DOI 10.1007/s10484-009-9091-4

Assigned Versus Random, Countermeasure-Like Responses in the P300 Based Complex Trial Protocol for Detection of Deception: Task Demand Effects

John B. Meixner · Alexander Haynes · Michael R. Winograd · Jordan Brown · J. Peter Rosenfeld

Published online: 20 June 2009

Abstract We recently introduced an accurate and countermeasure resistant P300-based deception detection test called the complex trial protocol (Rosenfeld et al. in Psychophysiology 45(6):906–919, 2008). When subjects use countermeasures to all irrelevant items in the test, the probe P300 is increased rather than reduced (as it was in previous P300-based deception protocols), allowing detection of countermeasure users. The current experiment examines the role of task demand on the complex trial protocol by forcing the subject to make countermeasure-like response to stimuli. Subjects made either a simple random button response to both probe and irrelevant stimuli (experiment 1) or a more complex, assigned, button response to probe and irrelevant stimuli (experiment 2). We found that an increase in task demand reduced the effectiveness of the test. Using random responses we found a simple guilty hit rate of 11/12 with no false positives, but only a 4/11 hit rate for countermeasure-users. Using assigned responses we found a simple guilty hit rate of 8/15 with no false positives, and a 7/16 hit rate for countermeasure-users. We herein suggest that the high level of task demand associated with these countermeasure-like responses causes reduced hit rates.

Keyword Task demand effects · ERP · P300 · Deception detection · CIT · GKT

J. B. Meixner (✉) · A. Haynes · M. R. Winograd · J. Brown · J. Peter Rosenfeld
Department of Psychology, Northwestern University,
2021 Sheridan Road, Evanston, IL 60208-2700, USA
e-mail: jmeixner@northwestern.edu

Introduction

In the past 20 years, the conventional control question test (CQT) technique for the detection of deception has come under much criticism (National Research Council 2003; Ben-Shakhar 2002; Lykken 1981). A more promising and scientifically sound method, the Concealed Information Test (CIT, also known as the Guilty Knowledge Test), was developed by Lykken (1959, 1960) for use with the polygraph. The CIT presents subjects with various stimuli, one of which is a concealed information item (such as the gun used to commit a crime). The other stimuli in the test consist of control items that are of the same class (such as other potentially deadly weapons: a knife, a bat, etc.) such that an innocent person would be unable to discriminate them from the concealed information item. If the subject's physiological response is greater for the concealed information item (as compared to the control items), then knowledge of the crime or other event is inferred.

The CIT has since been adapted to detect guilty knowledge using event related potentials (ERPs), specifically focusing on the P300 component (Rosenfeld et al. 1988; Farwell and Donchin 1991; Allen et al. 1992). P300 is known to be largest in amplitude in response to infrequently presented, personally meaningful items (Sutton et al. 1965; Donchin and Coles 1988, Johnson 1988). In the most familiar P300-based CIT protocol (hereafter referred to as the "Three-stimulus protocol"), subjects typically view test items of three types: the probe, which is the guilty knowledge item; the irrelevant, which is of the same class as the probe but with no relevance to the crime in question; and the target, which is an irrelevant item to which the subject must make a unique response to ensure that he/she is paying attention to the stimuli (Rosenfeld et al. 1988; Farwell and Donchin 1991; Allen et al. 1992).

Three stimulus protocols have yielded accuracy rates as high as 95% (Rosenfeld et al. 1988; Farwell and Donchin 1991; Allen et al. 1992) but these accuracy rates have been reduced to 50% or less when confronted with simple countermeasures (Rosenfeld et al. 2004; Mertens and Allen 2008). Countermeasures (CMs) are discrete responses that one makes to the irrelevant items, turning them into covert targets and thus enlarging their P300 amplitude. Because the critical comparison in the P300 based CIT is between the probe item and the irrelevant items, detection accuracy decreases as irrelevant P300 amplitude increases.

Rosenfeld et al. (2008) described a novel, CM resistant P300-based CIT called the *complex trial protocol*, which divides each trial into a first phase containing a single probe or irrelevant stimulus, followed by a second phase containing a single target or nontarget stimulus (see Fig. 1 for an example). The rationale behind this division is that during a single trial the subject's attention will no longer be divided between the implicit probe/irrelevant recognition task and the explicit target/nontarget discrimination task because the probe/irrelevant discrimination and the target/nontarget decision tasks are separated. This elimination of the competing target/nontarget task theoretically increases P300 amplitude to the probe (Donchin et al. 1986). Using the complex trial protocol, Rosenfeld et al. (2008) reported 100% detection accuracy with guilty subjects as well as 92% detection accuracy with CM-users. Additionally, Rosenfeld et al. (2008) found that P300 amplitude to the probe was larger in the countermeasure condition than in the simple guilty condition.

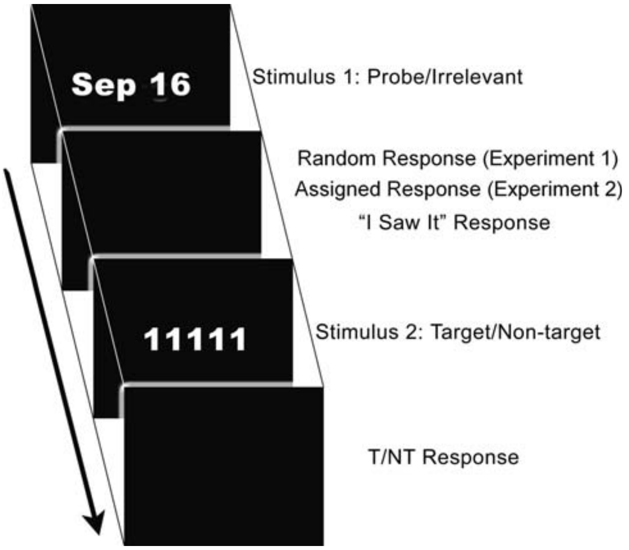
Sep 16
Stimulus 1: Probe/Irrelevant
Random Response (Experiment 1)
Assigned Response (Experiment 2)
"I Saw It" Response
11111
Stimulus 2: Target/Non-target
T/NT Response

Fig. 1 Structure of each trial. On each trial, subjects view 2 stimuli: one date (*probe or irrelevant*) and one string of numbers (*target or nontarget*). Using their left hand, Subjects press a random response button based on the date seen, followed by the "I saw it" button for all dates. When the string of numbers appear, subjects use their right hand to press the right mouse button if the string is all ones (*target*), and the left mouse button if the string is a series of any other numbers (*nontarget*)

Research by Donchin et al. (1986) showed that while an unrelated and competing task that is conducted simultaneously to a P300 eliciting task will reduce P300 amplitude, simultaneously performing a task that is highly related to a P300 eliciting task can increase P300 amplitude during the primary task. Donchin et al. referred to this as *embedding* of the secondary, related task within the primary task. Rosenfeld et al. (2008) postulated that the high accuracy rate for CM-users is because the removal of the target/nontarget discrimination task from the first phase of the trial causes CMs to be embedded in probe/irrelevant recognition task, thereby increasing the P300 amplitude of probe items (Donchin et al. 1986).

