (1) *probe* stimuli which were items of guilty knowledge that only "perpetrators" and others familiar with the crime (experimenters) would have; (2) *irrelevant* items which were unrelated to the "crime" but were of the same category as the probe; (3) *target* items which were unrelated to the "crime," but to which the subject was instructed to execute a unique response. Thus, subjects were instructed to press a yes-button to the targets, and a no-button to all other stimuli.

The subjects in this first experiment had participated in a mock crime espionage scenario in which briefcases were passed to confederates in operations that had particular names. The details of these activities generated six categories of stimuli, one example of which would be the name of the mock espionage operation. For each such category, the actual probe operation name might be operation "donkey." Various other animal names – tiger, cow, etc. – would comprise the rest of the set of six stimuli including the probe, four irrelevants and one target name. The six (categories) with six stimuli per category yielded thirty-six items that were randomly shuffled and presented twice per block. After each block, the stimuli were re-shuffled into a new random order and re-presented for a total of four blocks. The mock crime was committed one day before the P300 GKT. It is very important to note that prior to the P300 GKT and prior to performance of the mock crime scenario, each subject was trained and tested on the details of the mock crime in which he/she participated. The training was to a 100 percent correct criterion. Therefore the experimenters could be quite certain that these details would be remembered. Subjects were also trained to know the targets. Subjects were also run as their own innocent controls by being tested on scenarios of which they had no knowledge.

Farwell and Donchin (1991) reported that in the twenty guilty cases, correct decisions were possible in all but two cases which could not be unambiguously classified (as either guilty or innocent) and so were put in an "indeterminate" category. Indeed, this would be impressive except that, as just noted, the subjects were pre-trained to remember the details of their crimes, a procedure having limited ecological validity in field circumstances – in which training of a suspect on details of a crime he was denying is clearly impossible. In the innocent condition, only seventeen of twenty subjects were correctly classified yielding an overall detection rate of 87.5 percent with 12.5 percent "indeterminate" outcomes. Thus although the procedure of Farwell and Donchin (1991) did not have traditional false positive nor false negative outcomes, with accurate verdicts for all the classified cases, their procedure left 12.5% of the cases unclassified.

The second experiment of Farwell and Donchin (1991) had only four subjects. These four volunteering subjects were all previously admitted wrongdoers on the college campus. Their crime details were well-detected with P300, but these previously admitted wrongdoers probably had much rehearsal of their crimes at the hands of campus investigators, teachers, parents, etc. Thus – was the P300 test detecting incidentally acquired information or previously admitted, well-rehearsed information?

A very important contribution of the Farwell and Donchin (1991) paper was the introduction of *bootstrapping* in P300-based deception detection. This was a technique that allowed an accurate diagnosis within each individual. In the earlier Rosenfeld *et al.* (1987, 1988) papers, t-tests comparing individual probe and irrelevant averages were performed. That is, the t-test examined the significance of the difference between probe and irrelevant P300 means. We did not report the results of these t-tests, which afforded low diagnostic rates (<80 percent correct), and did not correspond with what our visual inspection of the waveforms showed. Now one realizes that since the database for such t-tests consists of *single trial* ERPs – which are typically very noisy – the t-tests may miss all but the largest intra-individual effects. Farwell and Donchin (1991) had appreciated that most analyses in ERP psychophysiology were based on group effects in which the grand average of the individual *averages* were compared between conditions. Thus, the database for these tests were *average* ERPs, rather than single sweeps. Farwell and Donchin appreciated also that to do such a test within an individual required multiple probe and irrelevant averages within that individual. These were not usually available since obtaining them would have required running an individual through multiple runs which would have doubtless led to confounding habituation effects, as well as loss of irrelevance of originally irrelevant stimuli which would become relevant via repetition. Bootstrapping was the answer: a *bootstrapped* distribution of *probe* averages could be obtained by repeatedly sampling *with replacement* from the original set of, say, N1 *probe* single sweeps. After each sample is drawn, it can be averaged, so that if one iterated the process 100 times, one would have a set of 100 bootstrapped average probe ERPs. The same procedure could be done with N2 *irrelevant* single sweeps. Then one would have distributions of 100 irrelevant and 100 probe *averages*. A t-test on these cleaner averages would be much more sensitive than such a test on single sweeps. (One usually doesn't need more than 100 iterations, and fifty might do well. N1 and N2 should usually be not much less than twenty-five in my experience, and as suggested by Polich, 1999; Fabiani *et al.*, 2000.)

In fact, once one has distributions of bootstrapped probe and irrelevant averages (which approach the respective actual average ERPs in the limit as developed by Efron, 1979), there are many possibilities for analysis: Farwell and Donchin (1991) reasoned that one ought to statistically compare two cross-correlation coefficients; the cross-correlation of (a) probe and irrelevant P300s with the cross-correlation of (b) probe and target P300s. The idea was that if the subject was guilty, there would be a large P300 in both target and probe ERPs, but not in irrelevant ERPs, so that correlation (b) would be greater than correlation (a). On the other hand, if the subject was innocent, then there would be no P300 in the probe ERP, so that the greater correlation would be (a). If results of ninety of 100 correlation subtractions (b-a) were > 0, then guilt could be inferred.

This method, however, has problems as pointed out by Rosenfeld *et al.* (1991, 2004, 2008) and demonstrated in Rosenfeld *et al.* (2004), even though the method had great success in the Farwell and Donchin

(1991) paper, noted above as having low external validity. One issue that poses a problem for this approach is that although probes and targets may both have P300s in guilty subjects, these waveforms may be out of phase and/or show other latency/morphology differences (as we illustrated in Figure 2 of Rosenfeld *et al.*, 2004). After all, although target P300s were treated as benchmark P300 waveforms by Farwell and Donchin (1991), in fact the psychological reactions to personally meaningful and concealed guilty knowledge probes vs. explicitly task-relevant but inherently neutral targets should differ for many reasons which could account for various morphology differences in the respective P300s. Our view of target stimuli, in summary, is that they are useful attention holders, but not good benchmark waveform producers for probe P300s.

Another problem with the cross-correlation comparison concerns the expectation (Farwell and Donchin, 1991) that the probe-irrelevant correlation will be lower than the probe-target correlation in a guilty party. Actually, in a guilty subject, irrelevant ERPs may contain small P300s as can be seen in Farwell and Donchin (1991), Allen *et al.* (1992) or Rosenfeld *et al.* (1991, 2004). The probe and irrelevant P300 *amplitudes* will differ, but the *shapes* may not, meaning that the Pearson correlation coefficient will scale away the probe and irrelevant amplitude differences, leaving two waveforms that have a *high* correlation. Farwell and Donchin applied a method (called "double centering") designed to correct this problem. The correction computes the grand average waveform (of all probes, irrelevants, and targets) and subtracts it from each probe, irrelevant, and target waveform prior to computation of cross-correlations. The method will be effective in making the probe-irrelevant correlation negative and the probe-target correlation positive if the probe and target P300 amplitudes are about the same size, with both larger than the irrelevant P300, and *if all three waveforms are in phase*. Obviously, this will make the probe-target correlation greater than the probe-irrelevant correlation. However, in cases in which probe and target are more than about 45 degrees out of phase (implying a P300 peak latency difference of 65 or more ms), then this double-centering correction begins to fail. (This observation is based on informal, unpublished simulations by John Allen, and the present author.) Thus I recommend the analysis method described next.

