Psychophysiology, 45 (2008), 906–919. Wiley Periodicals, Inc. Printed in the USA.
Copyright © 2008 Society for Psychophysiological Research
DOI: 10.1111/j.1469-8986.2008.00708.x

The Complex Trial Protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information

J. PETER ROSENFELD, ELENA LABKOVSKY, MICHAEL WINOGRAD,
MING A. LUI, CATHERINE VANDENBOOM AND ERICA CHEDID
Department of Psychology, Northwestern University, Evanston, Illinois, USA

Abstract

A new P300-based concealed information test is described. A rare probe or frequent irrelevant stimulus appears in the same trial in which a target or nontarget *later* appears. One response follows the first stimulus and uses the same button press regardless of stimulus type. A later second stimulus then appears: target or nontarget. The subject presses one button for a target, another for a nontarget. A P300 to the first stimulus indicates probe recognition. One group was tested in 3 weeks for denied recognition of familiar information. Weeks 1 and 3 were guilty conditions; Week 2 was a countermeasure (CM) condition. The probe–irrelevant differences were significant in all weeks, and percent hits were $>90\%$. Attempted CM use was detectable via elevated reaction time to the first stimulus. In a replication, results were similar. False positive rates for both studies varied from 0 to .08, yielding J. B. Grier (1971) A′ values from .9 to 1.0.

Descriptors: Psychophysiological detection of deception, P300, Event-related potentials, Guilty knowledge tests, Concealed information tests, Lie detection, Credibility assessment

We (Rosenfeld, Soskins, Bosh, & Ryan, 2004) and others (Mertens & Allen, 2008) reported that the deception detection protocols based on the oddball P300 recognition response to concealed information are vulnerable to countermeasures (CMs). In these earlier protocols (e.g., Allen, Iacono, & Danielson, 1992; Farwell & Donchin, 1991; Rosenfeld, Angell, Johnson, & Qian, 1991; Rosenfeld et al., 1988), three types of trials were used: probe, irrelevant, and target trials. The rare probe trials presented the suspected concealed information items that guilty suspects would (behaviorally) deny recognizing so as to deny their involvement in the crime. A murder weapon such as knife is an example of a probe. The frequent irrelevant items were items of the same type as the probe (e.g., other potential murder weapons such as pistols), but are not relevant to the crime under investigation and therefore should not be recognized by guilty suspects. The expectation was that only guilty persons would recognize the probes and respond to them with a P300. Thus there would be a difference between ERPs in probe versus irrelevant trials for guilty subjects, but not for innocent subjects. This protocol is related to the ANS-based Guilty Knowledge Test developed by Lykken (1959, 1998).

In the earlier P300-based tests, an additional type of rare stimulus trial, the target trial, was utilized: This target stimulus required a unique, instructed response in guilty and innocent suspects and was usually just another irrelevant item except for its assigned task relevance. A reason for using the target trial was to have a means of forcing attention. That is, the three types of stimuli were presented in a random order on separate trials, and because a subject never knew which trial type was about to occur, it was necessary that the subject attend all stimuli, lest the operator realize that the suspect was not cooperating, as evidenced by missing unique responses to targets. The early reports with these protocols (see above references) were very promising, showing hit rates of 85% and above in deceptive subjects. More recently, much lower rates were reported (Lefebvre, Marchand, Smith, & Connoly, 2007; Mertens & Allen, 2008; Miyake, Mizutanti, & Yamahura, 1993; Rosenfeld et al. 2004; Rosenfeld, Shue, & Singer, 2007), and, as already noted, these protocols were found to be vulnerable to CMs.

To improve the accuracy and increase CM resistance of the P300-based Concealed Information Test (CIT), we attempted to identify factors in the older P300 protocols that potentially compromised the test's sensitivities. 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 older protocol. That is, the subject's explicit task in the older protocol is to decide whether or not the stimulus is a target. However, it was also expected by previous workers 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,

Address reprint requests to: Address reprint requests to: J. Peter Rosenfeld, Department of Psychology, Northwestern University, Evanston, IL 60208, USA. E-mail: jp-rosenfeld@northwestern.edu because diversion of resources away from an oddball task by a second task reduces the oddball evoked P300 (Donchin, Kramer, & Wickens, 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.

This protocol is tested here in two studies for the first time. In each of its trials, there are two stimuli presented about 1–1.5 s apart. The subject responds to each in succession. (The protocol is called the Complex Trial Protocol, or CTP, because each trial has two distinct stimuli.) The response to the first stimulus, either a probe or irrelevant item, is a simple acknowledgement that the stimulus was seen. There is no explicit choice or discrimination to be made, as there is only one response button available. It is expected that probes will be salient and elicit P300, as in older studies. However, without a concurrent target–nontarget discrimination there is no diversion of resources from the probe recognition. We expected that the P300 recognition response to the probe in the less demanding CTP would be larger than those probe P300s seen in previous studies and thus lead to better detection of concealed information. We also reasoned that larger probe responses would remain larger than even the enhanced irrelevant P300s from subjects using a preferred CM method involving secret, specific behavioral responses to irrelevant stimuli, thus covertly changed to secret targets (Mertens & Allen, 2008; Rosenfeld et al. 2004). The target decision, still used to hold attention on each trial, is made following later presentation of the second, target or nontarget stimulus.

MAIN STUDY

Methods

Participants

The participants in the experimental group of the main study were 12 members (6 female) of a junior–senior level advanced laboratory class in psychophysiology. All had received B to A grades in two previous quarters of a neurobiology class. All had normal or corrected vision. The participants' ages were 18–22 years.

Procedures

All participants took part in three blocks of trials, one each week, for 3 weeks (as in Rosenfeld et al., 2004). In each week a different, self-referring type of information was probed for each participant, and for order counterbalance, each of six pairs of participants experienced a different order of information types across the 3 weeks. The information types in the present study were mothers' first names, family surnames, and home town names. In the first week, all participants were naïve as to the experimental design. In the second week, participants were instructed to use CMs described below. In the third week, the participants were told to repeat the first week, that is, to *not* use the previously learned CMs.

An *innocent* group of size $n = 12$ (10 women) took part in this experiment and its replication. Nine participants were from 18–26 years. One man was 44; two women were 51 and 62. These participants were obtained from a research agency in Chicago. The innocent group participants completed the experiment once and were treated exactly as described above for the first week of the main study, except there were no personally relevant, self-referring stimuli presented in the probe positions used for the stimulus lists in the guilty groups.

Regarding the CMs, 3–5 days prior to running the second (CM) week, we told each participant what the four irrelevant stimuli would be for his/her specific CM run. We told all participants—course enrollees familiar with P300 basics—in class and by e-mail to learn to associate one of the four specific CMs provided (by us to them) with each irrelevant stimulus. We also emphasized that the CM should be executed *before* the first "I saw it" button press response was made, in order that it be on time to impact the brain response to this critical first stimulus. We shared with the participants our hypothesis that if one made the "I saw it" button press response first and then executed the CM, the P300 to the irrelevant stimulus would be too late or absent to be effective. The reason for these procedures not used previously (Rosenfeld et al., 2004) and not likely to be available in the field was to prepare and enable the participants as fully as possible to defeat the test. The participants understood that the CMs worked by converting the irrelevant stimuli into covert targets. Such meaningful stimuli would evoke P300s, thus reducing the difference between probe and irrelevant P300s that ordinarily allows diagnosis of probe recognition. We reasoned that if such well-prepared participants could not defeat this test, then neither could participants in the field who lack this preparation (as was the case in the near replication).

As in Rosenfeld et al. (2004), the four CMs were (1) imperceptibly increasing pressure of the left index finger on the left leg where it rested (the response button box was under the right hand), (2) imperceptibly wiggling the left big toe inside the shoe, (3) imperceptibly wiggling the right big toe inside the shoe, and (4) imagining the operator slaps you in the face. When the participant arrived in the laboratory for the CM run, he/she was first tested about his/her CMs. (All knew them.) Then he/she was given 32 practice trials, as in Weeks 1 and 3. Left and right button presses were executed with index and middle fingers, respectively.

Detailed Trial Structure (See Figure 1)

Each trial began with a 100-ms baseline period during which prestimulus electroencephalogram (EEG) was recorded. Then, as EEG recording continued, a 0.5-cm-tall first stimulus word was presented for 300 ms in white font on a computer display 1 m from the participant's eyes. This word was either a probe or an irrelevant item. Participants were instructed to signal their having seen the first (probe or irrelevant) stimulus in white. They did so by pressing the left button immediately after they saw the white stimulus. Thus, no decision was made in response to this first stimulus; the response simply indicated the participant's having seen the stimulus, so we refer to this response as the "I saw it" response. One can run the protocol without it[1] with some success; however the protocol is more sensitive with it, and the reaction

[1] In two pilot studies, we ran the protocol almost exactly as in the main study except that the "I saw it" response was omitted. Also, in the pilot studies, the simple guilty, CM, and Innocent conditions were done on six (2 studies × 3 conditions) independent groups of subjects of mean $n = 13.2$. (There were other slight differences involving stimulus durations and target stimuli; further details will be provided on request to the senior author.) Using a bootstrap confidence level of .9, the false positive rate was 25% in both studies; the hit rate in one study was 87% in the simple guilty group and 100% in the other study. In the two CM groups, the hit rates were 87% and 93% in the two studies. At a confidence level of 95%, the false positive rates were 8% and 12.5% in the two studies; the hit rates in guilty subjects were 81% and 100% in the two studies and 95% in the CM groups of both the two studies. These yielded Grier (1971;see last paragraph of methods) A′ values of .89 to .93 across confidence levels and studies.