In the current study, we used a countermeasure-like embedded task to focus attention on the first stimulus of a CTP-style CIT. Subjects were instructed to perform overt countermeasure-like responses to probe and irrelevant stimuli. While subjects in Rosenfeld et al. (2008) responded to the probe and irrelevant stimuli with a single "I saw it" button press, subjects in the current study performed a somewhat more difficult task intended to force more attention to the first stimulus. We hypothesized that because the countermeasure-like task is embedded within the probe/irrelevant recognition task, P300 amplitude to the probe will be increased, thereby increasing detection accuracy. In two experiments, two tasks with different levels of difficulty were tested: one simple task with random countermeasure like responses (experiment 1) and one difficult task with assigned, countermeasure-like responses (experiment 2).

Method

Experiment 1

Subjects

Thirty seven students (average age: 19 years; 13 males) at Northwestern University were recruited for the study. Subjects gave written informed consent to participate. Subjects received introductory psychology course credit for their participation. All subjects had normal or corrected vision. The experiment was approved by the Northwestern Institutional Review Board.

Trial Structure

Trial structure was modeled after Rosenfeld et al. (2008). Each trial began with a 100 ms baseline period of empty black screen during which prestimulus EEG was recorded. Next, a date was presented in white text on a black background for 300 ms. Dates were presented in the form of MONTH, DAY, with the first three letters of the month used (e.g. Apr 12, Jan 23). Upon seeing the stimulus, subjects were instructed to press a random response button using the left hand (see Fig. 1). Random responses were made using a five-button response box where subjects placed each digit of the left hand on one of the buttons. The purpose of this random button response task was to increase attention to the first stimulus. Subjects were instructed to monitor their responses and be careful that they were not responding in a pattern. Following the random response, subjects were instructed to press the left mouse button with the right hand. Because this response indicates that the subject has seen the stimulus, regardless of whether he saw a probe or an irrelevant item, it is termed the “I saw it” button. Subjects were instructed to make the random response with the left hand prior to pressing the “I saw it” button.

After a 1,500 ms interval in which the subjects viewed a black screen, a string of six identical numbers ranging from 1 to 5 (i.e. 111111, 222222, etc.) was presented for 300 ms. Subjects were instructed to press the left mouse button with the index finger of the right hand when they saw a string of ones (the target), and the right mouse button with the middle finger of the right hand when they saw a string of any other numbers (nontargets). All stimuli were shown in white font 0.7 cm high on a monitor ~70 cm in front of the subject.

Procedure

After signing the consent form, subjects were seated in a comfortable chair and given written instructions for a practice task. The practice task was similar to the full task as described above, but included no random response to the first stimulus, which, for practice, was a name rather than a date (e.g. John, Cindy). Subjects were instructed to immediately press the “I saw it” button when they saw a name. Following the name, subjects completed a target/nontarget recognition task as described above in the detailed trial structure. Subjects practiced the task until they felt comfortable and made no errors. Following the practice tasks, subjects were given written instructions for the full task. Subjects read these instructions and asked questions as the experimenter was applying electrodes. Subjects were questioned to ensure that none of the irrelevant dates had any confounding personal relevance. Subjects then practiced the full task with all responses included until they felt comfortable to continue (typically 10–15 trials).

Subjects then completed 300–350 trials of the task (depending on the subject’s blink rate). The task lasted ~30 min. The task was paused each 50–60 trials at which point the subject was asked what the previous date was to ensure that he/she was paying attention. Prior to the run, the subject was alerted that missing more than one of these check-ups would result in test failure. Subjects were given three to four 30 s rest breaks, spaced evenly throughout the task.

Table 1 Stimulus probabilities

Stimulus type	Number	Probability
Probe target	33	0.09
Probe nontarget	33	0.09
Irrelevant target	33	0.09
Irrelevant nontarget	250	0.72
All probes	66	0.19
All irrelevants	283	0.81

Note: Probe target ratio = .50; Irrelevant target ratio = .11. A probe target trial is one in which a target follows a probe. An irrelevant target trial is one in which a target follows an irrelevant

The ratio of probe to irrelevant trials was 1:4, as shown in Table 1. It is noted that probe targets and probe non-targets have equal probabilities whereas irrelevant targets are much less frequent than irrelevant non-targets. This discrepancy could lead to a confound if the probability of a target following a probe being greater than that of a target following an irrelevant increased the salience of the probe item. This issue was examined in Rosenfeld et al. (2008) using an innocent control group in which the “probe” item was just another irrelevant item. If the asymmetry of conditional target probabilities caused an increase in salience of the probe, false positive outcomes would result. Rosenfeld et al. (2008) found 0–8% false positives; no more than in previous studies without this asymmetric probability matrix. Additionally, submitted data from our lab (Rosenfeld et al. 2009, in press) have shown that a nearly identical protocol (in which the only difference is the removal of this asymmetry) shows no difference in P300 amplitude or detection rates in comparison with the asymmetric probability protocol.

Subjects were randomly assigned to one of three groups:

1. *Simple Guilty*. Subjects in the *simple guilty* (SG) group ($n = 12$) were shown four irrelevant dates (irrelevants) and their respective birthdate (probe).
2. *Innocent*. Subjects in the *innocent* (IN) group ($n = 12$) were shown five irrelevant dates.
3. *Countermeasure*. Subjects in the *countermeasure* (CM) group ($n = 11$) were shown four irrelevant dates (irrelevants) and the respective birthdate (probe), as in the *simple guily* group. Subjects in the CM group were instructed to attempt to beat the P300 CIT by making covert responses to enhance the salience of two of the irrelevant items. After practicing the full task without CMs, subjects in the CM group were given an additional

set of instructions that specify two irrelevant dates that they were to counter. Subjects were told to silently say their first name to themselves when they saw one of the dates, and to silently say their last name to themselves when they saw the other date. Subjects were instructed to make these responses before the random button press and "I saw it" button press.

Experiment 2

Subjects

Forty six students (average age: 19 years; 22 males) at Northwestern University were recruited for the study. Subjects gave written informed consent to participate. Subjects received introductory psychology course credit for their participation. All subjects had normal or corrected vision. The experiment was approved by the Northwestern Institutional Review Board.

Trial Structure

The trial structure of experiment 2 was identical to that of experiment 1 except subjects made specific assigned responses to all stimuli rather than random responses. Using the same left hand 5 button box used in experiment 1, subjects pressed either the index or middle finger button to all stimuli based on response assignments that subjects were given prior to the experiment. The two earlier dates were assigned to the middle finger, and the three latter dates were assigned to the index finger. The purpose of this assigned button response task was to force more attention to the first stimulus, compared with a task lacking the stimulus classification requirement (such as experiment 1). Following the assigned response, subjects pressed the "I saw it" button and completed the target/nontarget task just as in experiment 1.