In our studies with bootstrap analysis (Rosenfeld *et al.*, 1991; Johnson and Rosenfeld, 1992; Rosenfeld *et al.* 2004, 2008), in order to avoid the problems associated with correlation comparison just noted, we utilized the simple probe-irrelevant *P300 amplitude differences*, rather than comparative cross-correlations. Thus our approach was to simply develop a distribution of difference values for bootstrapped average probe minus average irrelevant P300s (see Rosenfeld *et al.*, 2004, 2008). Each iterated computation of a bootstrapped probe and irrelevant average lead to a bootstrapped P300 difference calculation. If these bootstrapped differences were > 0 in 90 of 100 iterations, guilt was inferred. (Although this 0.9 criterion has been traditional, it is somewhat arbitrary.) I will re-visit the criterion issue below in the discussion of Meixner and Rosenfeld (2009d). Note that in computing the P300 value for each iterated average, the P300 maximum value is sought within a search window of about 300–700 ms post stimulus. Thus peak latency or phase differences cannot become problematic, as we compute the peak values wherever they fall within the search window. Most recently (Lui and Rosenfeld, 2008; Meixner and Rosenfeld, 2009b), instead of performing the difference computation iterations, we simply applied a t-test to the probe-irrelevant P300 bootstrapped average values.

I briefly cited Rosenfeld *et al.* (1991) above to make a point about bootstrapping. It is worth describing this paper in a bit more detail as part of the early history of P300-based deception detection, inasmuch as it was the first attempt to use P300 methods in diagnosing deception in a scenario that partly modeled the *employee screening Control Question Test*. This is a test that used to be the major application (in terms of number of tests given per year) of all protocols in field detection of deception in the US – until employee screening tests were banned by the US Congress in the federal Employee Polygraph Protection Act of 1988. (There were exceptions to this ban, and security agencies – CIA, NSA, FBI, etc. – still use these tests.)

By way of background, there are two protocols in use in psychophysiological detection of deception (PDD): the *Comparison Question Test* (CQT, formerly known as the *Control* Question Test) and the *Concealed Information Test* (CIT, formerly known as the *Guilty Knowledge Test* or GKT). The two protocols have been the subject of much bitter contention in academic and professional arenas of deception detection. The CQT, preferred by professional "polygraphers" in North America and elsewhere, involves "Did you do it?" type questions, e.g., "Did you take that $5000?" or "Did you kill your wife?" etc. This test is preferred by polygraph professionals because it is relatively easy to compose and apply in various situations, and because it tends to elicit confessions. On the other hand, this procedure is largely rejected by the PDD scientific research community as being unscientific (see Ben-Shakhar, 2002).

Actually our 1991 P300-based procedure contained elements of both CIT and CQT protocols. It was like a CQT in that it probed not about specific crime details, as in the usual CIT, but about past antisocial and illegal *acts* in which a test subject may have been involved. However the structure of the test was clearly that of a CIT: a subject entered the lab and was shown a list of thirteen antisocial or illegal acts aimed at our student-subject population (e.g., plagiarizing a paper, using a false ID for bar service, cheating on a test, smoking pot monthly, etc.). In a situation of perceived privacy (but in fact in a room whose ceiling contained a concealed video camera that relayed subject responses to the experimenter in an adjoining room), the subject checked boxes next to each act of which (s)he had been guilty within the past five years. This made it possible to construct a test containing just one probe item of which the subject was guilty, plus eight other items (including a target and seven irrelevants). We needed this tight control (that would not likely be possible in real field situations) for this first "proof of concept" study of a P300-based CQT-CIT hybrid screening analog. Before testing subjects for P300 responsiveness to probes vs. other stimuli, we ran a bogus recording session and then chose four items to accuse the subjects of having done: "Based on the previous [bogus] run, we think you did A, but you might also have done B, C, or D." The probe item was always either in the B or C position in that quotation, with an innocent, other item in the remaining (of B or C) position. A, B, C, and D were all items perceived by our subject pool to have an equal probability of occurrence in that subject pool. In each subject, we always compared P300s to items B and C in a bootstrap test, referring to these probes and irrelevants, respectively, as "relevants" and "controls" in the language of the CQT. We accurately identified about 87 percent of the thirteen guilty (12/13) and fifteen innocent (13/15) subjects in the study. However, it should be noted that there was a confound in this study: the subjects may have produced P300s to the items they checked "yes" to prior to the recording session because they checked "yes," rather than because these were the remembered guilty items. Clearly, one couldn't run the list checking manipulation prior to testing in the field. Thus, in a near replication study one year later (Johnson and Rosenfeld, 1992), we used one item for all subjects as the probe *prior* to the recording session. This was an item we knew from previous study would yield us about 50 percent guilty and 50 percent innocent subjects. Our diagnostic rate replicated. (We confirmed the "ground truth" by running the list checking session *after* the recording session.) Recently we (Lui and Rosenfeld, 2008) utilized these methods – enhanced with spatial-temporal principal component analysis – with subjects guilty of two and three probes (in two groups), detecting 86 percent and 71 percent of guilty subjects, respectively, although with about 30 percent false positives, yielding Grier (1971) A' AUC values of 0.87 and 0.76 for two and three probe groups, respectively.

Among the early P300 studies, one must also note the study by Allen *et al.*, (1992). This study was somewhat different than those reviewed previously in that it examined detection of newly acquired information, learned to perfection, which is often not as well detected as well rehearsed (self-referring) information (Rosenfeld *et al.*, 2006), but which was well detected in Allen *et al.* (1992), possibly because of the highly original Bayesian analysis they developed to detect concealed information within individuals. Thus, over three subject samples, 94 percent of the learned material was correctly classified, and 4 percent of the unlearned material was incorrectly classified.