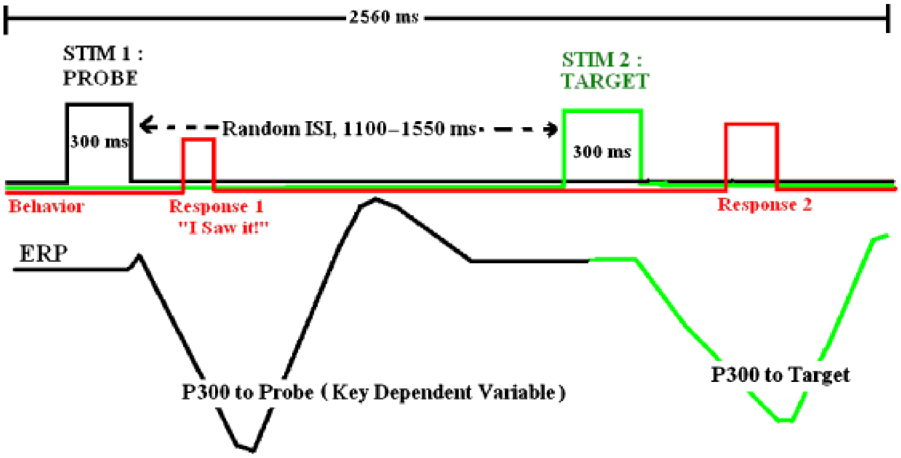

Figure 1. The structure of an example trial of the new CTP used in the main study is shown in terms of stimuli, responses, and ERPs as a f(time).

time to this "I saw it" response will be shown below to be a good indicator of countermeasure use.

The first stimulus was followed by a randomly varying interstimulus interval with a dark screen that endured for 1100 to 1550 ms. At the expiration of this dark interval, the original probe or irrelevant was presented again for 300 ms in one of five colors. Green was defined as the target color; the other colors (red, blue, yellow, purple) were defined as nontargets. Participants were instructed to press a right button for a rare target and a left button for a nontarget. Both probes and irrelevants could reoccur as targets or nontargets.

We also forced attention to the first stimulus by interrupting the run unpredictably every 20–30 trials when the first stimulus expired and requiring the participant to speak its identity. Prior to the run, the participant was alerted that missing more than one of these check-ups would result in test failure. This tended to discourage simple CMs such as vision blurring. The detailed trial events diagrammed in Figure 1 indicate a probe–target trial. Also shown is a hypothetical ERP channel. Note that because this diagram is of a probe–target trial, an early P300 in response to the probe is shown, followed by a later P300 in response to the target. We emphasize that the later P300 is of interest only in this first report to establish that the target did indeed function as a target normally does (forcing attention and eliciting a P300), but the key variable of interest with respect to concealed information detection is the response (or lack of same) to the first probe or irrelevant stimulus.

Table 1. *Stimulus Probabilities*

Stimulus type	Number	Probability
Probe target	33 (33)	.09
Probe nontarget	33 (35)	.09
Irrelevant target	33 (32)	.09
Irrelevant nontarget	260 (259)	.72
All probes	66 (68)	.18
All irrelevants	293 (291)	.82

Note: The intended original probabilities and numbers are given in bold text, and the average actual numbers of presented stimuli for which ERPs were stored (after removal of artifact-containing trials) are in parentheses.

For each block of trials (one per week), the ratio of probe to irrelevant trials was 1:4. The probabilities and numbers of the various stimuli are shown in Table 1. It is noted that probe targets and nontargets have equal probabilities, whereas irrelevant nontargets are much more probable than irrelevant targets. This was done in this first study because we wanted to confirm that irrelevant targets would evoke P300s to the targets, so we kept their probability rare. A possible confounding problem results: Probes could become much more salient than irrelevants because they are much more likely to be followed by a target; that is, the conditional probability of a target following a probe is much greater than the conditional probability of a target following an irrelevant. It was partly for this reason that we had innocent control participants complete the trial with the same conditional probabilities, but for whom probes were indistinguishable from irrelevants. High false positive rates in these participants would indicate operation of the putative conditional probability confound; it will be seen that this was not a problem.

Data Acquisition

EEG was recorded with Ag/AgCl electrodes attached to sites Fz, Cz, and Pz. Analysis here was confined to Pz. The scalp electrodes were referenced to linked mastoids. Electrooculogram (EOG) was recorded with Ag/AgCl electrodes above and below the right eye. They were placed intentionally diagonally so they would pick up both vertical and horizontal eye movements, as verified in a pilot study and in Rosenfeld et al. (2004) and Rosenfeld, Biroschack, and Furedy (2006). The artifact rejection criterion was 80 μV. The EEG electrodes were referentially recorded, but the EOG electrodes were differentially amplified. The forehead was connected to the chassis of the isolated side of the amplifier system ("ground"). Signals were passed through Grass P511K amplifiers with a 30-Hz low-pass filter setting, and high-pass filters set (3 db) at 0.3 Hz. Amplifier output was passed to a 12-bit Keithly Metrabyte A/D converter sampling at 100 Hz. For all analyses and displays, single sweeps and averages were digitally filtered off-line to remove higher frequencies; 3 db point = 4.23 Hz.

P300 at Pz was measured using the peak–peak (p-p) method, which we have repeatedly found to be the most sensitive in P300-based deception studies (e.g., Soskins, Rosenfeld, & Niendam, 2001): The algorithm searched within a window from 500 to 800 ms for the maximally positive segment average of 100 ms. The midpoint of the maximum positivity segment defined P300 latency. After the algorithm finds the maximum positivity, it searches from this P300 latency to 1300 ms for the maximum 100-ms negativity. The difference between the maximum positivity and negativity defines the p-p measure.

Analyses and Error Handling

Standard analyses of variance (ANOVAs) were run to determine group effects. Any within-subject tests with >1 df resulted in our use of the Greenhouse–Geisser (GG) corrected value of probability, p(GG), and the associated epsilon (ε) value. All error trials (as well as artifact trials) were discarded and replaced so that analyses were done only on error free trials. (An error occurred when the subject pressed the wrong button—in terms of the instructions—to a given stimulus.) This was also true for the within-subject analyses described in the next paragraph.

Within Individual Analysis: Bootstrapped Amplitude Difference Method

Standard ANOVA group analysis methods were applied to the usual P300 variables. Additionally, as this is a *diagnostic* deception detection method, we also diagnosed guilt or innocence *within individuals*. To determine whether or not the P300 evoked by one stimulus is greater than that evoked by another *within an individual*, the bootstrap method (Wasserman & Bockenholt, 1989) was used on the Pz site where P300 is typically largest. This will be illustrated with an example of a probe response being compared with an irrelevant response. The type of question answered by the bootstrap method is: "Is the probability more than 90 in 100 that the true difference between the average probe P300 and the average irrelevant P300 is greater than zero?" For each subject, however, one has available only one average probe P300 and one average irrelevant P300. Answering the statistical question requires distributions of average P300 waves, and these actual distributions are not available. One thus bootstraps these distributions, in the bootstrap variation used here, as follows: A computer program goes through the combined probe–target and probe nontarget set (all single sweeps) and draws at random, with replacement, a set of n1 waveforms. It averages these and calculates P300 amplitude from this single average using the maximum segment selection method as described above for the p-p index. Then a set of n2 waveforms is drawn randomly with replacement from the irrelevant set, from which an average P300 amplitude is calculated. The number n1 is the actual number of accepted probe (target and nontarget) sweeps for that subject, and n2 is the actual number of accepted irrelevant sweeps for that subject multiplied by a fraction (about .23 on average across subjects in the present report), which reduces the number of irrelevant trials to within one trial of the number of probe trials. The calculated irrelevant mean P300 is then subtracted from the comparable probe value, and one thus obtains a difference value to place in a distribution that will contain 100 values after 100 iterations of the process just described. Multiple iterations will yield differing (variable) means and mean differences due to the sampling-with-replacement process.

To state with 90% confidence (the criterion used in preceding studies; e.g., Farwell & Donchin, 1991; Rosenfeld, 2006b; Rosenfeld et al., 1991, 2004; Soskins et al., 2001) that probe and irrelevant evoked ERPs are indeed different, we require that the value of zero difference or less (a negative difference) not be > -1.29 *SD* below the mean of the distribution of differences. In other words, the lower boundary of the 90% confidence interval for the difference would be greater than zero. It is further noted that a one-tailed 1.29 criterion yields a $p<.1$ confidence level within the block because the hypothesis that the probe evoked P300 is greater than the irrelevant evoked P300 is rejected either if the two are not found significantly different or if the irrelevant P300 is found larger. (*T* tests on single sweeps are too insensitive to use to compare mean probe and irrelevant P300s within individuals; see Rosenfeld et al., 1991.)

In the present study, the bootstrap procedure just illustrated is applied to one block at a time of the three blocks run over the 3 weeks of the study. One obvious aim of this procedure is to compare diagnostic hit rates over the 3 weeks. We use a statistical procedure (bootstrapping) to determine that the probe P300 is larger than the irrelevant P300 with confidence level = .9, so it becomes possible to use .1 as the within-block chance hit or false positive rate, *providing we first show that none of the irrelevant P300s is larger than the others*, or providing we demonstrate that the *probe P300 is larger than the largest irrelevant P300*, which justifies the inference that it is larger than *all* irrelevant P300s. This is because, in comparing the probe P300 against the average of all four irrelevant P300s combined in the bootstrap, as we and others did in all previous studies, it is possible to obtain a positive outcome, even though one or more irrelevant P300s may be as large as or larger than the probe P300. We reasoned that if there is a rational method of providing evidence that all irrelevant P300s are of the same size, then it is reasonable to combine them into one irrelevant average against which to compare the probe P300 with a justifiable, within-block chance hit or false positive probability of .1. There are several ways in which one might do this. We have chosen here to simply compare the probe P300 to the largest irrelevant P300. We chose this bootstrap method of comparing the maximum irrelevant P300 to the probe P300 within a subject because it is uses the same approximate number of trials for each member of the comparison. However, it was also confirmed that the maximum irrelevant P300 amplitude was not associated with a statistically confirmed, unusual reaction time to the first ("I saw it") stimulus. As we show below, significant reaction time increases are associated strongly with CM use. (A detailed diagnostic algorithm will be provided on request to the senior author.)