Procedure

The procedure for experiment 2 was identical to that of experiment 1, with the assigned button response replacing the random button response. Experiment 2 had 15 subjects in the simple guilty group, 15 subjects in the innocent group, and 16 subjects in the countermeasure group.

Data Acquisition

EEG was recorded using Ag/AgCl electrodes attached to midline sites Fz, Cz, and Pz. Scalp electrodes were referenced to linked mastoids. Electrode impedances were held below 10 kΩ. EOG was recorded differentially via Ag/AgCl electrodes placed above and below the left eye. EOG electrodes were placed diagonally to allow for the recording of both vertical and horizontal eye movements as well as eye blinks. Artifact rejection criteria varied based on each subject's artifact amplitudes, always less than 50uv. Trials for which this threshold was exceeded were removed from both the ERP and reaction time analyses. Two subjects with fewer than 25 non-artifacted trials per stimulus were removed from the final analysis. The forehead was connected to the chassis of the isolated side of the amplifier system ("ground"). Signals were passed through Grass P511 K amplifiers with a 30 Hz low pass filter setting, and high pass filters set (3 db) at .3 Hz. Amplifier output was passed through a 16-bit A/D converter sampling at 500 Hz. After initial recording, single sweeps and averages were digitally filtered off-line to remove higher frequencies; 3 db point = 6 Hz.

Analysis Methods

P300 amplitude, our main dependent variable, was measured using the peak–peak method as described in Soskins et al. (2001). We and others have found this analysis method to be more sensitive for the detection of deception than the standard base-peak method used in earlier studies (Soskins et al. 2001; Meijer et al. 2007). Using in-house software designed for the Matlab platform, an algorithm searched a window of 400–650 ms to find the maximally positive segment of 100 ms, with the midpoint of this segment defined as P300 latency and its average amplitude defined as the positive P300 peak. Next, the algorithm searched a window from the P300 latency to 1,300 ms to find the maximally negative segment of 100 ms. The peak–peak amplitude of the P300 was defined as the difference between the positive P300 peak and the negative P300 peak.

Within Individuals Bootstrap Analysis

ANOVAs were applied to the Behavioral and ERP variables to assess the group effects of the study. Because this study relates to the detection of deception, individual diagnostic statistics are also essential. To determine whether the P300 evoked by a given stimulus is greater than that evoked by another stimulus within an individual, the bootstrap method (Wasserman and Bockenholt 1989) was used at the Pz site, where P300 is usually found to be largest (Fabiani et al. 1987). The typical bootstrap test compares the probe P300 to the average P300 of all irrelevant trials to determine whether the true difference between the average probe P300 and average irrelevant P300 is greater than zero (Iall bootstrap). Because the actual distributions of probe and irrelevant waves are not available, they must be bootstrapped from the existing data.

To do this, a computer program draws, with replacement, a set of individual probe waveforms equal to the number of accepted probe trials and also randomly draws (with replacement) an equal number of irrelevant waveforms. The program then subtracts the mean irrelevant P300 from the mean probe P300, and then repeats the process 100 times to create a distribution of bootstrapped probe minus irrelevant averages.

Additionally, a second and more rigorous bootstrap test was conducted, comparing the probe P300 to the largest maximum irrelevant stimulus P300 (Imax bootstrap). This process is identical to the Iall bootstrap method, except irrelevant waveforms were only drawn from trials of the irrelevant item with the largest average P300 amplitude.

Past studies (Rosenfeld et al. 2004, 2008; Farwell and Donchin 1991) have defined a p-value criterion of .1 in order to state that a probe waveform is significantly greater than an irrelevant waveform within an individual subject. Thus, 90% of the distribution must be greater than zero at -1.29 standard deviations from the mean of the distribution, which also means that at least 90 of the 100 iterations of the process described above must yield a positive number. In reporting bootstrap values, we report the number of iterations (out of 100) in which the probe average was greater than the irrelevant average.

Reaction Time Screen

Because the process of performing countermeasures tends to increase irrelevant reaction time during the task compared to the probe (Rosenfeld et al. 2004, 2008), we performed a reaction time screen in an attempt to reduce the effectiveness of countermeasures. We compared the random response RT (or assigned response RT for experiment 2) between the probe and the irrelevant item with the largest P300 amplitude (Imax) using a t-test. If the Imax RT was *significantly* greater than the probe reaction time, we performed the same procedure with next largest irrelevant, until we found the irrelevant item largest in P300 amplitude while not being significantly greater than the probe in RT. This irrelevant item's P300 amplitude was then compared with that of the probe using the Imax bootstrap method as described above. If the original Imax RT was not significantly greater than the probe RT, the screen was reported as not significant, and the original Imax bootstrap test was kept as the final test. Additionally, if all irrelevant reaction times were significantly greater than the probe reaction time, the original Imax bootstrap was kept as the final test (see Table 3). The reaction time screening procedure was conducted on all subjects across all groups. If a subject was detected by the Imax bootstrap test at a .9 confidence, the reaction time screen was not necessary and thus not conducted.

Results

Experiment 1

All within subjects ANOVA p-values reported are Greenhouse-Geisser (GG) corrected if $df > 1$. Partial Eta squared values (η^2) are reported where applicable.

Behavioral: Reaction Times

Figure 2 shows the mean random response reaction times to both probes and the average of all irrelevants (Iall) for each group. Note that CM reaction times are clearly greater than both simple guilty and innocent RTs. Irrelevant reaction times were collapsed over all four irrelevant stimuli as there was no significant difference between RT values for any single irrelevant item in a 1 × 4 ANOVA comparing each irrelevant item; $F(3, 136) = .08$, $p > .9$. A mixed model 2 × 3 ANOVA (Stimulus × Group) revealed a significant main effect of group, $F(2, 32) = 32.4$, $p < .001$, $\eta^2 = .67$, but no effect of stimulus ($p > .6$) and no significant interaction ($p > .3$). Tukey's post hoc tests revealed significant differences between the CM and simple guilty groups ($p < .001$) and between the CM and innocent groups ($p < .001$), indicating significantly slower random response reaction times for the CM group.