It should be added here that various methods of individual diagnosis have been compared by Allen and Iacono (1997), Rosenfeld *et al.* (2004), Abootalebi *et al.* (2006) who also introduced an original wavelet classifier method, and by Mertens and Allen (2008). Allen and Iacono (1997) compared Bayesian analysis, bootstrapped cross-correlations, and bootstrapped amplitude differences applied to the data of Allen *et al.* (1992). They found no difference in the effectiveness of the first two methods but found both to be superior to the bootstrapped amplitude difference method. However Allen and Iacono (reporting an overall accuracy of 87 percent) utilized the *baseline-to-peak* index of P300 amplitude (in their amplitude difference computations) which we never use since we and others (such as Meijer *et al.*, 2007) found it to be at least 25 percent less accurate than our *peak-to-peak* method. In the Abootalebi *et al.* (2006) paper, the ROC curves displayed for the wavelet classifier, bootstrapped cross-correlation, and bootstrapped *peak-to-peak* amplitude methods show considerable overlap, although small differences can be seen favoring either the bootstrapped amplitude difference method (e.g., Rosenfeld *et al.*, 2008) or the wavelet classifier method depending upon the location in the curve in ROC space. The bootstrapped cross-correlation method of Farwell and Donchin (1991) performed consistently worst, although the differences among the three methods were small. (The three methods correctly detected 74 percent to 80 percent of the subjects in a mock crime protocol. However it is difficult to compare accuracy levels obtained in different studies since protocols and thresholds of classification differ.) Rosenfeld *et al.* (2004) *consistently* found that the peak-to-peak amplitude difference method outperformed the cross-correlation method. This study was the only one in which comparisons were made on two stimulus sets, one involving autobiographical data, and the other involving mock crime details, neither stimulus set being pre-learned to perfection. By contrast, Mertens and Allen (2008) found that the bootstrapped peak-to-peak amplitudes performed worse than either the cross-correlation or Bayesian approaches. However this was demonstrated in *only* some comparisons involving countermeasure groups, and when innocent subjects were considered, the bootstrapped peak-to-peak amplitude difference method performed better than the cross-correlation method. All these results from the four comparative studies reviewed together suggest that no one method is ubiquitously superior. However, comparisons were indeed difficult within the Mertens and Allen (2008) study because their correlation approach uniquely used an indeterminate category, and bootstrap criteria differed between their two bootstrap methods. Moreover, only the Mertens and Allen (2008) study reported poor rates of accuracy overall (48 percent at *best*) in the virtual reality environment which only they used. They also used the highest number of irrelevant stimuli (ten) deployed by any of these studies, which no doubt made for a uniquely demanding task, possibly accounting for their low accuracy. More systematic work of this type is certainly in order. Although it appears presently that when the tested material is learned to perfection prior to testing, all methods work equally well, other situations that lead to greater P300 latency variation, perhaps related to uncertainty of stimulus recognition, favor the peak-to peak amplitude difference method. This tentative conclusion needs further verification.

All the protocols used in previous sections may be denoted as *3-stimulus protocols* (3SPs), in that they all present subjects on a given trial with either a probe, an irrelevant stimulus, or a target stimulus (requiring a unique response) in the same temporal position on each trial. During the past two decades, there were multiple applications of the original 3SPs in which the type of information to be detected varied according to the anticipated needs of various agencies. There were also various technical questions addressed that concerned P300 measurement and analysis in the CIT context. Our lab became interested in detecting simulated malingering in modeled head injury populations. We were concerned also with utilizing a different dependent measure related to P300, namely the distribution of amplitudes across the scalp. This foray was reviewed previously (Rosenfeld, 2002) so I will say no more about it now beyond the fact that although P300 *scalp distribution* seemed to distinguish malingering and non-malingering as a robust group effect, the individual detection rates never exceeded 75 percent, which is not very impressive. Our work indicated, however, that P300 *amplitude* did consistently well in identifying memory malingerers (Rosenfeld *et al.*, 2002). Moreover, in an extensive series of related papers of high quality, Van Hooff and colleagues have pursued the use of P300 in memory assessment with very positive results (e.g., Van Hooff *et al.*, 1996; Van Hooff and Golden, 2002; Van Hooff *et al.*, 2009).

In a different line of original research, Lefebvre *et al.* (2007) applied the 3SP to the measurement of eyewitness identification accuracy in a clever model of identification of culprits in a simulated police lineup. Subjects observed a mock crime on a videotape and then were tested on their abilities to identify culprits as opposed to bystanders and/or other lineup members after varying time delays between crime and test up to a week. The P300s elicited by probe (culprit) faces confirmed recognition of correct faces. The authors concluded, "P300 provided a reliable index of recognition of the culprit relative to the other lineup members across all time delay conditions. Although participants' accuracy decreased at the 1-week time delay compared to no delay and the 1-h time delay, the P300 effect remained strong for participants that made correct identifications irrespective of the time delay." This novel fact that face stimuli could be used to elicit P300 as an index of concealed *pictorial* information was replicated by Meijer *et al.* (2007).

Regarding the technical developments in measuring and analyzing P300 from 1992–2004, the papers by Allen and Iacono (1997), as well as by Abootalebi *et al.* (2006), in which various analytic methods were compared, have already been discussed. It remains to detail that in my experience, the best method of measuring P300 *solely for purposes of detecting concealed information* is to measure it from its positive peak to its subsequent negative peak using filter settings of 0.3Hz to 30Hz. We discussed why this is so in Soskins *et al.* (2001). One major reason is that we always detect at least 25 percent more guilty subjects with no additional false positives using this peak-peak (p-p) method than we do using the standard baseline-to-peak method. This is based on at least ten of our studies. We have had independent confirmation recently from Meijer *et al.* (2007). I want to make clear that *I am not advocating this p-p method for any other uses of P300, especially regarding theoretical questions*, since the p-p method may measure more than pure P300 as noted in Soskins *et al.* (2001).

Other technical issues came up recently regarding approaches to P300 bootstrap analysis, and regarding the confidence interval criteria used in these tests. They arose in the context of our later described, novel protocol for P300-based detection of concealed information (Rosenfeld *et al.*, 2008). This protocol was devised to deal with the serious issue of countermeasures (CMs) to P300-based detection of concealed information in the 3SP. It is best to delay this discussion until the CM issue is covered next.

Countermeasures to P300 as a concealed recognition index

Many eminent people assumed for many years that the P300 CIT (in its original 3SP format) would be unbeatable because the stimuli were presented so rapidly (every 2–4 s) and responded to by the brain so quickly (300–700 ms) that subjects would have no way to utilize CMs. Lykken (1998, p. 293) put it this way, "Because such potentials are derived from brain signals that occur only a few hundred milliseconds 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) similarly wrote, "ERP measures seem to be immune against countermeasures because they are based on a repeated rapid presentation of the items (e.g., one item per second). When items are presented at such a rapid pace, it is virtually impossible to execute countermeasures to the control items." Our eminent colleague, Donchin, has repeatedly expressed this view to me in email correspondence even after publication of Rosenfeld *et al.* (2004), to be reviewed below, and after its approximate replication and extension by Mertens and Allen (2008). Our original 2004 demonstration of CMs to the 3SP arose from some simple reflections about that 3SP.

