*Psychophysiology, 41* (2004), 205–219. Blackwell Publishing Inc. Printed in the USA.

DOI: 10.1111/j.1469-8986.2004.00158.x

# Simple, effective countermeasures to P300-based tests of detection of concealed information

J. PETER ROSENFELD,[a] MATTHEW SOSKINS,[a] GREGORY BOSH,[a] AND ANDREW RYAN[b]
[a]Department of Psychology, Northwestern University, Evanston, Illinois, USA
[b]Department of Defense Polygraph Institute, Charleston, South Carolina, USA

**Abstract**

We found countermeasures to protocols using P300 in concealed information tests. One, the "six-probe" protocol, in Experiment 1, uses six different crime details in one run. The countermeasure: generate covert responses to irrelevant stimuli for each probe category. Hit rates were 82% in the guilty group; 18% in the countermeasure group. The *average* reaction time (RT) distinguished these two groups, but with overlap in RT distributions. The "one-probe" protocol, in the second experiment, uses one crime detail as a probe. Here, one group was run in 3 weeks as a guilty group, a countermeasure group, and again as in Week 1. Countermeasure: Covert responses to irrelevant stimuli. In Week 1, hit rate was 92%. In Week 2, it was 50%. In Week 3, 58%. There was no overlap in the irrelevant RT distribution in Week 2: Countermeasure use was detectable. However, in Week 3, the RT distributions resembled those of Week 1; test-beaters could not be caught. These studies have shown that tests of deception detection based on P300 amplitude as a recognition index may be readily defeated with simple countermeasures that can be easily learned.

**Descriptors:** Psychophysiological detection of deception, P300, Event-related potentials, Guilty knowledge tests, Lie detection

Polygraphic tests of deception based upon autonomic responses have been repeatedly challenged for decades, the most recent critique from the National Academy of Science (National Research Council, 2003). Among the problems with polygraphy raised by the National Research Council report is its potential susceptibility to countermeasures. As stated by Honts, Devitt, Winbush, and Kircher (1996, p. 84), "Countermeasures are anything that an individual might do in an effort to defeat or distort a polygraph test." (See also Honts, Amato, & Gordon, 2001.) The National Research Council report concluded, "Countermeasures pose a serious threat to the performance of polygraph testing because all the physiological indicators measured by the polygraph can be altered by conscious efforts through cognitive or physical means" (National Research Council, 2003, p. 4).

In recent years, alternative approaches to polygraphic deception detection have been developed (National Research Council, 2003, chap. 6). In the academic psychophysiology community, the use of the P300 event-related potential (ERP) is probably the most familiar of the alternative approaches (e.g., Allen, Iacono, & Danielson, 1992; Farwell & Donchin, 1991; Johnson & Rosenfeld, 1992; Rosenfeld, Angell, Johnson, & Qian, 1991; Rosenfeld et al., 1988; see review by Rosenfeld, 2002). Most of these approaches are *concealed information tests* or *guilty knowledge tests*, which utilize P300 amplitude as an index of recognition of critical details of a crime or other concealed information. The National Research Council report suggested that such novel approaches offered promise since "there is an established tradition of using brain electrical activity measures to make inferences about neural correlates of cognitive and affective processes…" and that this approach "provides a potentially powerful tool for investigating the neural correlates of deception" (National Research Council, 2003, p. 161). Nevertheless, the report tempered this enthusiasm with caveats, one of which was "In addition, it is not known whether simple countermeasures could potentially defeat this approach by generating brain electrical responses to comparison questions that mimic those that occur with relevant questions" (National Research Council, 2003, p. 162). The primary goal and major interest of the present study was to address precisely this question.

It seemed timely to investigate countermeasures to ERP-based tests also because although there have been many laboratory studies claiming 85–95% accuracy, only one *field* study has been published, but it reported approximately chance

This effort was supported, in part, by funds from the Department of Defense Polygraph Institute as project DoDPI98-P-0001. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of Defense of the U.S. Government. Thanks to Andrew B. Dollins of the Department of Defense Polygraph Institute for his review and comments on an earlier version of this article. We are extremely grateful to Gershon Ben Shakhar for a thorough review of earlier versions of this article.

Address reprint requests to: J. P. Rosenfeld, Department of Psychology, Northwestern University, Evanston, IL 60208, USA. E-mail: jp-rosenfeld@northwestern.edu.

accuracy (Miyake, Mizutani, & Yamahura, 1993). Nevertheless, one user of these methods claims 100% accuracy and is presently attempting to commercialize them (see http://www.brainwave-science.com/). Finally, the ERP approach has now surfaced in popular novels, for example, Coonts (2003), as a foolproof method.