We also separately compared the probe P300 against the *average of all* irrelevants, so as to allow comparisons to results in previous studies using this less rigorous method.

Finally, in describing diagnostic accuracy results of experiments, we made use of the signal detection theoretical parameter, A′, based on Grier (1971). This is a function of the distance between a receiver operating characteristic (ROC) curve and the main diagonal of a ROC plot of hits and false alarms. It makes no assumptions about the shape or variances of the distributions of the key variables (such as P-I P300 amplitude differences). A′ varies from .5 (null effect) to 1.0 (maximum effect). $A' = 1/2 + ((yx)(1+yx)/(4y(1x)))$, where y is the hit rate and × is the false alarm rate.

Results

Behavioral: Error Rates

Table 2 (Panels A and B) gives the error rates for both responses for all stimulus types over the 3 weeks. An error for the first stimulus is pressing the wrong button, rather than the sole, de-

Table 2. *Error Rates Sorted by Response Types, Weeks, and Stimulus Types for Main Study (Panels A and B), Replication (Panel C), and Innocents (Panel D).*

	Week 1	Week 2	Week 3
Panel A: Response 1			
PT	.003	.005	.000
PN	.005	.000	.000
IT	.005	.000	.003
IN	.000	.000	.001
Panel B: Response 2			
PT	.043	.079	.048
PN	.020	.019	.007
IT	.222	.179	.226
IN	.005	.003	.001
Panel C: Replication: Response 2			
PT	.048	.025	.039
PN	.003	.033	.000
IT	.110	.126	.064
IN	.002	.002	.000
Panel D: Innocent Group: Week 1			
	Response 1	Response 2	
PT	.003	.048	
PN	.000	.003	
IT	.000	.047	
IN	.000	.002	

Note: PT: probe–target trial; PN: probe–nontarget; IT: irrelevant–target; IN: irrelevant–nontarget.

fined ("I saw it") button, per instructions. The error rates for the first response are trivial, probably because no decision was necessary. No significant effects were obtained in a three-way ANOVA (all Fs <2, ps $> .2$) as next described for the second response. Regarding this second (target vs. nontarget) response, there were higher error rates to the targets, especially irrelevant–target stimuli. (Here, an error means responding incorrectly either to a target or nontarget.) However, none of the stimulus types appears to show systematic changes over weeks. A three-way, completely within-subject ANOVA was applied to these data for the second response. The independent variables were target versus nontarget (two levels), weeks (three levels; this tests the effect of CMs), and probe versus irrelevant (two levels). The important finding was that the main effect of weeks was far from significant, $F(2,22) = 0.022$, p(GG) $> .97$, $\varepsilon = .99$: Subjects did not tend to make more errors during the CM week, supporting the fact of their cooperation with instructions as well as their ability to do the task. There were effects, evident in the table, of target versus nontarget, $F(1,11) = 19.535$, $p < .002$, and of probe versus irrelevant, $F(1,11) = 30$, $p < .001$, probably carried by the irrelevant targets. The triple interaction was not significant, $F = 2.8$, p(GG) $= .09$, $\varepsilon = .91$. There was a large interaction, likewise evident in the table, of target versus nontarget with probe versus irrelevant, $F(1,11) = 37$, $p < .001$. This interaction reflects the large difference in error rates between irrelevant targets and probe targets, but not irrelevant nontargets and probe nontargets. The only other significant interaction effect was of weeks with irrelevant versus probe, $F(2,22) = 5.7$, p(GG) $< .02$, $\varepsilon = .96$.

Reaction time (RT) data are discussed below, as RT to the first "I saw it" stimulus was important in diagnosing CM use. RTs to the second (target/nontarget) stimulus were of no diagnostic use in this article, and all analyses, text references, and figures showing RTs from here on are for the first response to the first stimulus.

ERPs: Qualitative

Figure 2 (top half) shows the grand averaged ERPs (and simultaneously recorded EOG waves) sorted by weeks, probes versus irrelevants, and targets versus nontargets. Numbers of trials (12 subjects combined) in each grand average are also indicated. The target versus nontarget distinction (upper row vs. bottom row) is not relevant for detection of concealed information, and, in general, the upper row resembles the lower row with regard to the critical first P300 in response to probe or irrelevant. The late (target-evoked) P300, more evident in the lower (target) row than in upper (nontarget) row (see especially the Week 2 column) is as expected, although there is an apparently small P300 in the probe–nontarget average. These effects are not germane to deception. Important here, as expected, are the rather larger probe P300s than irrelevant P300s (in both target and nontarget averages), particularly in the first and third weeks, when CMs were not used. In the second week, it is clear that at least some subjects were attempting to execute the covert CM responses to irrelevants, because the irrelevant ERP averages show clear P300s in Week 2, though not in Weeks 1 and 3. Nevertheless, the probe P300s in Week 2 are not reduced from Week 1, and actually appear even larger than in Week 1, as will be clear in the line graphs, presented next. There do seem to be effects of time passage from Weeks 1 to 3, although the probe–irrelevant differences still appear potentially diagnostic in Week 3.

Figure 3a (top, left) shows line graphs of computer-calculated mean (p-p) P300 amplitudes for probe–targets and probe–nontargets combined, because analyses below shows no effect of targets versus nontargets on ERP data. Assuming no conditional probability confound as discussed earlier, this is as expected, because when the first stimulus is presented, the subject does not know whether it will be followed by a target or nontarget. Figure 3a shows more clearly that the probes increased slightly in amplitude from the first (no CM) to the second (CM) weeks than does Figure 2 (top). This is probably because Figure 2 is based on grand averages that do not take individual P300 latencies into account, whereas the values in Figure 3a (top, left) are based on individual P300 amplitude computations that utilize values for each subject at each individual peak latency. Figure 3a (top, left) also shows clearly that the irrelevant P300 grows at an even greater rate in the second CM week, as most subjects apparently executed the specific covert CMs for each separate irrelevant. Nonetheless, the probe–irrelevant difference is clearly large across all 3 weeks.

ERPs: Quantitative Group Data

In support of the above observations, a 2 (probe vs. irrelevant) × 2 (target vs. nontarget) × 3 (weeks) ANOVA was applied to the individual average P300 values. The effect of stimulus type (probe vs. irrelevant) yielded $F(1,11) = 62.1$, $p < .001$; the effect of target versus nontarget was not significant, $F(1,11) < 1$, $p > .4$; the effect of weeks yielded $F(2,22) = 12.3$, p(GG) $< .002$, $\varepsilon = .80$; and the interaction of probe–irrelevant and weeks was also significant, $F(2,22) = 6.6$, p(GG) $< .008$, $\varepsilon = .93$, probably reflecting the greater increase in irrelevant (vs. probe) increases in the second week. The interaction of weeks × target versus nontarget was not significant, $F(2,22) < 1$, p(GG) $> .7$, $\varepsilon = .96$. The interaction of probes versus irrelevant × targets versus nontargets was not significant, $F(1,11) < 1$, $p > .5$, and the triple interaction was not significant, $F(1,11) < 1$, $p = .09$, $\varepsilon = .88$. This three-way ANOVA was also done in the replication, but because target and nontarget waves never differ (and logically cannot differ), in