The instructions to subjects in the 3SP are to press a target button when targets are presented, and an alternative, non-target button on all other trials, both irrelevant and probe. (Verbal responses such as "Yes" for target or "No" for non-target may be substituted.) It is expected that rare probes will evoke a P300 because even though they are not explicitly task-relevant, their crime-related or personal significance makes them meaningful and salient only to guilty subjects. Targets will also evoke a P300 because they *are* explicitly task-relevant. This is why Farwell and Donchin (1991) expected probe and target P300s to look alike and, therefore, cross-correlate in guilty subjects.

It occurred to us that if simply making a unique, *overt* response to an irrelevant, *experimenter-designated* target stimulus could endow it with P300-eliciting properties, it also ought to be possible for subjects to learn to instruct themselves to make unique *covert* responses to *self-designated* irrelevant stimuli. When these formerly irrelevant stimuli become covert, relevant targets, they too should evoke P300s, making their averages indistinguishable from probe P300 averages. Once the probe-irrelevant difference is lost, the 3SP should no longer work, since now the probe-irrelevant correlation should be not appreciably different than the probe-target correlation.

All this is what we showed in Rosenfeld *et al.* (2004), utilizing either a *multiple probe protocol* (as in Farwell and Donchin, 1991, and as described in Rosenfeld *et al.*, 2007) or one of our own one-probe protocols, and utilizing either bootstrapped cross-correlation differences (an 82 percent hit rate was reduced to an 18 percent hit rate with CMs) or bootstrapped simple amplitude differences (a 92 percent hit rate was reduced to 50 percent with CMs). Mertens and Allen (2008) used a somewhat different scenario involving simulated mock crimes represented in virtual reality software, but showed that similarly conceived CMs dramatically reduce hit rates obtained without CMs.

The Rosenfeld *et al.* (2004) report was critically reviewed by Iacono in an archival volume (Iacono, 2007), but some of the critical points were inaccurate and/or misleading. For example, "the classification hit rate difference between guilty and countermeasure subjects was not statistically significant." The large numerical differences were actually given above, and in fact varied in significance from $p < 0.05$ to $p < 0.08$ across two studies. It would have been more accurate to state that the significance varied from marginally to *actually statistically significant*.

Iacono continued his critique by emphasizing that "*no test should be accepted as valid if the irrelevant and target stimuli cannot be easily differentiated*." In fact, a glance at Figure 1 of Rosenfeld *et al.* (2004) makes it clear that this italicized phrase, *does not actually apply to our 2004 study* since that figure shows in the CM group that *the target P300s tower over the irrelevant (and probe) P300s* [my italics]. Indeed, the superimposed P300s for the three stimuli in the CM group greatly resemble the neighboring comparable superimpositions from the innocent group in the same figure and that is precisely the idea of the CM strategy – to make the CM-using guilty subject look innocent!

Iacono also casts doubt about the "salience of [our] probes" because they "did not elicit responses as large as the targets." In fact in the Farwell and Donchin (1991) first experiment, the task-relevant target P300s are clearly larger than probe P300s – as expected – in eighteen of twenty cases. The data of Rosenfeld *et al.* (1991) are similar. It appears to be an individual matter as to which stimulus, probe, or target will be more salient for a particular subject. It is also worth noting that although the Rosenfeld *et al.* (2004) paper contained two studies, all Iacono's criticisms refer only to the first. There were other matters, but overall we appreciated that Iacono did conclude positively about our "important" CM study that it was "the first to explore how countermeasures might affect this type of GKT."

Thus, it became apparent to us that the next major and long overdue challenge for P300-based information detection was to come up with a CM-resistant protocol. As similarly stated in Rosenfeld *et al.* (2008), to increase CM resistance of the P300-based CIT, we attempted to identify factors in the older P300 protocols that potentially compromised the test's sensitivity. The most obvious factor seemed to be the combination of the explicit target–nontarget decision with the implicit probe-irrelevant discrimination, both of which occur in response to the sole stimulus presented in each trial of the original 3ST protocol. That is, the subject's explicit task in each trial of the 3ST is to decide whether or not the stimulus is a target. However, it is also expected that the inherent salience of a probe stimulus (due to its personal or crime relevance) would nevertheless lead to an enhanced P300 as the target–nontarget discrimination was made. This meant that processing resources would have to be divided between the explicit target task and the implicit probe recognition. We reasoned that, because diversion of resources away from an oddball task by a second task reduces the oddball evoked P300 (Donchin *et al.*, 1986), likewise the probe P300 may be reduced by a concurrent target discrimination task. Thus we developed a novel protocol in which the probe-irrelevant discrimination would be separated from a time-delayed target–nontarget discrimination. We referred to the new protocol accordingly developed as the *complex trial protocol* or CTP.

A novel CM-resistant protocol

In the CTP, each trial begins with the presentation of either a rare ($p = 0.2$) probe or a frequent ($p = 0.8$) irrelevant (stimulus 1 or S1) and the subject is instructed to respond as soon as possible on a single response box with a single button press (Response 1 or R1) no matter whether probe or irrelevant is presented. This is called the "I saw it" response because the response simply signals the operator that the stimulus was perceived regardless of its type. Then, after a random quiet interval of about 1.2 to 1.8 s, a second stimulus (S2), either a target or non-target, appears and the subject must give a specific differential response (R2) to signal target or non-target. (Recently, we have used number strings for target and non-targets.) The protocol is called complex because there are two separated tasks (S1/R1 and S2/R2) on each trial. The S1/R1 task allows us to compare probe to irrelevant P300s. The target task, though delayed, maintains attention and helps us enforce task compliance. (See Figure 4.2.)

The protocol was the most accurate we have ever reported for the detection of self-referring (birth date) stimuli. With no CMs (a "simple guilty" or SG condition) we detected 12/12 subjects with a flexible P300 search window; or 11/12 with one fixed search window for all subjects. Using the same kind of CMs as in Rosenfeld *et al.* (2004), in which

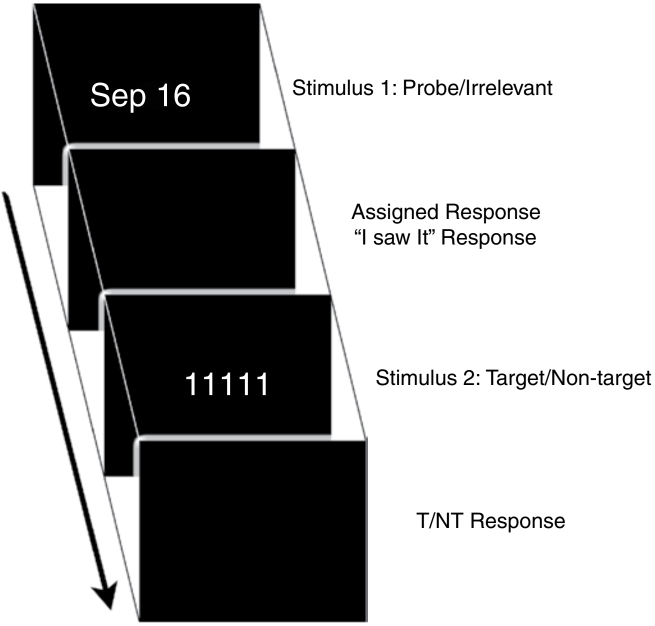


Figure 4.2 The events in the complex trial protocol using a date for S1 and a number string for S2.

subjects made a unique, assigned CM to each of the four irrelevants, we detected 11/12 subjects (flexible search window) or 10/12 (fixed search window). In a replication study in the same report, we detected 12/12 subjects in both SG and CM conditions with fixed search windows.