Figure 3 shows the mean "I saw it" reaction times to both probe and Iall stimuli. Irrelevant reaction times were collapsed over all four irrelevant stimuli as there was no significant difference between RT values for any single irrelevant item in a 1 × 4 ANOVA comparing each

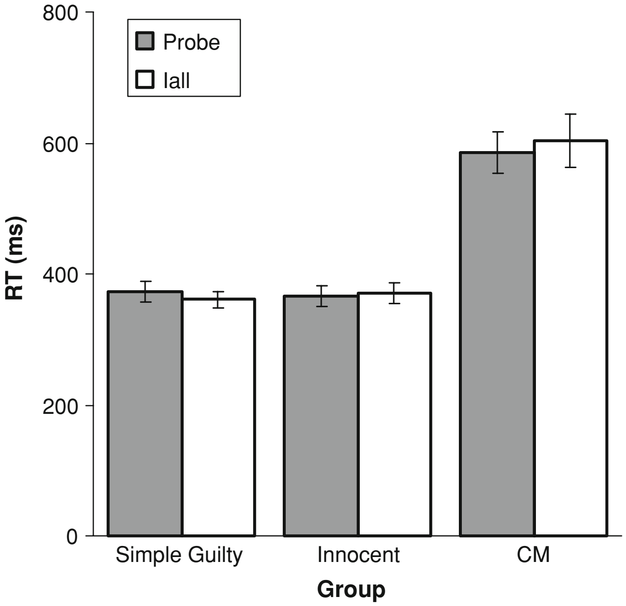

Fig. 2 Experiment 1 random response reaction times to the first stimulus

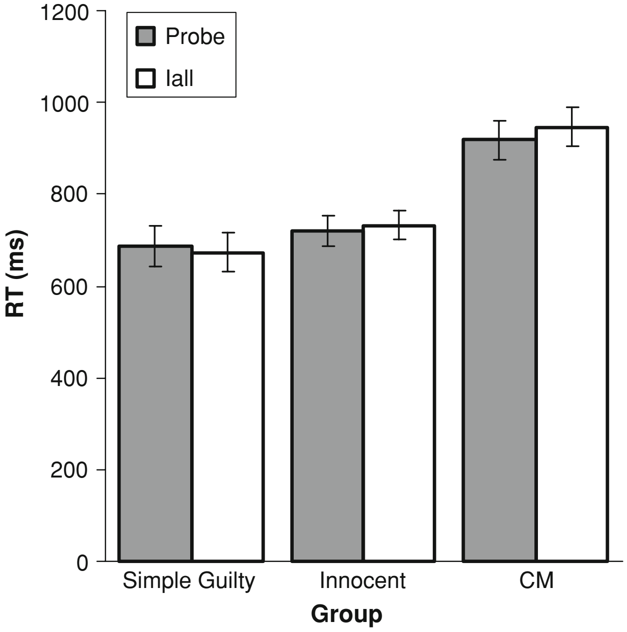

Fig. 3 Experiment 1 "I saw it" response reaction times to the first stimulus

irrelevant item; $F(3, 136) = .08$, $p > .9$. A mixed model 2 × 3 ANOVA (Stimulus × Group) revealed a significant main effect of group, $F(2, 32) = 10.98, p < .001, \eta^2 = .41$, but no main effect of stimulus, $F(1, 32) = 3$, $p > .09$, $\eta^2 = .086$. There was a significant stimulus × group interaction, $F(2, 32) = 6.55$, $p < .005$, $\eta^2 = .29$, with irrelevant RTs slightly greater than probe RTs in the CM and innocent groups, while irrelevant RTs were faster than probe RTs in the simple guilty group. Tukey's post hoc tests revealed significant differences between the CM and simple guilty groups ($p < .001$) and between the CM and innocent groups ($p < .005$), indicating significantly slower "I saw it" reaction times for the CM group.

ERPs: Qualitative

Figure 4 shows grand average waveforms at site Pz for each group. Waveforms are shown for the probe item in each group, as well as for the average of all irrelevant items (Iall) for each group. Grand averages are restricted to the first 1,500 ms of each trial, containing only the P300 response to the first stimulus (probe/irrelevant). Probe P300 amplitude is clearly larger than Iall amplitude in the simple guilty as well as the CM groups (though to a lesser extent), while probe and Iall P300 amplitudes are nearly indistinguishable in the innocent group.

ERPs: Quantitative Group Data

To examine observations about the grand averages quantitatively, a 2 × 3 (stimulus × group) ANOVA was run (see Fig. 5) on the peak–peak P300 amplitudes across groups. There was a main effect of stimulus, $F(1, 32) = 51.1$, $p < .001$, $\eta^2 = .615$ with probe amplitude exceeding Iall amplitude, as well as a trend toward a main effect of group, $F(2, 32) = 3$, $p < .07$, $\eta^2 = .158$ with simple guilty subjects having the largest overall P300 amplitudes, followed by CM subjects, and innocent subjects having the smallest P300 amplitude. The stimulus × group interaction was highly significant, $F(2, 32) = 13.91$, $p < .001$, $\eta^2 = .465$. To decompose this interaction, we subtracted the Iall amplitude from the probe amplitude in each subject to compute the average probe/Iall difference for each subject.

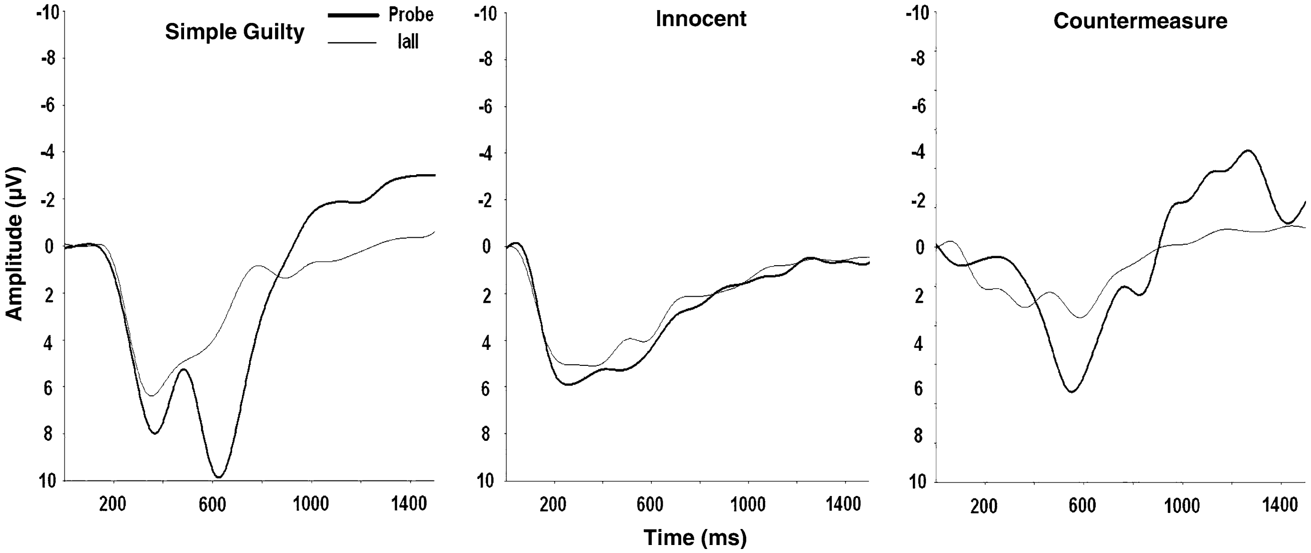

Fig. 4 Experiment 1 grand average probe and irrelevant ERPs at Pz for each group, including 100 ms baseline before stimulus presentation