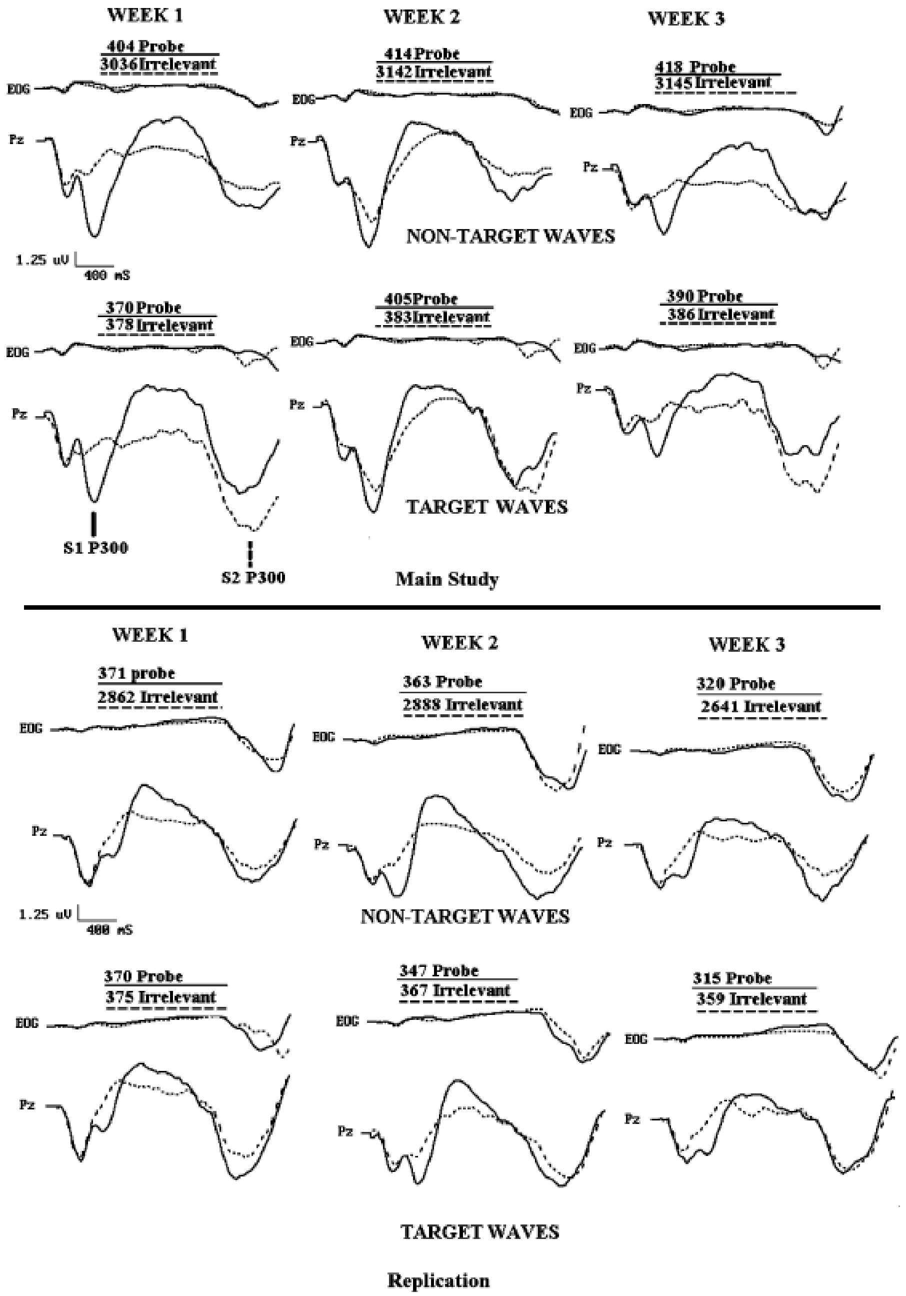

Figure 2. Grand averaged probe and irrelevant ERPs at Pz for target and nontarget trials in each of the 3 weeks of the experiment in the main (top) and near replication (bottom) studies are shown. The numbers next to the legends indicate numbers of trials across all subjects.

other analyses, target and nontarget data are combined. One post hoc test of interest concerned the probe–irrelevant difference in the CM Week 2, only: $t(11) = 4.99$, $p < .001$; the probe was still greater than the irrelevant despite CM use. (This effect was not seen in the older protocol of Rosenfeld et al., 2004, in which the probe–irrelevant difference declined to insignificance in the CM week.)

ERPs: Quantitative Individual Data (Hit Rates Based on Probe Against All Irrelevants)

Table 3a gives the detection rates, based on the bootstrapping procedures as used in previous studies described for these guilty subjects in the single CM week (Week 2) and in the two non-CM weeks (Weeks 1 and 3). As seen in Tables 3a and 3b, the results for guilty subjects are the same at both the .9 and .95 bootstrap confidence levels. These methods compared the probe P300 average against the average of all irrelevants within a subject. (The results with the more rigorous comparison of probe P300 against the largest irrelevant P300 *additionally screened with RT analysis* are given later after presentation of RT results.) The first column of numbers in Table 3a shows the results using one set of search windows for all subjects as described in the Methods section. Here, 11/12 (92%) are correctly diagnosed in the first and third weeks. In the second (CM) week, 10 subjects are correctly diagnosed. One subject had a P300 with component latencies different than those of others (as often happens; Rosenfeld et al.,

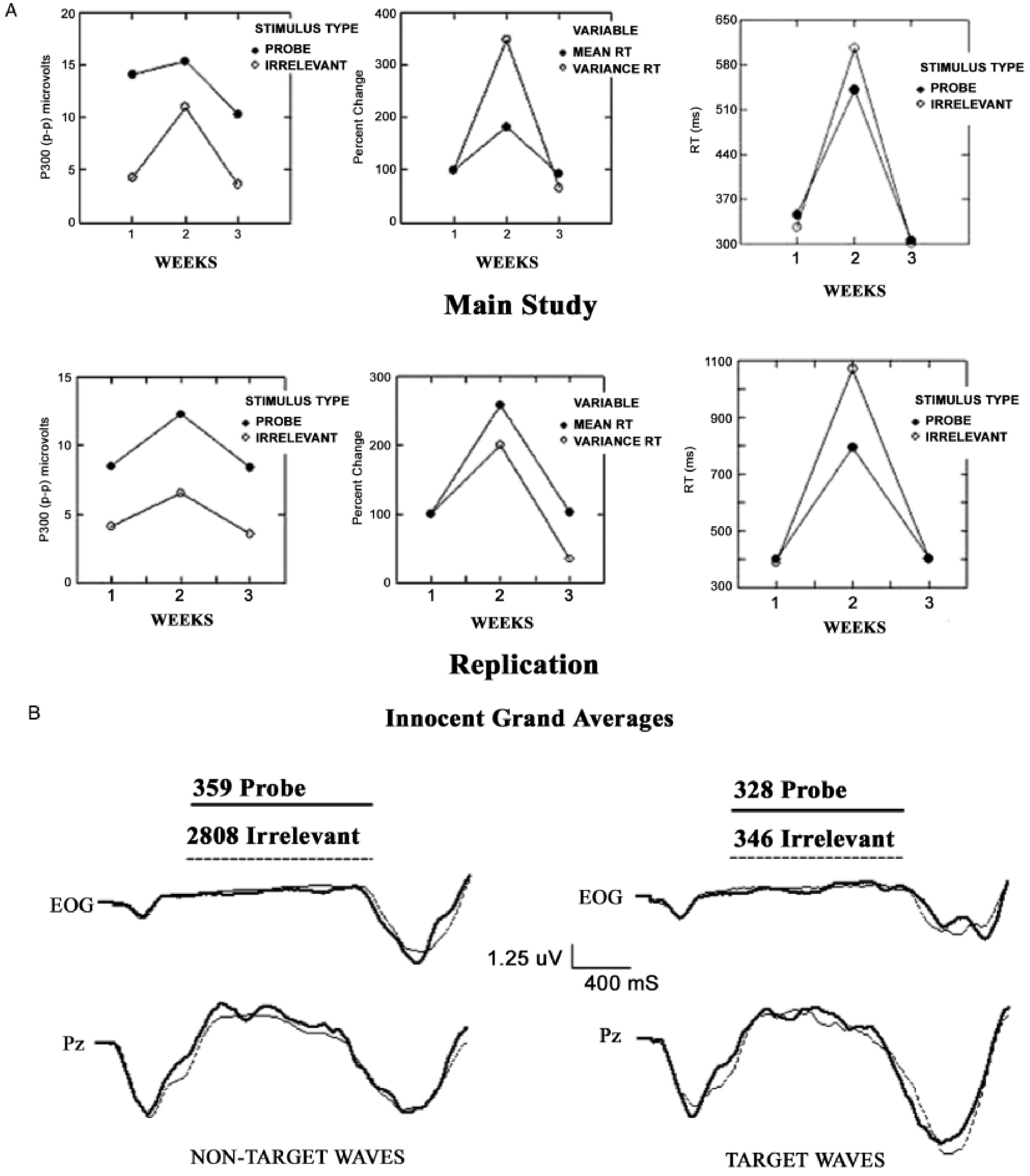

Figure 3. A: Main (top) and near replication (bottom) studies. Left: Computer calculated Pz-P300s averaged across all individual subjects in the 3 weeks of the study. Center: RT means and variances across all subjects and weeks given in terms of percent change from the first block (Week 1). Right: These are the RTs in milliseconds in the main and replication studies for stimuli and weeks as shown. These and all RTs shown in all figures to follow are RTs to the first ("I saw it") stimulus. B: Innocent group grand averages as in Figure 2.

2004), and also, much smaller amplitudes than seen in other subjects in all 3 weeks. If we used slightly different search windows for him (see Table 3a), then despite his aberrant P300s, he was indeed detected in all 3 weeks, and the rightmost column of numbers in Table 3a gives the rather impressive results overall. Either way, however, 1 of the group of 12 remaining subjects *did* defeat the test in Week 2. Nevertheless, as seen below, her RT profile reveals her attempt to use the CMs.

Table 3a. *Within-Subject Correct Detections of Guilty Subjects Based on Bootstrap Comparison of Probe P300 against the Average of All Irrelevant P300s over 3 Weeks in the Main Study*[a]

Week	Hit rate[b]	Hit rate[c]
Week 1 (no CM)	11/12 (92%)	12/12 (100%)
Week 2 (CM)	10/12 (83%)	11/12 (92%)
Week 3 (no CM)	11/12 (92%)	12/12 (100%)

[a] Results for both .9 and .95 bootstrap confidence levels were identical.
[b] These numbers are based on using one set of look windows for all subjects, 500 ms to 800 ms for the positive P300, and P300 latency to 1300 ms for the subsequent negative peak, as stated in the Methods section.
[c] Values are based on using an individually tailored pair of search windows for one case, 500 ms to 700 ms, and P300 latency to 1600 ms.

Reaction Time Group Data

Figure 3a (top, center) shows the averaged RT data (in percentage units) based on the response on all trials (probe–target, probe–nontarget, irrelevant–target, and irrelevant–nontarget combined) to the first ("I saw it") stimulus. Mean of each individual's variance values are also plotted. The Week 1 value for both RTand RTvariance was defined as 100% so that one graph with plots of both variables could be reasonably presented. This was done only for graphic representation; analyses were on RTs in milliseconds. A separate one-way (three levels) ANOVA for each variable across weeks was performed, $F(2,22) = 26.9$, $p(\text{GG}) < .001$, $\varepsilon = .52$. For RT variance, $F(2,22) = 28$, $p(\text{GG}) < .001$, $\varepsilon = .54$. RT data in milliseconds, sorted into separate probes and irrelevant values, are given in Figure 3a (right).

Reaction Time Individual Data: Combined Irrelevant Stimuli

Figure 4a shows frequency distributions of RTs for irrelevant (combined target and nontarget) trials in the first and second weeks plotted on opposite sides of one vertically oriented abscissa. It is clear that the mean RT with CMs is larger than that without CMs, as analyzed above. It is also obvious that the variance of the group is greater with CM use. However, it is also clear that that there are three individuals whose RTs during the CM week are within the *group* distribution for the first (no CM) week. None of these three defeated the test, and all three were clearly diagnosed as guilty.