The diagnoses are based on bootstrap tests in which the probe P300 (peak-to-peak) was compared against the average of all irrelevant P300s (P vs. Iall test) with a p = 0.9 confidence interval. We reported further that (1) the CM condition caused increased probe and irrelevant (see Figure 4.3) P300 waves in each subject in both replications (an effect which we also saw in Winograd and Rosenfeld, 2008, who extended the CTP protocol for use in the detection of incidental details of a mock crime), and (2) reaction time (RT) was dramatically increased to probes and more so to irrelevants when CMs were used, so much that just within the CM run, irrelevant RTs were larger than probe RTs.

Regarding the CM condition's increasing effect on probe P300s, we suggested that the CM task forced greater attention to S1 as subjects needed to decide on each trial whether or not to execute a CM. This increased attention, we suggested, led to enhanced probe P300s. This effect may have been occurring but we also saw possibilities for still

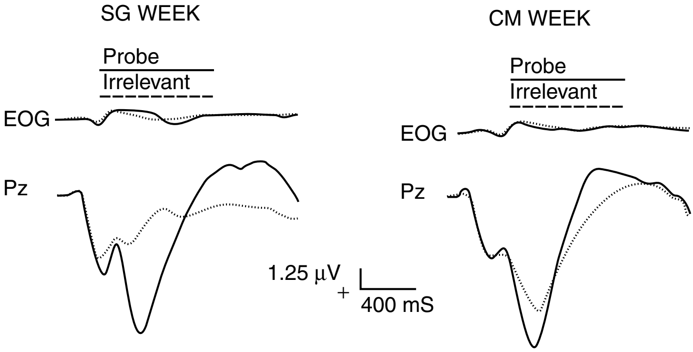


Figure 4.3 These are the Pz and EOG superimposed probe and irrelevant ERPs from the no countermeasure or simple guilty (SG) and countermeasure (CM) conditions of Rosenfeld *et al.* (2008) which were run over two successive weeks with the same subjects. The large down-going deflections are P300s. The enhanced irrelevant P300 in the CM week results from CM efforts (positive is plotted down).

another effect: we had reasoned in Rosenfeld *et al.* (2008) that since subjects made unique CM responses to each of four irrelevant stimuli, but of course, not to the probe, the latter could become the sole stimulus not requiring the added CM response, which would add salience to the probe beyond its inherent personal meaningfulness, thereby increasing its P300 amplitude. Meixner and Rosenfeld (2009b) demonstrated this *omit effect* rather unequivocally by showing that when a particular irrelevant stimulus, "IO," is presented in a series among four other irrelevant stimuli, each of which requires a specific assigned CM-like response, but not IO (Omitted Irrelevant), IO alone evokes a P300. If a meaningful probe-like item is substituted for IO, then the P300 virtually doubles in size. If CM-like responses are assigned to *all* five stimuli (omit effect removed), then the probe-like item elicits a P300 of about the same size as that evoked by IO.

It seemed to us at the time of the 2008 paper that having subjects generate a unique CM to each irrelevant stimulus was the best way to beat the test, since all irrelevants would become P300-eliciting covert targets, as in Rosenfeld *et al.* (2004). After finding the *omit effect*, we appreciated that from the guilty perpetrator's perspective, the best CM strategy might be to attempt to counter only a subset of irrelevant stimuli. Thus, Rosenfeld and Labkovsky (in press) conducted a new experiment with a modified version of the CTP in which only two of four irrelevants were countered. We also decided in this study to make the CMs purely mental, in an attempt to beat our recently observed RT effect of CM use (which were observed mostly with physical CMs in Rosenfeld *et al.* 2004, 2008). Subjects said silently to themselves either their first or last names upon seeing specifically one or the other irrelevant stimuli to be countered. The CTP was modified in such a way that a random five-button response box was used for R1. That is, the subjects said "I saw it" by pressing one of five randomly chosen buttons. The reasons for this novel R1 are given in Rosenfeld and Labkovsky (in press).

The results of this study were that the enormous increase in probe P300 associated with CM use seen in Rosenfeld *et al.* (2008) evaporated, a finding further supporting the *omit* hypothesis. Nevertheless we still detected 100 percent of SG subjects and CM users using the P vs. Iall test at p = 0.9 confidence. We have now run a total of five new studies in which one, two, three, four, and five CMs (the last included CMs against probe also) were used against four irrelevants in five groups of twelve subjects each. Hit rates in these groups were always 11/12 (92 percent) or 12/12 (100 percent) in P vs. Iall bootstraps at p = 0.9 (Labkovsky and Rosenfeld, 2009).

A promising alternative approach to a CM-resistant protocol was introduced by Lui and Rosenfeld (2009). We simply presented the key stimuli subliminally, the conceptualization being that stimuli not consciously perceived cannot be countered. The method accurately classified guilty and innocent subjects at 86 percent overall, with ROC-based discrimination of 0.88. This protocol was ERP-based, but could be tried with other dependent measures.

Recent developments in bootstrap tests about probe size

It was noted earlier that one robust finding in both Rosenfeld *et al.* (2008) and Rosenfeld *et al.* (2004) was that the use of CMs to irrelevants led to increased RTs. This proved to be a useful though not unsurprising finding, since it seems obvious in retrospect that if a subject needs to stop and reflect on which assigned CM response is to be used when a given irrelevant is presented, time will pass. Actually, the probe-irrelevant discrimination appears to take longer also as RT for probes is also elevated by CM use, though not as much as for irrelevants, in which case the extra cognitive step of CM response selection must occur.

The test we and others have mostly used in making decisions about a subject's guilt asked if the probe P300 was greater than an average of P300s to all irrelevant stimuli, the P vs. Iall test. The result of this test establishes that that probe P300 is greater than the irrelevant P300 average, but does not establish that the probe P300 is larger than each and every irrelevant P300, which one may want to know in some situations (e.g., in the "search CIT," when examiners are unaware of the critical items). Such a test requires that the probe P300 be greater than the largest irrelevant P300 at some level of confidence. This is the P vs. Imax test, and is obviously extremely rigorous, and could acceptably require use of a confidence level below 0.9. The confidence level for any test is acceptable so long as it is not lowered to a point at which one begins to see appreciable numbers of false positives in innocent control subjects. Thus the ROC curve from signal detection theory becomes useful as it relates the hit rate to the false positive rate, and it becomes possible to specify a confidence level for bootstrap testing in a manner such that the area under the ROC curve must exceed a certain amount, say, 0.9. This issue will be reconsidered when we discuss situations in which one does not know the identity of the probe in advance, such as in cases of suspected terrorist arrests, *prior to* commission of an act of terror (Meixner and Rosenfeld, 2009a).