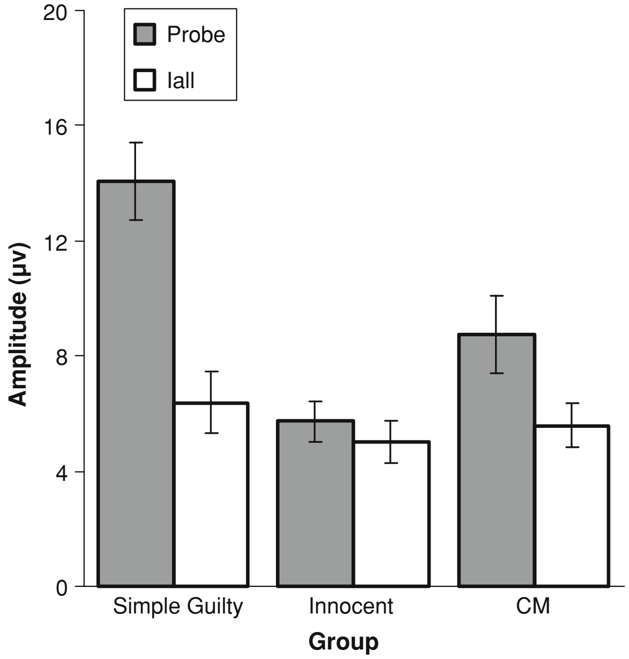

Fig. 5 Experiment 1 average probe and irrelevant peak–peak P300 amplitudes at Pz by group

Table 2 Experiment 1 individual bootstrap detection rates

Iall			Imax			Screened Imax		
SG	Innocent	CM	SG	Innocent	CM	SG	Innocent	CM
100	15	100	85	9	93	NS	NS	–
100	17	76	98	34	55	–	NS	79
100	34	14	100	52	2	–	NS	NS
99	23	100	98	4	83	–	NS	100
100	70	77	99	26	31	–	NS	NS
100	76	76	100	35	1	–	NS	NS
99	99	99	96	46	77	–	NS	NS
100	79	100	100	43	99	–	NS	–
100	98	71	100	85	5	–	NS	NS
98	51	100	96	18	99	–	NS	–
100	85	45	100	64	4	–	NS	NS
100	58		100	0		–	NS	
12/12	**2/12**	**5/11**	**11/12**	**0/12**	**3/11**	**11/12**	**0/12**	**4/11**

Note: Detection rates based on a .9 confidence interval

Numbers indicate the number of iterations of the bootstrap process in which probe was greater than Iall, Imax, or the screened Imax value (depending on column). "NS" indicates that the Imax reaction time was not significantly greater than the probe reaction time. A dash indicates that the Imax value was greater than 90, so the screened test was not performed

A 1 × 3 ANOVA was conducted comparing each group's probe/Iall difference, and yielded a highly significant group effect, $F(2, 32) = 13.91$, $p < .001$. Tukey's post hoc tests revealed that the probe/Iall difference of the simple guilty group was greater than that of the innocent group ($p < .001$) and of the CM group ($p < .01$). There was no significant difference between the innocent and CM groups ($p > .15$), suggesting that CMs were effective.

To examine differences between the sensitivities between methods of experiments 1 and 2, we compared the probe/Iall difference of each group between experiments using an independent samples *t*-test. There was a significant difference between the simple guilty groups, $t(24) = 1.966$, $p < .05$, while there was no difference between the innocent groups ($p > .9$) or the countermeasure groups ($p > .85$).

ERPs: Quantitative Individual Data

Table 2 shows detection rates within subjects for each group at the .9 confidence level, as well as the number of significant iterations out of 100 in the bootstrap distribution for each subject. Bootstrap detection rates are shown for each of the three tests: Iall, Imax, and screened Imax, as described above. The Iall analysis method was the most sensitive, correctly classifying 12/12 SG subjects while yielding two false positives but catching only 5/11 CM subjects. The Imax method provided a more conservative test, catching 11/12 SG subjects with no false positives, and catching 3/11 CM subjects. The addition of the Imax screen allowed the detection of one additional CM subject (4/11 total) and did not change the detection rates of the other groups.

Experiment 2

All within subjects ANOVA *p*-values reported are Greenhouse-Geisser (GG) corrected if $df > 1$. Partial Eta squared values (η^2) are reported where applicable.

Behavioral: Reaction Times

Figure 6 shows the mean assigned response reaction times to both probes and the average of all irrelevants (Iall) for each group. Irrelevant reaction times were collapsed over all four irrelevant stimuli as there was no significant difference between RT values for any single irrelevant item in a 1 × 4 ANOVA comparing each irrelevant item; $F(3, 180) = 2.18$, $p > .1$. A mixed model 2 × 3 ANOVA (Stimulus × Group) revealed a significant main effect of simulus, $F(1, 43) = 10.62$, $p < .005$, $\eta^2 = .2$ with Iall RTs greater than probe RTs, but no main effect of group ($p > .14$). There was a significant stimulus × group interaction, $F(2, 43) = 8.58$, $p < .001$, $\eta^2 = .29$. To examine this interaction, we conducted a paired *t*-test of probe versus Iall RT in each individual group, revealing significantly faster RTs for probe stimuli in the simple guilty group, $t(14) = 3.41$, $p < .005$ and in the countermeasure group, $t(15) = 3.61$, $p < .005$. There was no effect of

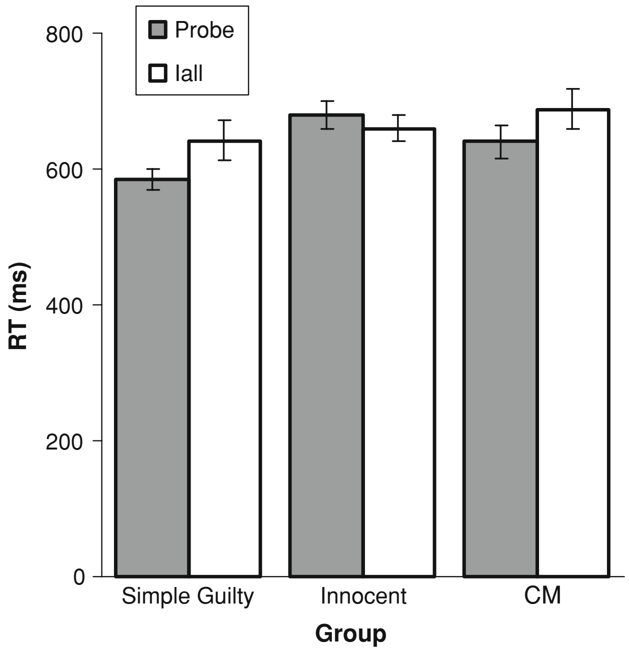

Fig. 6 Experiment 2 assigned response reaction times to the first stimulus

stimulus in the innocent group, $p > .15$. These reaction time results are strikingly different than those found in experiment 1, as discussed below.