The distributional overlap does not exclude a possible difference *within each individual* between the RT during CM use as opposed to the RTvalue in the absence of CM use. Every subject showed a decreased RT from CM use to CM-free performance. More quantitatively, each of these changes was statistically significant ($p < .001$) in each subject in repeated measures t tests comparing *individual* RTs from Week 1 to Week 2. The mean t value was 27.9 across the 12 subjects, and for 1 subject who defeated the test, $t = 9.9$, with a mean RT difference from Week 1 to Week 2 of 59 ms. For the other subject who, based on one set of look windows for all subjects, defeated the test, $t = 4.7$, with a mean RT difference from Week 1 to Week 2 of 26 ms. These data support the expectation that all subjects followed instructions by attempting to use CMs, mostly without success.

Table 3b. *Hit Rates and Corresponding Grier A′ (Grier, 1971) Value from Week 1 (= Week 3) Data (Most Conservative Results from Table 3a) at .9 and .95 Confidence Levels and Using Probe versus All Irrelevant Bootstrap Tests (Iall) and Probe versus Maximum Irrelevant (Imax) Bootstrap Tests from the Main Study, Innocent Group*

	Confidence = .9			Confidence = .95		
Test	FPs[a]	Hits	A′	FPs	Hits	A′
Iall	.08[b]	.92[c]	.95	0	.92	.98
Imax	0	.92	.98	0	.92	.98

[a] FPs: false positive rate.
[b] .08 = 1/12.
[c] .92 = 11/12.

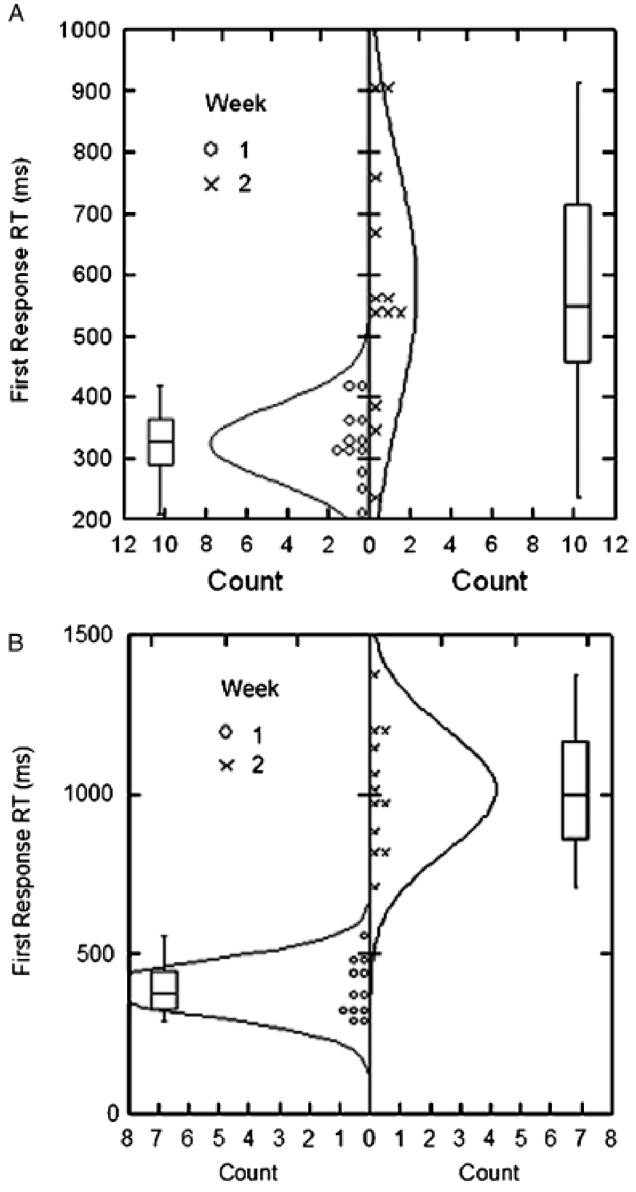

Figure 4. Main (A) and near replication (B) studies. RT distributions for irrelevant (combined target and nontarget) trials plotted vertically for the first 2 weeks. Left is Week 1 (no CMs), right is with CMs.

Reaction Time Individual Data: Probe versus Each Irrelevant

Figure 5a shows mean RTs for probes versus irrelevants in the non-CM Weeks 1 and 3 at the left and for the CM Week 2 at the right. It is clear (see y-axis numbers) that, as noted above, all RTs are elevated in the CM week. It appears also that there is some

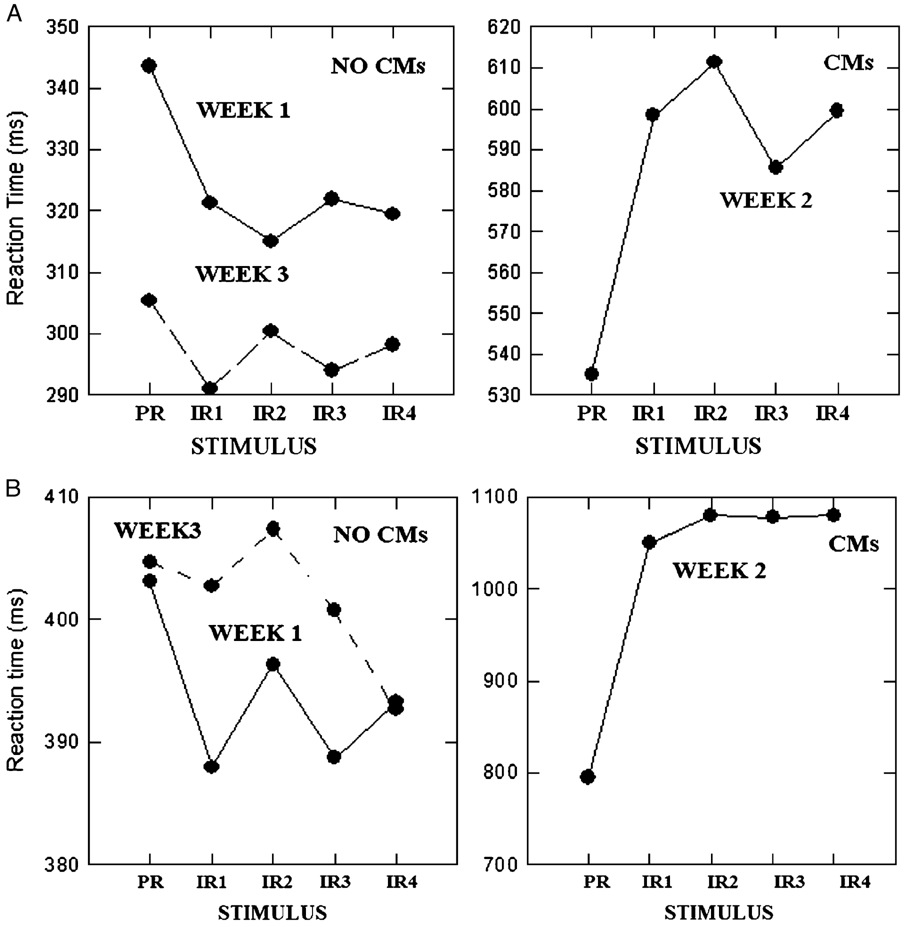

Figure 5. Main (a) and near replication (b) studies. Mean RTs for probes versus four irrelevants (IR1, IR2, etc.) in the non-CM Weeks 1 and 3 at left and for the CM Week 2 at right.

practice effect from Week 1 to Week 3 on all stimuli, although this was not significant, perhaps due to a floor effect. The main datum here, however, is that in the non-CM weeks, the probe RT appears greater than the irrelevant RTs (as previously reported by Seymour, Seifert, Mosmann, & Shafto, 2000, and others). In contrast, it is clear that even though the probe RT is elevated along with irrelevant RTs in Week 2, the irrelevant RTs are increased to a greater extent, probably reflecting the demand of having to recall and select which CM response to make. A 2 (condition: Week 1 vs. Week 2) × 5 (stimulus: probe and four irrelevants) repeated measures ANOVA was performed. The $F(1,11)$ for condition was 17.826, $p<.002$. The $F(4,44)$ for stimulus was not significant, $F=1.8$, $p>.2$, $\varepsilon=.42$; this was probably related to the significant interaction, which was $F(4,44)=7.86$, $p(\mathrm{GG})<.003$, $\varepsilon=.56$: That is, the probe RT was greater than the irrelevant RTs in Week 1 and, though this was not originally analyzed, in Week 3, but smaller in CM Week 2. In confirmation, and also including Week 3 in the analysis, a paired group t test comparing probe and average of all irrelevant RTs for Weeks 1 and 3 combined yielded $t(23)=+3.158$, $p<.005$; probes >irrelevants. The same paired t test for Week 2 yielded $t(11)=-2.688$, $p<.03$, irrelevants > probes.