The extreme rigor and sometimes concomitant loss of sensitivity of the P vs. Imax test can be somewhat tempered if one utilizes the elevated RT effect of CM use as a screen: that is, in both Rosenfeld *et al.* (2008) and Rosenfeld and Labkovsky (in press), we utilized the elevating CM effect on RTs as follows: if the test of P vs. Imax was significant, the subject was simply diagnosed as guilty. (If additionally, the RT for Imax was significantly greater than the probe RT, we noted also the probable use of CMs.) If the test of P vs. Imax failed, but the RT for Imax was greater than the probe RT, CM use was assumed, and the bootstrap test was performed on the probe vs. the *second largest* irrelevant – assuming its RT was not greater than the probe RT – and the result was referred to as an *RT-screened* P vs. Imax test result. If the RT for the second largest irrelevant was greater than that to the probe, one could proceed to the next largest irrelevant, and so on. The rationale for this screening procedure is the assumption (now backed by much data) that if the RT to an irrelevant stimulus is significantly elevated relative to the probe, one can assume a CM was used for that irrelevant, and there is no point testing a probe P300 against an irrelevant P300 likely to have been enlarged by a covert CM response. In both Rosenfeld *et al.* (2008) and Rosenfeld and Labkovsky (in press), the screened P vs. Imax tests yielded 92 percent (11/12) detection of CM-users.

In the RT screening procedure, we have described a comparison of the probe RT with irrelevant RTs, *all collected within one block*. It

Table 4.1 *Probe and Iall RTs (ms) in baseline and experimental blocks in SG, CM, and innocent (IC) groups.*

Group	Base-P	Base-Iall	Exp-P	Exp-Iall
SG	401.8	390.8	392.8	393.4
CM	483.9	460.3	806.5	1196.9
IC	469.1	442.2	393.7	397.3

is often the case that this procedure reveals CM use, particularly if a subject executes even a purely mental CM response *discretely*, *prior to* the "I saw it" response. Very recently, we have trained a novel batch of special CM users who execute CMs (two for four irrelevants) *simultaneously* with their "I saw it" responses, (Sokolovsky *et al.*, 2009). This "lumping" strategy reduces the within-session RT differences between probes and countered irrelevants (although guilt is detected in 83 percent of the CM-users with P300 amplitude anyway). However no matter what type of CM strategies subjects use, the RT average in a CM session is always at least 150 ms larger than in non-CM sessions. Much larger SG vs. CM differences in RT were found in Rosenfeld *et al.* (2008) and in Winograd and Rosenfeld (2008) in which all four irrelevants were countered. This suggests that the irrelevant RTs in a test session could be compared to RT data from some kind of baseline/control session so as to index CM use. This is what we did in Winograd and Rosenfeld (2008). In this study, prior to the mock crime execution, subjects were given a baseline RT task in the same format as the CTP except they were tested on knowledge of selected playing cards. In Table 4.1 are the RTs in the baseline and post-mock crime experimental sessions for probe (P) and irrelevant stimuli (all averaged together = Iall).

It is evident that in the CM experimental session, probe RTs are elevated by about 400 ms and Iall RTs by 800 ms in comparison to all other RTs. These dramatic increases can obviously be used as CM indexes.

It is appreciated that there is the possibility that a subject planning to use CMs will simply intentionally delay RTs in the *baseline* test. However, using upper limits for RTs will prevent such efforts from being successful (Ratcliff and McKoon, 1981; Seymour *et al.*, 2000). Moreover, there is no reason why a normative distribution of RTs in a situation with no motivation for CMs cannot be collected. Such "norms" could be used to test suspected CM-inflated RTs.

What if you lack advance knowledge of "ground truth"? Are there attendant analytic questions?

Since September 11, 2001, interest in deception detection in general increased, and in particular, interest especially developed in situations in which suspected terrorists might be captured and questioned prior to commission of a specific terrorist act. Such a situation poses new challenges to P300-based detection, even using the putatively powerful new CTP protocol. For example, one may arrest a suspected terrorist, but one doesn't necessarily know where, when, or how he plans to strike. However, intelligence may provide a few ideas about reasonable and probable answers to these questions, so that one can construct lists of plausible item sets for each category of information in which one is interested, for example, a set of US cities likely to be attacked. This is not necessarily a simple matter, and would likely be based on extensive prior investigation, including analysis of the "chatter" monitored in terrorist networks by authorities, as well as on results of interrogation of other suspects in custody, and so on. However one then is faced with the question of identifying which is the probe item to be used in tests of whether or not it elicits the largest P300 among a set of such stimuli.

John Meixner in our lab recently (the data are still being analyzed) undertook to model this situation (Meixner and Rosenfeld, 2009a). A subject in a guilty (SG) group (n = 12) was given a briefing document we prepared explaining that he was to play the role of a terrorist agent and plan a mock terrorist attack on the United States. The document detailed several different possible options he could choose regarding how to carry out the attack. The subject then read detailed descriptions of four types of bombs that could be used, four locations in the city of Houston that could be attacked, and four dates in July when the attack could take place. The descriptions contained pros and cons of each potential choice and instructed subjects to choose one type of bomb, one location in Houston, and one date on which to attack. After reading the briefing document, the subject was instructed to compose a letter to the fictitious superior in the terrorist organization describing the choices made. Subjects in the innocent (IN) group (n = 12) completed a similar task, but planned a vacation instead of a terrorist attack. Then after electrode attachment, a subject completed three separate blocks of the CTP task, with each block testing for a separate concealed information item. Subjects were shown potential cities where the terrorist attack could occur (Houston was the correct item), potential types of terrorist attacks (with Bomb as the correct item) and potential months the attack could occur in (with July as the correct item).

The data for each block were analyzed in three ways: in one way the correct item was considered the probe and its P300 was tested against the average P300 of five other irrelevant items in each block for this study (the P vs. Iall test). The second analysis (P vs. Imax) tested the known probe P300 against the maximum irrelevant (without RT screening, as we have not yet analyzed RT data). *Finally, we did an analysis for situations in which ground truth was lacking* (Allen *et al.*, 1992, did something similar as demanded by their Bayesian approach to the question determining the probability that a word was from a learned list, given its evocation of a P300). We simply assumed that if the subject was concealing information concerning one item of the six tested in each block, it would evoke the largest P300, so we tested the largest P300 (the hypothesized probe P300) against the next largest P300 (the "Blind" Imax test; we assumed this second largest P300 to be the largest evoked by an irrelevant item). This was actually a conservative test since we might have tested the largest P300 against the average of *all* the remaining P300s for the other stimuli. Such a test, however, might have had a cost in specificity. We used 1,000 bootstrapped iterations for each block, then combined data from three blocks and averaged across blocks to yield the following table of results (see Table 4.2).