Figure 7 shows the mean "I saw it" reaction times to both probe and Iall stimuli. Irrelevant reaction times were collapsed over all four irrelevant stimuli as there was no significant difference between RT values for any single irrelevant item in a 1 × 4 ANOVA comparing each irrelevant item; $F(3, 180) = 1.72, p > .16$. A mixed model 2 × 3 ANOVA (Stimulus × Group) revealed a significant main effect of group, $F(2, 43) = 4.26$, $p < .02$, $\eta^2 = .165$, with innocent RTs greater than simple guilty and CM RTs. There was also a main effect of stimulus, $F(1, 43) = 7.17, p < .01$, $\eta^2 = .14$, with irrelevant reaction times greater than probe reaction times. There was a significant stimulus × group interaction, $F(2, 43) = 8.41$, $p < .001$, $\eta^2 = .281$. Tukey's post hoc tests revealed significant differences between the CM and innocent groups ($p < .05$), indicating significantly slower "I saw it" reaction times for the innocent group. To further examine the interaction, we conducted a paired *t*-test of probe versus Iall RT in each individual group, revealing significantly faster RTs for probe stimuli in the simple guilty group, $t(14) = 3.76$, $p < .005$ and in the countermeasure group, $t(15) = 2.4, p < .05$. There was no effect of stimulus in the innocent group, $p > .05$.

ERPs: Qualitative

Figure 8 shows grand average waveforms at site Pz for each group. Waveforms are shown for the probe item in

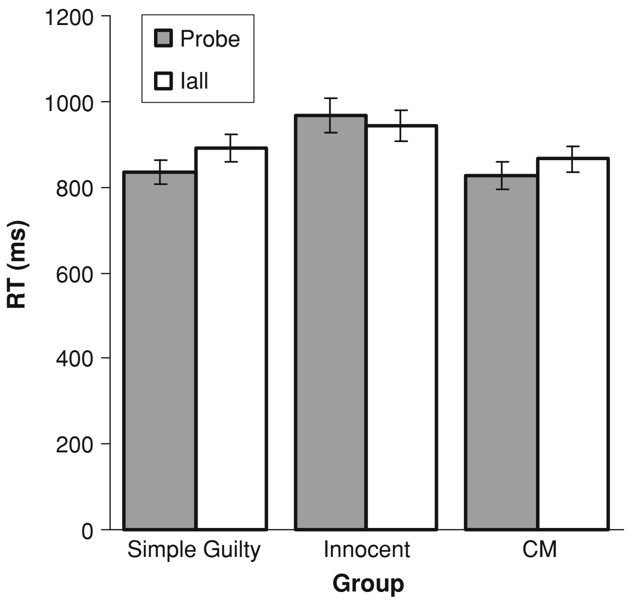

Fig. 7 Experiment 2 "I saw it" response reaction times to the first stimulus

each group, as well as for the average of all irrelevant items (Iall) for each group. Grand averages are restricted to the first 1,500 ms of each trial, containing only the P300 response to the first stimulus (probe/irrelevant). Probe P300 amplitude is clearly larger than Iall amplitude in the simple guilty as well as the CM groups, while probe and Iall P300 amplitudes are nearly indistinguishable in the innocent group. It is important to note that it appears more difficult here to distinguish probes from irrelevants for both the simple guilty and CM groups than it was when examining the experiment 1 grand averages.

ERPs: Quantitative Group Data

To examine observations about the grand averages quantitatively, a 2 × 3 (stimulus × group) ANOVA was run on the peak–peak P300 amplitudes across groups (see Fig. 9). There was a main effect of stimulus, $F(1, 43) = 68.4$, $p < .001$, $\eta^2 = .614$, as well and a main effect of group, $F(2, 43) = 3.23, p < .05, \eta^2 = .131$. The stimulus × group interaction was highly significant, $F(2, 43) = 13.69$, $p < .001$, $\eta^2 = .389$. To decompose this interaction, we subtracted the Iall amplitude from the probe amplitude in each subject to compute the average probe/Iall difference for each subject. A 1 × 3 ANOVA was conducted comparing each group's probe/Iall difference, and yielded a highly significant group effect, $F(2, 43) = 13.69, p < .001$. Tukey's post hoc tests revealed that the probe/Iall difference (with probe greater than irrelevant in all cases) of the simple guilty group was greater than that of the innocent group ($p < .001$) and the CM group ($p < .02$). There was

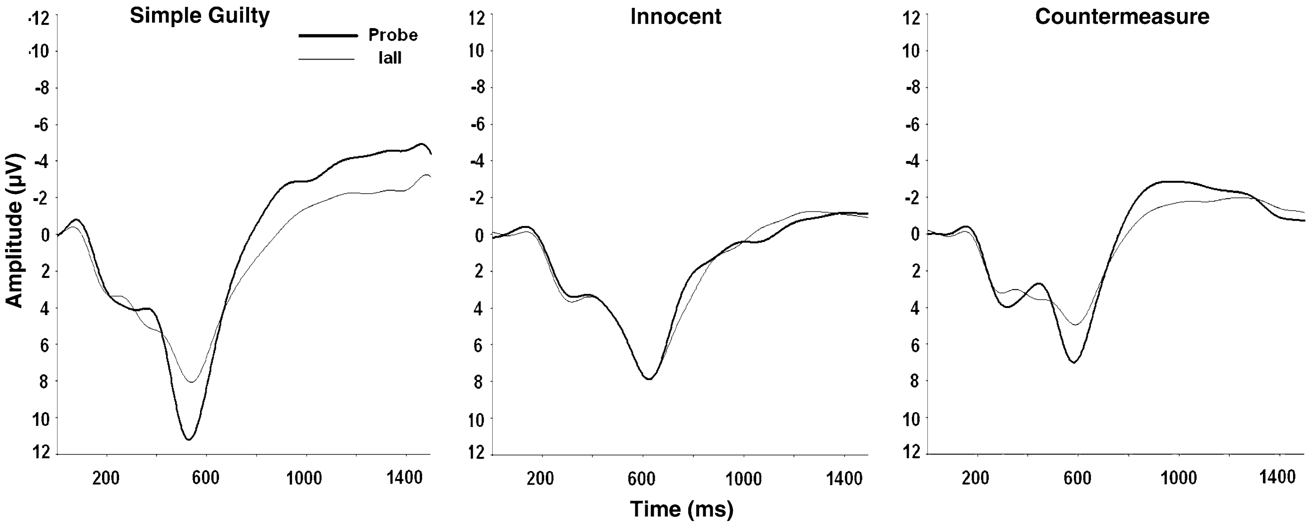

Fig. 8 Experiment 2 grand average probe and irrelevant ERPs at Pz for each group, including 100 ms baseline before stimulus presentation