ERPs: Quantitative Individual Data (Hit Rates Based on Probe Against Largest Irrelevants)

As noted in the Methods section, the most rigorous, within-individual bootstrapping analysis, used here in a P300-based concealed information test for the first time, compares probe versus *maximum irrelevant* (*Imax*) *P300 values* (Table 3c). In view of the preceding data regarding RTs during CM use, we modified these analyses using the following algorithm: (1) If the probe P300 was greater than the maximum irrelevant P300, then a guilty decision was declared with no further analysis (as stated above). We refer to this diagnosis (Table 3c) as a "simple hit." We also added to the diagnosis the fact of CM use if the average irrelevant RT was greater than the probe RT. (The detailed algorithm for diagnosis is available on request from the senior author.) (2) If a simple hit diagnosis was not obtained and if, in the test week (Week 2 here), one or more irrelevant P300s was associated with RTs that were significantly greater than the probe RT, we then chose for comparison with the probe P300 the largest irrelevant P300 whose associated RT was not greater than the probe RT. If we used this alternative Imax, CM use was of course diagnosed, and if the just

Table 3c. *Simple and RT-Qualified Diagnoses in the Main Study (at Confidence = .9), Probe versus Imax (or RT-qualified Imax) across 3 Weeks, and CM Use*

	Week 1 ($n=12$)	Week 2 ($n=12$)	Week 3 ($n=12$)
Simple hits	.92	.67	.92
Hits/RT qualified	.92	.83	.92
CM use	0.0	1.0	0.0

described comparison was positive (probe P300 > RT-screened Imax) an alternative diagnosis was made, referred to in Table 3c as an "RT-qualified hit." (3) If every irrelevant RT was significantly greater than the probe RT, then CM use was also diagnosed, and the alternative diagnosis "RT-qualified hit" was also made based on the test of probe versus the average of all irrelevants (as in Table 3c).

These modifications were based on the following reasoning: If a subject generated irrelevant RTs that were significantly larger than the probe RT, it is reasonable to conclude, based upon the preceding RT data (Figures 4 and 5), that the subject was successfully utilizing CMs. It makes no sense to compare the probe P300 amplitude to a single irrelevant P300 amplitude when there is reason, based on RT data, to believe that a CM was successfully used with this irrelevant stimulus. It makes more sense in these cases to compare the probe P300 to the maximum irrelevant P300 not associated with a significantly elevated RT or with the average irrelevant P300 as in previous studies. This is less rigorous, but still relatively conservative in first requiring the RT screen. It could be argued that it might make most sense to simply diagnose the subject as uncooperative by reason of CM use; however, we wanted to try a P300-based analysis even though we knew and made the diagnosis that CM use was being attempted. If we had simply prescreened subjects for CM use in this way and then not analyzed guilt versus innocence with P300, we would have reduced the numbers of subjects available for ERP analysis in Week 2.

Modifications 2 and 3 above were utilized only for the CM Week 2, as in Weeks 1 and 3, the simple Step 1 above was adequate for accurate diagnoses: As shown in Table 3c, using only the simple, most rigorous algorithm (probe vs. maximum irrelevant P300 amplitude), 11/12 (92%) of the subjects were again correctly diagnosed in Weeks 1 and 3, utilizing a common set of look windows for all. In Week 2, 8/12 (67%) of the subjects were still detected ("simple hits") using the rigorous, bootstrap tests in the CM Week 2, even though we knew a priori (from RT data) that all subjects in Week 2 used a CM. One of these subjects was the same one who beat the test using the less rigorous bootstrap criterion in which the probe was simply compared with the average of all irrelevants. The other three subjects were previously detected as guilty with the less rigorous analysis, but not with the more rigorous testing of the probe P300 against the maximum irrelevant P300. (The data from these three subjects are good evidence that a probe can be larger than the average irrelevant P300 while still being smaller than one or more irrelevant P300s.) Moreover, as seen in Table 3c, using the RT-qualified diagnoses, two of the four subjects undetected as "simple hits" were accurately classified as guilty ("RT-qualified hits"), raising accuracy to 83%. CM use was detected in all subjects in Week 2: For each, the average RT in Week 1 was significantly smaller ($p < .001$) than in Week 2. In summary, no matter how rigorous the diagnostic criteria, in Week 2, all subjects were diagnosed as either guilty despite CM use or noncooperative in using CMs.

The Innocent Group: False Positive and Error Rates

Table 3b gives the results obtained in the guilty group along with those of the innocent group, whose grand averages showing no probe–irrelevant differences are in Figure 3b. The point of this group was the allowance of calculations of false positive rates, which, combined with Week 1 data from the guilty group, allows calculation of Grier's (1971) A′ values also shown in Table 3b. These rather impressive results were computed at two levels of confidence (.9 and .95) and using the bootstrap tests of probe versus irrelevant average (the method used in previous studies), as well as the more rigorous test of probe versus maximum irrelevant. The A′ value allows comparison of diagnostic accuracies across various studies in various laboratories. Because the innocent group had only one session, A′ values are shown for Week 1 data only. The impressively low false positive rates also argue against operation of a conditional probability confound previously noted.

Table 2 (Panel D) shows errors for the innocent group (1 week only). These results are *notably different* than the corresponding group results from the main study and replication (next detailed) Week 1 (guilty) groups, in that irrelevant targets and probe targets have similar error rates. A 2 (probe vs. irrelevant) × 2 (target vs. nontarget) ANOVA yielded a significant effect only for the target–nontarget manipulation, $F(1,11) = 23.8, p < .001$. The probe and interaction effects were not significant, $F(1,11) < .1$, $p > .9$.

A NEAR REPLICATION STUDY

Methods

The final set of data presented is from a study that virtually replicated the main study with three differences: (1)The stimulus parameters differed. In the previously reported study, a stimulus would appear in white for 300 ms and then return again, after a randomly enduring dark screen interval, in a target or nontarget color for 300 ms. In this near replication, the first stimulus would appear in white and remain on the screen for the same random interval (1100–1550 ms) as in the main study, until it was replaced seamlessly with the change to color lasting 300 ms. Thus we here examined if the persistent white stimulus on the screen would improve performance by functioning as a fixation locus forcing attention to the expected target (nontarget) stimulus. (2) The subjects were not from an advanced university class, but, as for the innocent group, were instead recruited from the general Chicago area population by a research subject recruiting agency (ResearchChicago.Com). This group allowed us to ascertain that subjects from a general population who were unacquainted with the senior author and lacking a bias to cooperate would be as detectable as the student subjects. Three men and nine women between the ages of 21 and 35 took part in the experiment. (3) In the main study, subjects in Week 2 were taught the CM responses to each irrelevant stimulus prior to the P300 recording test. In the presently described study, CM subjects learned the CMs just before the test session. Moreover, they did not know which CM was to be associated with which irrelevant stimulus in advance of the test, but had to form these connections during the test, as in Rosenfeld et al. (2004) and as in the field. Thus this replication allowed us to observe any differences between the laboratory and field conditions of CM training. False positive data are reused from the same innocent group which participated previously in calculation of Grier (1971) A′ values.

Results

Behavioral: Error Rates

The error rate results were similar to those in the main study. The results for the first "I saw it" response were that most rates were 0.0 and the highest was .005. A three-way, completely within-subject ANOVA was applied to these data for the first response. The independent variables were target versus nontarget (two levels), weeks (three levels; this tests the effect of CMs), and probe versus irrelevant (two levels). No significant effects were found, all Fs <2, ps$>.2$. The results on the second response are shown in Table 2 (Panel C). The pattern resembles that seen immediately above for the main study, although visual inspection reveals mostly lower error rates in the replication than in the main study. A three-way, completely within-subject ANOVA was applied to these data for the second response. Again, the independent variables were target versus nontarget (two levels), weeks (three levels; this tests the effect of CMs), and probe versus irrelevant (two levels). The attempted ANOVA failed because there was insufficient variance in these data. We therefore applied this analysis to Week 1 data only, yielding a 2 × 2 ANOVA (with the weeks factor removed). Here there were significant effects of probe versus irrelevant, $F(1,11) = 5.34$, $p<.05$, target versus nontarget, $F(1,11) = 13.5$, $p<.005$, and a marginal interaction, $F(1,11) = 3.9$, $p<.08$. This is very similar to what was reported for the main study with the factor of weeks included, except that in the main study the interaction of probe–irrelevant by target–nontarget was $p<.05$ (vs. $p<.08$ here).

ERPs: Qualitative

The grand averages (in Figure 2, bottom half) in this near replication appeared generally similar to those in main study (Figure 2, top half), although probe–irrelevant differences in the replication are not as dramatic looking as in the main study. The computer-calculated grand-average P300 values across weeks are shown in Figure 3a (bottom left) and appear generally similar to those in the original study seen in Figure 3a (top, left). Both probe and irrelevant P300 amplitudes are elevated in Week 2, although the elevation appears greater for the probes than for the irrelevants in the replication from Week 1 to Week 2, and the probe decline in Week 3 appears less in the replication than in the original study. The largest apparent difference between the original and near replication is the greater increase in the irrelevant P300 from Week 1 to Week 2 in the original study. (Statistical comparison of original and replication studies is given below.)

ERPs: Quantitative Group Data

As in the original study, a 2 (stimulus type: probe vs. irrelevant) × 3 (week) × 2 (target vs. nontarget) ANOVA on the P300 group means yielded $F(1,10) = 82.41$, $p<.001$, for the stimulus type factor and $F(2,20) = 6.64$, $p(\mathrm{GG})<.02$, $\varepsilon = .84$, on the week factor. The nonsignificant interaction of weeks and stimulus type was $F(2,20) = 1.56$, $p(\mathrm{GG}) > .2$, $\varepsilon = .99$. Clearly, P300s evoked by both stimuli increase during CM use, and CMs lead to larger P300s for both probes and irrelevants. This is different from what was seen in the original study, in which the greater irrelevant increase in Week 2 (vs. the probe increase) led to a significant interaction of weeks and stimulus type. The effect of target versus nontarget was not significant, $F(1,10) = 2.25$, $p>.16$, as in the main study. Again the interaction of weeks and target/nontarget was not significant, $F(2,20) = 1$, $p>.38$, and the interaction of probe versus irrelevant × target versus nontarget was not significant, $F(1,10) = 3.3$, $p = .1$. The triple interaction was not significant, $F(2,20) = 0.24$, $p>.78$. A critical follow-up test, as for the original experiment, involved a test of probes (targets and nontargets combined) versus irrelevants (targets and nontargets combined) during Week 2, the CM week. The results, as in the main study, were $t(11) = 8.9$, $p<.001$.