The numbers under the guilty and innocent designations show the three-block average number (maximum = 1,000) of bootstrap iterations in which the bootstrapped average probe or hypothetical probe (for blind Imax) tested as greater than the average of other P300s as designated. Each of the twelve rows represents a subject in each column for guilty and innocent groups. Means are shown in third row from bottom. Guilty diagnostic fractions are shown in second row from bottom, and the respective areas under ROC curves (AUC) are shown in the bottom row. It is apparent that we obtained perfect guilty-innocent discrimination in P vs. Iall and P vs. Imax tests, and excellent discrimination (AUC = 0.979) in the blind tests.

It should be added that in the above experiment, we were again using the CTP approach with incidentally acquired, newly learned information, which we have previously shown (Rosenfeld *et al.*, 2006, 2007) to be not well detected with the 3SP. It is furthermore to be emphasized that in this study, subjects studied their newly acquired information for only thirty minutes, whereas it is likely that a real terrorist would have repeatedly rehearsed the details of a planned terrorist act to greater levels of processing depth. Thus it is quite possible that the signal to noise ratio in probe vs. irrelevant stimulus comparisons was probably less than in other situations we have worked with, and likely less than in field situations. In this connection, we note that although for the first P

Table 4.2 *The number (maximum = 1,000) of bootstrap iterations in which the bootstrapped average (hypothetical) probe was greater than that of the irrelevant items (Iall/Imax)*

P vs. Iall		**P vs. Imax**		**Blind Imax**	
Guilty	**Innocent**	**Guilty**	**Innocent**	**Guilty**	**Innocent**
1,000	648	985	287	985	603
1,000	610	999	416	998	602
955	598	889	476	892	649
996	611	898	430	893	605
994	150	946	17	943	689
909	475	698	284	761	547
945	600	677	365	702	536
997	555	959	250	961	569
999	586	908	217	907	565
985	690	888	382	886	706
912	390	667	129	698	650
903	644	837	215	842	702
966	**546**	**863**	**289**	**872**	**619**
12/12	**0/12**	**12/12**	**0/12**	**10/12**	**0/12**
AUC = 1.0		**AUC = 1.0**		**AUC = 0.979**	

vs. Iall test, we were able to use a bootstrap confidence interval of 0.9, as in previous studies, for the unscreened and rigorous P vs. Imax test, we had to drop our confidence interval to any value from 0.5 to 0.65 in order to achieve an AUC = 1.0. There is nothing inherently wrong with this adjustment since it is seen by examining the means of the third and fourth columns of the above table, that the numbers of positive iterations for *both* probe and Imax values are well below those seen under the P vs. Iall columns. Similarly, for the blind Imax test, we used a 0.75 confidence interval to get the best discrimination.

A guilty decision in the above study was based on totals for three blocks of data. It is certainly also of interest to know how many (of three possible in each subject) details of the planned terrorist act could be discerned. For that datum, one needs to know how many *individual blocks* led to positive outcomes on bootstrap tests. Using a confidence interval of 0.9, with no a priori specification of the probe, we were able to correctly identify twenty-one of thirty possible terrorist act details in the ten of twelve subjects correctly identified in blind Imax tests. The CTP appears to hold promise for the anti-terrorist challenge. (Other CIT-based attempts to deal with this challenge in very different ways were reported by Lui and Rosenfeld, 2009, reviewed briefly above, and Meijer *et al.*, in press.)

Conclusions

It is clear that whatever promise any of the P300-based protocols hold for real-world application, there are problems yet to be solved, despite the progress of the past two decades. The CTP is a new method that appears to show more CM resistance than any of its predecessors, but it too needs further research. Rosenfeld and Labkovsky (in press) observed for the first time a possibly novel ERP component called "P900" (with a latency of about 900 ms) that is maximal at Fz and Cz. It is seen only in countermeasure users in probe ERPs and sometimes in non-countered irrelevant ERPs, but not in countered irrelevant ERPs. This may prove a useful CM index in situations that may yet be seen in which RT is not a useful CM index. For example, the new "lumping" CM noted above (in which subjects make CM and "I saw it" responses simultaneously; Sokolovsky *et al.*, 2009) seems to pose problems for use of a probe-irrelevant, within-session RT difference to detect CM use. Clearly we need to fully document the phenomenology of and better understand P900 before its profitable application. Obviously, the application of the CTP to anti-terror situations needs much more work; in particular, the effect of CMs needs documentation in the anti-terror protocol described above. Finally, and this is true for all deception detection protocols, certainly not excepting the CTP, the effect of time passage between crime (or crime planning) and testing is not well known, but is clearly critical. There have been some preliminary efforts which are promising (e.g., Hamamoto *et al.*, 2009; Lefebvre *et al.*, 2007) but there is much work needing to be done on this crucial variable.

REFERENCES

Abootalebi, V., Moradi, M. H., and Khalilzadeh, M. A. (2006). A comparison of methods for ERP assessment in a P300-based GKT. *International Journal of Psychophysiology*, 62, 309–320.

Allen, J. J. B., and Iacono, W. G. (1997). A comparison of methods for the analysis of event-related potentials in deception detection. *Psychophysiology*, 34, 234–240.

Allen, J. J. B., Iacono, W. G., and 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, 504–522.

Ben-Shakhar, G. (2002). A critical review of the Control Questions Test (CQT). In M. Kleiner (ed.), *Handbook of Polygraph Testing* (pp. 103–126). San Diego: Academic Press.

Ben-Shakhar, G., and Elaad, E. (2002). The Guilty Knowledge Test (GKT) as an application of psychophysiology: future prospects and obstacles.

In M. Kleiner (ed.), *Handbook of Polygraph Testing* (pp. 87–102). San Diego: Academic Press.

Donchin, E., Kramer, A., and Wickens, C. (1986). Applications of brain event related potentials to problems in engineering psychology. In M. Coles, S. Porges, and E. Donchin (eds.), *Psychophysiology: Systems, Processes and Applications* (pp. 702–710). New York: Guilford.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.

Fabiani, M., Gratton, G., and Coles, M. G. H. (2000). Event-related brain potentials: methods, theory, and applications. In J. T. Cacioppo, L. G. Tassinary, and G. Berntsen (eds.), *Handbook of Psychophysiology* (pp. 85–119). New York: Cambridge University Press.

Fabiani, M., Karis, D., and Donchin, E., (1983). P300 and memory: individual differences in the von Restorff effect. *Psychophysiology*, 558 (abstract).

Farwell, L. A., and Donchin, E. (1986) The brain detector: P300 in the detection of deception. *Psychophysiology*, 24, S34 (abstract).