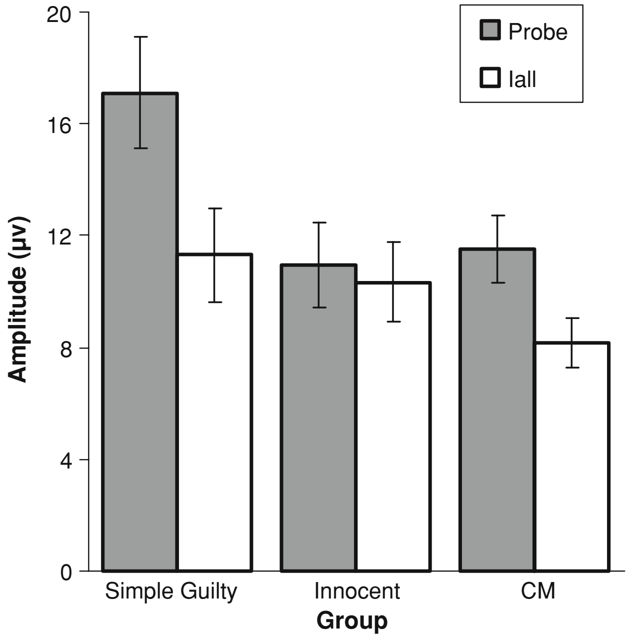

Fig. 9 Experiment 2 average probe and irrelevant peak–peak P300 amplitudes at Pz by group

also a significant difference between the innocent and CM groups ($p < .05$).

ERPs: Quantitative Individual Data

Table 3 shows detection rates within subjects for each group in experiment 2 at the .9 confidence level, as well as the number of significant iterations out of 100 in the bootstrap distribution for each subject. Bootstrap detection rates are shown for each of the three tests: Iall, Imax, and screened Imax, as described above. Iall analysis allowed for correct classification of 14/15 simple guilty subjects but yielded three false positives while correctly classifying 10/16 CM users. The Imax method once again provided a more conservative test, removing all false positives but reducing simple guilty detection rate to only 6/15 and CM detection rate to 6/16. The addition of the Imax screen allowed the detection of 1 additional CM subject (7/16 total) as well as the detection of two additional simple guilty subjects.

Grier A' Values

To evaluate and compare each test's ability to correctly discriminate between guilty and innocent subjects, we calculated the A' parameter based on the formula described by Grier (1971), $A' = .5 + \{(y - x) * (1 + y - x)/[4 * y * (1 - x)]\}$, where y is the hit rate and x is the false alarm rate. A' is a function of the distance between a receiver operating characteristic (ROC) curve and the main diagonal of the plot of hits against false alarms. It varies between 1.0, indicating perfect discrimination between honest and dishonest responders, and 0.5, indicating random discrimination. Here, the hit rate was the detection rate in either the simple guilty or CM group, and the false alarm rate was the false positive rate in the innocent group. Table 4 displays the A' value for each of the three analysis methods conducted (Iall, Imax, Screened Imax) for both simple guilty and countermeasure groups as they compare to the innocent group in experiments 1 and 2. As can be seen in the table, A' is greater for experiment 1 when examining the hit rate of simple guilty subjects and the false alarm rate of innocent subjects, while the A' is nearly identical for both experiments (slightly greater for experiment 2) when examining the hit rate of CM subjects and

Table 3 Experiment 2 individual bootstrap detection rates

Iall			Imax			Screened Imax		
SG	Innocent	CM	SG	Innocent	CM	SG	Innocent	CM
100	14	100	77	0	100	AS	NS	–
100	33	100	97	0	98	–	NS	–
84	78	100	33	20	98	AS	NS	–
100	2	20	86	0	6	NS	NS	NS
100	28	86	95	1	39	–	NS	NS
92	100	99	12	73	17	82	NS	98
100	62	100	83	14	100	99	NS	–
99	27	99	68	2	94	NS	7	–
100	30	86	100	3	56	–	NS	NS
100	34	97	91	0	83	–	AS	NS
99	97	100	97	59	99	–	NS	–
100	48	80	97	9	9	–	NS	NS
90	84	80	12	24	58	87	NS	NS
100	100	93	83	58	76	NS	NS	NS
100	43	54	22	25	36	100	NS	NS
		99			81			NS
14/15	**3/15**	**10/16**	**6/15**	**0/15**	**6/16**	**8/15**	**0/15**	**7/16**

Note: Detection rates based on a .9 confidence interval

Numbers indicate the number of iterations of the bootstrap process in which probe was greater than Iall, Imax, or the screened Imax value (depending on column). "NS" indicates that the Imax reaction time was not significantly greater than the probe reaction time. "AS" indicates that all irrelevant reaction times were significantly greater than the probe reaction time. A dash indicates that the Imax value was greater than 90, so the screened test was not performed

Table 4 Grier A' values comparing discriminative efficiency of experiments 1 and 2

	Simple guilty & innocent			Countermeasure & innocent		
	Iall	Imax	Screened	Iall	Imax	Screened
Experiment 1	0.96	0.98	0.98	0.74	0.82	0.84
Experiment 2	0.93	0.85	0.88	0.80	0.84	0.86

the false alarm rate of innocent subjects. Additionally, it is noted that the *A*' values generally increase when using the Imax analysis method as compared to the Iall analysis method.

Discussion

The studies reported here demonstrate two novel modifications of the complex trial protocol as developed by Rosenfeld et al. (2008). The modifications to the CTP described here were designed to take advantage of the fact that a task embedded within an oddball task should increase P300 amplitude. However, detection rates in the countermeasure groups for both experiments reported here are considerably lower than those found using the original CTP (Rosenfeld et al. 2008). We herein suggest that the reason for these reduced detection rates is that the increase in task demand produced by the assigned/random responses, added to the "I saw it" button presses in the current experiments increase the task demand and reduce the amount of attention paid to the critical first stimulus, thereby reducing P300 amplitude to probe items.

The removal of the target/nontarget decision from the first stimulus in the original CTP (Rosenfeld et al. 2008) reduced task demand, allowing subjects to focus more attentional resources on processing the stimulus. While the addition of countermeasure responses (an increase in task demand) caused an increase in probe P300 amplitude in the original Rosenfeld et al. (2008) experiment, this was in part mediated by the fact that subjects performed a countermeasure to every irrelevant item, but not the probe, causing a P300 amplifying effect of omitting a countermeasure to only the probe (Meixner and Rosenfeld 2009). Thus, the findings of Rosenfeld et al. (2008) imply that a decrease in task demand during a P300-based concealed information test will result in greater P300 amplitude. It is likely that the modifications made to the CTP in the current experiment caused a significant enough increase in task demand to reduce P300 amplitude to probe items and thereby reduce detection rates, especially in the countermeasure condition.