ERPs: Quantitative Individual Diagnostic Data Based on Probe versus All and Maximum Irrelevants

The "simple hit" detection rates for all bootstrap comparisons at two levels of confidence are in Table 4. Without RT screening, 92%–100% of subjects were detected both with and without CMs in the first 2 weeks, although use of RT screening did detect one more subject undiagnosed as a simple hit in the CM week (Week 2). Also shown in Table 4 are false positive rates based on the innocent control group as in the original experiment and Grier (1971) A′ values of test discrimination efficiency. Because the innocent group had only one session, A′ values are shown for Week 1 data only. These values are quite strong and comparable to (or better than) those in Table 3b from the main experiment.

Reaction Time: Group Data

Figure 3a (center row, bottom) shows the same RT information for the near replication study as Figure 3a (center row, top) shows for the main study. (Figure 3a is in terms of percent of baseline, and probes and irrelevant are combined; Figure 3a, right column bottom shows RTs in milliseconds for probes and irrelevants, separately.) The results are similar: RT and RT variance increase during CM use in Week 2. The three-level (Week 1 vs. Week 2 vs. Week 3), one-way ANOVAs on both variables yielded significance: For RT, $F(2,20) = 106.2$, $p(\mathrm{GG}) < .001$, $\varepsilon = .54$; for RT variance, $F(2,20) = 4.87$, $p(\mathrm{GG}) < .05$, $\varepsilon = .59$.

Individual Reaction Time Data: Combined Stimuli

Figure 4b shows for the replication the same type of data as seen for Figure 4a for the main study. The basic trends are the same for both studies: Within each individual, the RT is elevated in the CM week compared to the first week. A difference between the two studies is the fact that Figure 4a shows a slight overlap between weeks, whereas in the replication data (Figure 4b), there

Table 4. *Within-Subject Correct Detections (Simple Hits) of Guilty Subjects Based on Bootstrap Comparison of Probe (P) P300 Against the Average of All Irrelevant P300s (I-All) over Weeks and Against the Largest Irrelevant P300 (I-Max) in the Near Replication Study*

	P vs. I-All			P vs. I-Max		
Week	Hits	FPs	A′	Hits	FPs	A′
Confidence level .90						
1	12/12 (100%)	8%	.91	11/12 (92%)	0%	.98
2	12/12 (100%)			11/12 (92%)[a]		
3	9/10 (90%)			7/10 (70%)		
Confidence level .95						
1	11/12 (92%)	0%	.98	11/12 (92%)	0%	.98
2	12/12 (100%)			11/12 (92%)[a]		
3	9/11 (82%)[b]			8/11 (73%)[b]		

Note: Diagnoses are uncorrected with RT, but see note a. Data in this table are, with two exceptions (one subject in both weeks, each with no apparent irrelevant P300s or positivity whatsoever), based on using one set of look windows for all subjects, 500 ms to 800 ms for the positive P300, and P300 latency to 1700 ms for the subsequent negative peak. The exceptional subjects' windows were determined by examining the probe P300 and finding its positive peak, then using peak +50 and − 50 ms as the look window for the P300 in probe and irrelevant waves. This was also done with the subsequent negative wave following P300, proper. False positives (FPs) and Grier's (1971) A′ values also given based on Week 1 Hits and FPs.

[a] The single undiagnosed, CM-using guilty subject in Week 2 used an I-max value associated with a significantly elevated RT. When correction was applied, the subject was correctly detected.

[b] One subject's file was lost in Week 3.

was no overlap, as in Rosenfeld et al. (2004). Also, as in the main study, within each individual in the replication, the mean of all RTs in Week 2 was significantly greater than the Week 1 mean. The *p* values were all <.001. The *t* values varied from 10 to 30 in five subjects and from 31 to 60 in the remaining seven.

Individual Reaction Time Data: Probe versus Each Irrelevant

Figure 5b shows data comparable to the main study data in Figure 5a, but for the near replication. The trends seen are similar to the main study's findings: (1) The probe RTs are usually greater than the irrelevant RTs without CMs, but smaller with CMs. (2) There is a large elevation of all RTs in the CM week. This elevation appears more marked in the near replication than in the main study. As in the main study, we did a 2 (Week 1 vs. Week 2) × 5 (stimulus type: 1 probe and 4 irrelevants, targets and nontargets combined) ANOVA. The effect of weeks was $F(1,10) = 148.882$, $p < .001$. The effect of stimulus type was also significant, $F(4,40) = 9.827$, $p(GG) < .002$, $\varepsilon = .54$, as was the interaction, $F(4,40) = 14.7$, $p(GG) < .001$, $\varepsilon = .55$. The only difference in this pattern of results from those in the main study is the significant effect of stimulus type in the near replication but not in the main study. We showed evidence that in the main study, the interaction obscured the main effect of stimulus type; however, in the replication, the greater elevation of RTs due to CM use was apparently able to overcome this interaction effect.

Replication versus Main Study: ERPs

To statistically compare this near replication with the original study, two 2 (group: original vs. replication) × 3 (week) ANOVAS were performed separately for each stimulus type (probe, irrelevant). For the probes, the group effect was in the right direction, that is, reflecting the larger probe in the main study than in the replication, but not significant, $F(1,21) = 2.54$, $p < .13$. The week effect was $F(2,42) = 8.46$, $p(GG) < .002$, $\varepsilon = .94$. The interaction was not significant, $F(2,42) < 1.3$, $p > .3$. Results were different with the irrelevants. Here, the group effect was not significant, $F(1,21) < 1.1$, $p > .3$. The week effect was significant as with the probes, $F(2,42) = 30.4$, $p(GG) < .001$, $\varepsilon = .72$. However, the interaction of weeks × group with the irrelevants was $F(2,42) = 5.8$, $p(GG) < .012$, $\varepsilon = .72$. This reflects the visually greater increase in the original than the replication study of the irrelevant P300 amplitude in the CM week, as noted above. The effect of weeks in both studies, as suggested by Figure 3a, is also carried by the increase in probe P300s in the second (CM) week, as compared with the first and third weeks. In confirmation, a paired *t* test comparing probe P300s (targets and nontargets combined) between the first and third weeks combined versus the P300 of the CM week gave $t(22) = 3.39$, $p < .003$. To confirm that the effect of weeks in the immediately preceding *t* test was not carried, in the main experiment, by the reduced Week 3 probe level, a 2 (Week 1 vs. Week 2) × 2 (group: original vs. replication) ANOVA yielded a main effect of weeks, $F(1,22) = 4.3$, $p < .05$, showing a greater probe size in Week 2 than in Week 1 (13.8 μV > 11.3 μV). The interaction was not significant, $F(1,22) = 1.1$, $p > .3$, indicating that no difference in the probe increase over the 2 weeks was apparent, and the effect of group was marginal, $F(1,22) = 3.59$, $p < .08$.

Replication versus Main Study: RTs

Figure 3a presents the group RT data for all stimulus types combined as a function of weeks in terms of percentages of Week 1 values. It is evident that the increment of RTs in the second (CM) week was greater for the subjects in the replication than for those in the main study. (See also Figure 3a, right column.) A 2 (group) × 3 (week) ANOVA yielded $F(1,21) = 25.9$, $p < .001$, for the group effect of main versus replication study. The main effect of weeks was $F(2,42) = 119.5$, $p(GG) < .001$, $\varepsilon = .53$, and the interaction of group × weeks was $F(2,42) = 17.9$, $p(GG) < .001$, $\varepsilon = .53$. Clearly RTs were greater over all stimuli and weeks for subjects in the replication, and the interaction probably reflects the greater increase in Week 2 for the replication subjects than for those in the main study.

Discussion

The studies reported here suggest that the CTP is more accurate and resistant to CMs than previously published ERP-based studies in detecting concealed information. (We do not include here more recent reports and claims of Farwell—e.g., Farwell & Smith, 2001, and on his web site called "Brain Fingerprinting"—for reasons detailed in Rosenfeld, 2006a.) We hypothesize that the reasons for the improved performance of the CTP are partly related to the psychophysiological mechanisms engaged by the CTP, but not by the earlier protocols. Our major evidence for the idea that different mechanisms are involved is that (1) the CTP is resistant to CMs of the type used here, unlike what was reported in the earlier protocols as in Rosenfeld et al. (2004), and (2) the P300 elicited by the probe stimulus is *increased* when a CM is used against the CTP (a novel finding here), whereas, as reported by Rosenfeld et al. (2004), the probe P300 is drastically *reduced* when a CM is used against the older protocols. Indeed the latter fact is *partly* related to why CMs are effective against the older but not the CTP protocols. It is only partly related, because the detection of guilt depends on the probe P300 being larger than the irrelevant P300, and although the irrelevant P300 does increase during CM usage in both the older and CTP protocols, the probe P300 decreases in the older but increases in the newer CTP protocol, which appears to compensate for the irrelevant P300 increase. The reduction of the probe P300 in the older protocols, in contrast to its augmentation in the CTP, is in part why the augmentation of the irrelevant P300 with a CM defeats the older but not the newer protocol. The theoretical challenge thus becomes identification of the special attributes of the CTP that could account for its augmenting effect on P300 in general and, in particular, when CMs are used.

Briefly, we suggest that the new protocol engages more attention (or allocation of perceptual resources) to the critical first stimuli. In the older protocol, the subject's attention to the probe versus irrelevant nature of the sole stimulus may be diverted by the need to decide whether or not the *same* stimulus is a target. No decision about probe versus irrelevant is required and, indeed, such a distinction may be ignored by the subject. The older protocols are based on the hope that probe stimuli will have enough potency to engage the subject's attention anyway, due to their hypothetically inherent salience as guilty knowledge items. In the CTP, there is no such (target/nontarget) decision required because the target/nontarget decision has been postponed until the second stimulus event occurs. Thus all the subject's attention is available to notice the probe should it be the first stimulus, as there is no concurrent target decision.