Farwell, L. A ., and Donchin, E. (1991). The truth will out: interrogative polygraphy ("lie detection") with event-related potentials. *Psychophysiology*, 28, 531–547.

Hamamoto, Y., Hira, S., and Furumitsu, I. (2009) Effects of refreshing memory on P300-based GKT administered one month after a mock crime for repeated offenders. Poster presented at 49th Ann. Meeting, Society for Psychophysiolological Research. Berlin.

Iacono, W. G. (2007). Detection of deception. In J. T. Cacioppo, L. G. Tassinary, and G. Berntsen (eds.), *Handbook of Psychophysiology* (pp. 668–703). New York: Cambridge University Press.

Johnson, M. M., and Rosenfeld, J. P. (1992). Oddball-evoked P300-based method of deception detection in the laboratory II: utilization of nonselective activation of relevant knowledge. *International Journal of Psychophysiology*, 12, 289–306.

Labkovsky, E. B., and Rosenfeld, J. P. (2009) Accuracy of the P300-based complex trial protocol for detection of deception as a function of number of countered irrelevant stimuli. Poster presented at 49th Ann. Meeting, Society for Psychophysiolological Research. Berlin.

Lefebvre, C. D., Marchand, Y., Smith, S. M., and Connolly, J. F. (2007) Determining eyewitness identification accuracy using event-related brain potentials (ERPs). *Psychophysiology*, 44, 894–904.

Lui, M. and 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, 721–730.

Lui, M. and Rosenfeld, J. P. (2009). The application of subliminal priming in lie detection: scenario for identification of members of a terrorist ring. *Psychophysiology*, 46, 889–903.

Lykken, D. T. (1998). *A Tremor in the Blood: Uses and Abuses of the Lie Detector*, 2nd edn. New York: Plenum Press.

Meijer, E. H., Smulders, F. T., and Merckelbach, H. L. (in press). Extracting concealed information from groups. *Journal of Forensic Sciences*.

Meijer, E. H., Smulders, F. T., Merckelbach, H. L., and Wolf, A. G. (2007). The P300 is sensitive to concealed face recognition. *International Journal of Psychophysiology*, 66, 231–237.

Meixner, J. B., and Rosenfeld, J. P. (2009a). Identifying terrorist information using the P300 ERP component. Poster presented at 49th Ann. Meeting, Society for Psychophysiolological Research. Berlin.

Meixner, J. B., and Rosenfeld, J. P. (2009b). Countermeasure mechanisms in a P300-based concealed information test. *Psychophysiology*, 47, 57–65.

Mertens, R., and Allen, J. J. (2008). The role of psychophysiology in forensic assessments: deception detection, ERPs, and virtual reality mock crime scenarios. *Psychophysiology*, 45, 286–298.

Polich, J. (1999). P300 in clinical applications. In E. Niedermeyer and F. Lopes da Silva (eds.), *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, 4th edn. (pp. 1073–1091). Baltimore and Munich: Urban & Schwarzenberg.

Ratcliff, R., and McKoon, G. (1981). Automatic and strategic priming in recognition. *Journal of Verbal Learning and Verbal Behavior*, 20, 204–215.

Rosenfeld, J. P. (2002). Event-related potentials in the detection of deception, malingering, and false memories. In M. Kleiner (ed.), *Handbook of Polygraph Testing* (pp. 265–286). New York: Academic Press.

Rosenfeld, J. P., and Labkovsky, E. (in press) 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*.

Rosenfeld, J. P., Biroschak, J. R., and Furedy, J. J. (2006). P300-based detection of concealed autobiographical versus incidentally acquired information in target and non-target paradigms. *International Journal of Psychophysiology*, 60, 251–259.

Rosenfeld, J. P., Shue, E., and Singer, E. (2007). Single versus multiple probe blocks of P300-based concealed information tests for autobiographical versus incidentally learned information. *Biological Psychology*, 74, 396–404.

Rosenfeld, J. P., Angell, A., Johnson, M., and Qian, J. H. (1991). An ERP based, control-question lie detector analog: algorithms for discriminating effects within individuals' average wave forms. *Psychophysiology*, 32, 319–335.

Rosenfeld, J. P., Rao, A., Soskins, M., and Miller, A. R. (2002). P300 scalp distribution as an index of deception: control for task demand. *Journal of Credibility Assessment and Witness Psychology*, 3(1), 1–22.

Rosenfeld, J. P., Soskins, M., Bosh, G., and Ryan, A. (2004). Simple effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, 41, 205–219.

Rosenfeld, J. P., Nasman, V. T., Whalen, I., Cantwell, B., and Mazzeri, L. (1987). Late vertex positivity in event-related potentials as a guilty knowledge indicator: a new method of lie detection. *International Journal of Neuroscience*, 34, 125–129.

Rosenfeld, J. P., Cantwell, G., Nasman, V. T., Wojdac, V., Ivanov, S., and Mazzeri, L. (1988). A modified, event-related potential-based guilty knowledge test. *International Journal of Neuroscience*, 24, 157–161.

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., and Chedid, E. (2008). The Complex Trial Protocol (CTP): a new, countermeasure-resistant, accurate P300-based method for detection of concealed information. *Psychophysiology*, 45, 906–919.

Seymour, T. L., Seifert, C. M., Mosmann, A. M., and Shafto, M. G. (2000). Using response time measures to assess "guilty knowledge." *Journal of Applied Psychology*, 85, 30–37.

Sokolovsky, A. W., Rothenberg, J., Meixner, J. B., and Rosenfeld, J. P . (2009). Sequential versus simultaneous stimulus acknowledgement and countermeasure responses in P300-based detection of deception. Poster presented at 49th Ann. Meeting, Society for Psychophysiolological Research. Berlin.

Soskins, M., Rosenfeld, J. P., and Niendam, T. (2001). The case for peak-to-peak measurement of P300 recorded at .3 Hz high pass filter settings in detection of deception. *International Journal of Psychophysiology*, 40, 173–180.

Van Hooff, J. C., and Golden, S. (2002). Validation of an event-related potential memory assessment procedure: intentional learning as opposed to simple repletion. *International Journal of Psychophysiology*, 16, 12–22.

Van Hooff, J. C., Brunia, C. H. M., and Allen, J. J. B. (1996). Event-related potentials as indirect measures of recognition memory. *International Journal of Psychophysiology*, 21, 15–31.

Van Hooff, J. C., Sargeant, E., Foster, J. K., and Schmand, B. A. (2009). Identifying deliberate attempts to fake memory impairment through the combined use of reaction time and event-related potential measures. *International Journal of Psychophysiology*. Online April 15.

Winograd, M. R., and Rosenfeld, J. P. (2008). Mock crime application of the complex trial protocol P300-based concealed information test. *Psychophysiology*, 45, S62 (abstract).