As would be anticipated by this hypothesis, the less demanding task described in experiment 1 (random responses) yielded greater detection rates than the more difficult assigned response task described in experiment 2 (see Tables 2, 3). Additionally, experiment 1 yielded a significantly greater probe/Iall difference than experiment 2. This probe/Iall difference is the critical factor that leads to high detection rates of guilty subjects. In the CM group, experiment 1 yielded a lower detection rate than experiment 2 using both the Iall and Imax methods (see Tables 2, 3) but there was no significant difference between experiments in the probe/Iall difference of the CM group. Additionally, inspection of the grand averages of both experiments (Figs. 4, 8) reveals apparently better distinction between probe and Iall in experiment 1 than in experiment 2, though this is not reflected in the detection rates. It is possible that the high level of task demand created by performing a countermeasure in addition to either the random or assigned response overwhelms the P300-eliciting oddball effect, leading to reduced detection rates. In contrast, the original CTP as reported by Rosenfeld et al. (2008) provides a very simple task for subjects (the pressing of a single "I saw it" button regardless of the stimulus) and detected CM-users at a high rate. Additionally, submitted data from our lab (Rosenfeld and

Labkovsky 2009) have found that eliminating the "I saw it" response from the protocol described in experiment 1 allows for the detection of 100% of countermeasure users. This suggests that the increase in task demand caused by the inclusion of the "I saw it" button in addition to the random response decreases the sensitivity of the test.

Regarding reaction times to both the assigned/random response and the "I saw it" button, experiment 1 showed the previously reported effect of significantly increased RTs for CM-users (Rosenfeld et al. 2004, 2008). However, within the CM group these studies also reported elevated reaction times to irrelevant items as compared to probe items, allowing the diagnosis of CM-use via RT. Experiment 1 found no significant difference between probe and irrelevant RT within the CM group (Figs. 2, 3) and the RT screening process found significant RT effects of CMs in only 2/11 subjects, leading to only 1 additional detection. It is unclear why this is, as the countermeasure instructions were highly similar to those of Rosenfeld et al. (2008) and we anticipated similar RTs. It is possible that subjects found it difficult to execute the CM response in addition to the demanding random response, and instead executed these two responses simultaneously, which would theoretically lead to the RT effects observed as subjects combine the additional CM response with the original response that a non-CM user must also make.

Experiment 2 also yielded interesting RT results in that probe RTs for both the assigned response and "I saw it" response were significantly faster than irrelevant responses for both the simple guilty and countermeasure groups, but not the innocent group. We suspect that probe RT in the simple guilty group was faster than irrelevant RT because subjects found it easier to remember the button assignment corresponding to their birthdate than to irrelevant dates. Thus, when the subject's birthdate appeared on the screen, the subject remembered which button to press more quickly than he would with an irrelevant date, leading to reduced probe RTs. This effect may have occurred in the CM group as well, though the execution of CMs would be expected to increase Iall reaction time significantly. Because of this effect, the RT screening protocol in experiment 2 was highly ineffective at diagnosing CM use. Only two screened CM subjects showed a significantly elevated RT to Imax, while two innocent subjects and six simple guilty subjects showed significantly elevated RT to Imax as compared to the probe (see Table 3). The net effect of this was positive, leading to one additional countermeasure detection and two additional simple guilty detections, but the rationale of the RT screening protocol is no longer sound when non-CM users show elevated reaction times to irrelevant items as compared to probe items. In the assigned response paradigm of experiment 2, reaction time instead served somewhat as an index of guilt—subjects who recognize the probe item as familiar are more likely to have a reduced RT to the probe. Other researchers have found reaction time to be a reliable indicator of concealed information (Seymour et al. 2000) though this may be susceptible to strategic countermeasures (Gronau et al. 2005). The reaction time effect found in experiment 2 implies that a RT screen for countermeasure use may be ineffective in any protocol requiring assigned responses to individual items in a concealed information test.

The studies described above have shown the influence of two unique modifications to the task demand in the complex trial protocol as described by Rosenfeld et al. (2008). Increases in task difficulty here reduced ability to discriminate between the probe and irrelevant items, thereby reducing detection rates. In the less demanding random response task (experiment 1) we observed detection rates similar to those found in Rosenfeld et al. (2008) in the simple guilty and innocent groups, but reduced detection rates in the countermeasure group, likely due to overloading of task demand when combining CMs with the random response. In the more demanding assigned response task (experiment 2) we observed reduced detection rates as compared to experiment 1 and Rosenfeld et al. (2008), likely due to the high level of task demand involved. Further research is still needed to better determine the role of task demand in the P300-based concealed information test, as the most sensitive protocol must focus as much attention on probe items as possible without being so difficult as to reduce P300 amplitude because of high task demand.

Acknowledgments This research was supported by the Defense Academy for Credibility Assessment DODP198-P-0001 and DoD-PI04-P-0002 awarded to J. Peter Rosenfeld.

References

Allen, 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*, 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). London: Academic Press.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences, 11*, 355–372.

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

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 P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 2, pp. 1–78). Greenwich, CT: JAI Press.

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

Grier, J. B. (1971). Non-parametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*, 424–429.

Gronau, N., Ben-Shakhar, G., & Cohen, A. (2005). Behavioral and physiological measures in the detection of concealed information. *Journal of Applied Psychology, 90*, 147–158.

Johnson, R. (1988). The amplitude of the P300 component of the event-related potential: Review and synthesis. In P. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology: A research annual* (Vol. 3, pp. 69–137). Greenwich, CT: JAI Press, Inc.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43*, 385–388.

Lykken, D. T. (1960). The validity of the guilty knowledge technique: The effects of faking. *Journal of Applied Psychology, 44*, 258–262.

Lykken, D. T. (1981). *A tremor in the blood*. Reading, Mass: Perseus Books.

Meijer, E. H., Smulders, F. T. Y., Merckelbach, H. L. G. J., & Wolf, A. G. (2007). The P300 is sensitive to face recognition. *International Journal of Psychophysiology, 66*(3), 231–237.

Meixner, J. B., & Rosenfeld, J. P. (2009). Countermeasure mechanisms in a P300-based concealed information test. *Psychophysiology* (in press).

Mertens, R., & Allen, J. B. (2008). The role of psychophysiology in forensic assessments: Deception detection, ERPs, and virtual mock crime scenarios. *Psychophysiology, 45*(2), 286–298.

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

Rosenfeld, J. P, & Labkovsky, E. (2009). Updated P300-based complex trial protocol to detect concealed information: Resistance to mental countermeasures against only half the irrelevant stimuli. *Psychophysiology* (submitted).

Rosenfeld, J. P., Labkovsky, E., Lui, M. A., Winograd, M., Vandenboom, C., & Chedid, K. (2008). The Complex Trial Protocol (CTP): A new, countermeasure-resistant, accurate P300-based method for detection of concealed information. *Psychophysiology, 45*(6), 906–919.

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

Rosenfeld, J. P., Tang, M., Meixner, J. B., Winograd, M., & Labkovsky, E. (2009). The effects of asymmetric versus symmetric probability of targets following probe and irrelevant stimuli in the complex trial protocol (CTP) for detection of concealed information with P300. *Physiology and Behavior* (in press).

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

Soskins, M., Rosenfeld, J. P., & 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*(17), 3–180.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked potential correlates of stimulus uncertainty. *Science, 150*, 1187–1188.

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

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology, 26*, 208–221.