When a CM is optionally executed immediately following the first stimulus and before the first response, as done here, the stimulus event is likely to evoke even more attention to its probe versus irrelevant status because now the subject needs to decide whether or not to execute a CM if and only if the stimulus is an irrelevant. In other words, if a CM is planned to the irrelevant, more attention is required to make a now explicit probe–irrelevant discrimination. Donchin et al. (1986) summarized research that demonstrated that, although a task competing with an oddball task leads to diversion of perceptual processing resources and a concomitant *reduction* in oddball evoked P300 if the two tasks are *unrelated*, the *embedding of one task inside the oddball task* will lead to P300 *augmentation* by concentrating processing resources on the stimuli shared by both tasks. It is suggested that this is what is happening in the CTP on presentation of the first stimulus in a situation where CMs are used. That would account for P300 augmentation seen here in the CTP during the CM week. This P300 augmentation might, however, also be expected to occur in the older protocols. However, we suggest that the task *competition* required by the simultaneous target/nontarget decision in the older protocols *increases* this *independent* task demand and thereby diverts resources such that the probe P300 is depressed during a CM session, as is empirically observed (Rosenfeld at al., 2004).

We also suggest that delaying the target decision within the trial (in the CTP) is as effective at maintaining attention and cooperation with the task as having the target decision come in response to the sole first stimulus, as in the older protocols. (We also periodically tested subjects about the first stimulus, which helps maintain attention and also prevents simple CMs such as blurring vision to the first stimulus.) This is supported by the negligible error rates in response to the first stimuli and the accuracy of the CTP in detecting concealed information—which depends on subjects' attending to stimuli. There were relatively higher error rates in response to irrelevant targets as second stimuli. These did not differ over weeks (i.e., with vs. without a CM), so there is no reason to suspect that doing the CMs increased task demand to the point where cooperation with the task became problematic. We would suggest simply that because there were many more irrelevant nontargets than irrelevant targets in all 3 weeks, the subjects rightly expected the more frequent event (nontarget) to follow an irrelevant first stimulus and perseverated on this response tendency when the irrelevant was followed by a target. This does not occur in innocent groups.

This discussion, based on Table 2 and its analysis, raises a question about a possible confound operating in this study, as noted in the Methods section. Given the stimulus probabilities shown in Table 1, it may be suggested that the large P300s in response to probes seen in the present study could be indeed related to their having a greater salience, but unrelated to their containing concealed information: It is clear that a probe presentation has a 50–50 chance of being followed by a target, whereas an irrelevant presentation is at least four times as likely to be followed by a nontarget than by a target. Subjects could, during the run, come to recognize this nonexplicit asymmetry of conditional target probability and then become much more alert for the target on probe trials than on irrelevant trials, thereby endowing the probes with an oddball salience having nothing to do with concealed information. Such a confound is also consistent with the differential error rates seen in Table 2. If such a confound were operating and affecting diagnoses based on ERPs, however, one would expect to see much higher false positive rates in the innocent control subjects than what is seen in Tables 3b and 4. The false positive rates shown in these tables are mostly 0%, and even with the most liberal method of computing this rate (e.g., using 90% vs. 95% confidence levels), it was 8%. These data support the greater sensitivity of the CTP for detection of concealed information and are not consistent with the simple operation of the confound just noted.

However, there could be an interaction of this putative conditional probability effect and the guilty status of the subject: If the probe is recognized by the guilty subject as concealed information, this would facilitate recognition of the probe's greater probability of being followed by a target. To the innocent subject, the probe is simply another irrelevant stimulus. Error rate analyses are consistent with this interaction hypothesis: There was no difference between the target probe and irrelevant error rates, nor between the nontarget probe and irrelevant error rates (i.e., no interaction) in the innocent group, but this interaction did obtain in guilty subjects.

There is a reservation about the preceding conclusions comparing main study data with data from the innocent control group: As described, the former group were all students aged 18–22 from an advanced laboratory class, and the latter group was provided by a Chicago area recruiting agency. This latter group contained nine students aged 18–26, one aged 44, and two aged 51 and 62 (mean = 29.6). Although these groups are not ideally comparable, the baseline RTs of the two oldest members of the latter group were well within the group's RT distribution, and their ERPs looked typical of the whole group. Moreover, the replication study subjects were recruited from the same agency as were the innocents, and their age mean was 27.5, quite comparable ($p > .29$) to that of the innocents. Neither was there an RT difference ($p > .39$). As noted previously, A′ values based on replication and innocent subjects (Table 4) were larger than those based on the main study subjects, so concern about group comparability may be tempered.

We stated above that in the present CTP studies, the subject probably executes the CM prior to making the first "I saw it" response to the first stimulus. This order of behaviors was expected first because it was in our instructions to use this order. Second, we informed our subjects in the main study and in the replication that using the other order ("I saw it" response before CM) would probably result in the "critical brain wave" in response to the first stimulus having ended before the CM could be effective, a hypothesis that we believe to be virtually self-evident and that we have recently tested with confirmatory results. Finally, the subjects in the first, main study were advanced students in a psychophysiology laboratory course, and they reported that they agreed with our rationale for using the instructed response order.

In the CTP, the probe P300 increases from the first no-CM week to the second, CM week. However, this increase was greater in the near replication than in the main experiment reported above. Moreover, it is obvious from Figure 3a (and was statistically confirmed) that the irrelevant increase in the main experiment of the CTP was greater in the CM week than for the near replication. Also, that Figure 4 shows no overlap of RT distributions in the replication, but overlap in the main studies, suggests that CM use is more readily detected in the replication. These results are consistent with the interpretation that the subjects in the main experiment were more intelligent and thus more effective in CM use than those subjects in the near replication. This is also supported by the faster RTs in the main study and by the greater increase in RTs in the replication subjects during the CM Week 2. Finally, it is consistent with the fact that, whereas the subjects in the CM week of the main study above were taught the CM responses to each irrelevant stimulus prior to the P300 recording test, the replication subjects learned the CMs just before the test session. Moreover, they did not know which CM was to be associated with which irrelevant stimulus in advance of the test, but had to form these connections during the test, as in Rosenfeld et al. (2004).

It is clear that this novel protocol needs further research: (1) It will be necessary to run at least three blocks within a day, so as to allow testing of multiple probes when using the superior single probe (per block) protocol (Rosenfeld et al., 2007). This will make it possible to determine minimal, tolerable habituation effects. Multiple blocks also make possible the determination of the overall false positive rate, given a known, within-block chance error rate (Rosenfeld et al. 2007). (2) It will be useful to test this protocol in detection of mock crime details in addition to the self-referring knowledge items used here. This is especially needed in view of recent findings showing that (mock) crime details and other incidental knowledge items are not as readily detected by the older protocols of P300-based tests, as are self-referring items (Rosenfeld et al., 2006, 2007). This may *not* be the case for the CTP. (3) It will also be of interest to some workers to explore the use of the CTP in *scientifically valid* comparison question tests, as we did with older P300 protocols (Johnson & Rosenfeld, 1992; Rosenfeld et al., 1991). (4) Other CM approaches also need to be explored. For example, what if a subject intentionally attempted to delay the "I saw it" response to the probe items? This might tend to equilibrate RT of probe and irrelevant, making it difficult to detect CM use. On the other hand, using RT upper limit (1000 ms) cutoff methods as in Seymour et al. (2000), based on the demonstration of Ratcliff and McKoon (1981) that time-limited (800 ms) RT obviates its voluntary control, might prevent such delayed probe RTs from occurring. Gronau, Ben-Shakhar, and Cohen (2005) differed, presenting evidence that such voluntary RT manipulation may be effective, but they used much longer time limits (1500 ms) and a Stroop protocol. The issue remains controversial, however, and in all these aforementioned studies, RT was used as a concealed information detector; in the CTP protocol, RT is used as a CM detector. Indeed, it seems quite likely that adding such extra relevance to the probe by having subjects try to increase their probe RTs should increase its P300 beyond its usual size, increasing the probe–irrelevant difference and thereby aiding detection.

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.

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). New York: Guilford.

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

Farwell, L. A., & Smith, S. S. (2001). Using brain MERMER testing to detect knowledge despite efforts to conceal. *Journal of Forensic Sciences*, *46*, 135–143.

Grier, J. B. (1971). Non-parametric indexes for sensitivity and bias: Computing formulas. *Psychology 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, M. M., & Rosenfeld, J. P. (1992). A new ERP-based deception detector analog II: Utilization of non-selective activation of relevant knowledge. *International Journal of Psychophysiology*, *12*, 289–306.

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

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

Lykken, D. T. (1998). *A tremor in the blood.* Reading, MA: Perseus Books.

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

Miyake, Y., Mizutanti, M., & Yamahura, T. (1993). Event related potentials as an indicator of detecting information in field polygraph examinations. *Polygraph*, *22*, 131–149.

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

Rosenfeld, J. P. (2006a). "Brain fingerprinting:" A critical analysis. *Scientific Review of Mental Health Practice*, *4*, 20–37.

Rosenfeld, J. P. (2006b). The complex trial (CT) protocol: A new protocol for deception detection. *International Journal of Psychophysiology* [abstract], *61*, 305.

Rosenfeld, J. P., Angell, A., Johnson, M., & Qian, J. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average waveforms. *Psychophysiology*, *38*, 319–335.

Rosenfeld, J. P., Biroschak, J. R., & Furedy, J. J. (2006). P-300-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., 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., Shue, E., & 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., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology*, *41*, 205–219.

Seymour, T. L., Seifert, C. M., Mosmann, A. M., & Shafto, M. G. (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*, 173–180.

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

(Received July 9, 2007; Accepted March 25, 2008)