The P300-based concealed information test presents rare probe stimuli, which represent guilty knowledge elements, in a Bernoulli series of more frequent crime-irrelevant stimuli. As guilty subjects are expected to recognize guilty knowledge items as meaningful stimuli that are relatively rare, these rare and meaningful attributes of P300-eliciting stimuli are expected to elicit P300 s in response to probe stimuli, but not in response to frequent, meaningless irrelevant stimuli. A previous study of countermeasures to P300-based concealed information tests utilized the distraction procedure of having subjects count backwards by sevens (Sasaki, Hira, & Matsuda, 2002). However, just as we found in a pilot study (noted below) for the present report, Sasaki et al. found this mental countermeasure to be largely ineffective. In this present report, our effective countermeasure strategy involved making irrelevant stimuli task relevant (i.e., meaningful) by assigning covert responses to them, thus defeating the intended probe–oddball paradigm.

There were other, secondary issues also less formally explored: (1) It will be noted that there have been at least two somewhat different paradigms utilized: The ''one-probe'' method presents separate blocks of trials, with only one critical (crime-related) detail used per block (e.g., Rosenfeld et al., 1988, 1991). The ''six-probe'' method presents just one trial block with multiple critical details (six in the first exemplar of this protocol; Farwell & Donchin, 1991). We would expect the latter to involve greater task demand, leading to smaller P300 s (Kramer, Sirevaag, & Braune, 1987), and thus poorer detection rates. (2) The two protocols have been associated with different methods of analysis. The comparative accuracy of these methods will also be explored. (3) A third issue concerns the nature of the subject utilized in these laboratory analogs. We would expect advanced students and collaborators of experimenters, as in Farwell and Donchin (1991) to be more highly motivated to perform the tasks correctly and pay closer attention to instructions. (4) We also formally study reaction time as an adjunct method of indicating countermeasure use.

## General Methods

In the studies of P300 amplitude as a recognition index for concealed information, there are typically three kinds of stimuli presented to subjects: probes, which concern concealed information known only to guilty persons and authorities; irrelevants, which are items irrelevant to the interests of authorities and unrelated to criminal acts; and targets, which are irrelevant items, but to which subjects are asked to press a ''yes'' button, so as to signal that they are paying attention, and cooperating with the task. In this report and in previous studies, probes and targets have a 1/6 probability, irrelevants have a 4/6 probability. The items are randomly presented one at a time on a video display screen every 3 s (as recommended by Farwell & Smith, 2001). The dishonest subject will press a ''no'' button to each probe occurrence, falsely signaling nonrecognition, even though he recognizes the item. Our instructions make it explicit that by pressing the ''no'' button to a probe, the subject is making a dishonest response, that is, is telling a lie with the button. He will press the ''no'' button honestly to the irrelevant stimuli, and the ''yes'' button honestly, as instructed, to the target stimuli. The target stimuli serve two purposes: First, they force the subject to attend to the display, as failure to respond appropriately to target stimuli will suggest noncooperation. Second, the target is a rare and task-relevant stimulus that evokes a benchmark P300 with which other ERPs can be compared in some analyses (described below).

The basic assumption of the P300 concealed information test is that the probe is recognized (even if behaviorally denied) by the dishonest subject, and is thus a rare but meaningful stimulus capable of evoking P300. For the innocent subject, the probe is simply another irrelevant and should evoke a small or no P300. As noted, there were originally two analytic approaches taken (until Allen et al., 1992, added a Bayesian method) in order to diagnose guilt or innocence. Ours (Soskins, Rosenfeld, & Niendam, 2001) has been to compare the amplitudes of probe and irrelevant P300 responses; in guilty subjects, one expects probe > irrelevant; in innocent subjects, probe is just another irrelevant and so no probe–irrelevant difference is expected. We use what is described in the next paragraph as the bootstrapped amplitude difference (SIZE) method. The other approach, introduced by Farwell and Donchin (1991), is based on the expectation that in guilty persons, the rare and meaningful probe and target stimuli should evoke similar P300 responses, whereas in the innocent subject, probe responses will look more like irrelevant responses. Thus in this approach, called here bootstrapped correlation analysis of disparity (FIT), the cross correlation of probe and target is compared with that of probe and irrelevant. In guilty subjects, the probe–target correlation is expected to exceed the probe–irrelevant correlation. The opposite is expected in innocents. Both of these analytic methods will be described next because they are utilized and compared in both the studies to be presented here.

### *Bootstrapped Amplitude Difference (SIZE)*

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) is usually 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 question answered by the bootstrap method is: ''Is the probability more than 95 in 100 (or 90 in 100 or whatever) 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 the distributions, in the bootstrap variation used here, as follows: A computer program goes through the probe 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 below. 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 sweeps for that subject, and n2 is the actual number of accepted irrelevant sweeps for that subject. The calculated irrelevant mean P300 is 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 95% confidence that probe- and irrelevant-evoked ERPs are indeed different, one requires that the value of zero difference or less (a negative difference) *not* be $> -1.65$ standard deviations below the mean of the distribution of differences (1.29 standard deviations for 90% confidence). In other words, the lower boundary of the 95% (or 90%) confidence interval for the difference would be greater than 0. It is noted that sampling different numbers of probes and irrelevants could result in differing errors of measurement; however, studies have shown a false positive rate of zero utilizing this method (Ellwanger, Rosenfeld, Sweet, & Bhatt, 1996) and others have taken a similar approach (Farwell & Donchin, 1991) with success. This method has the advantage of utilizing all the data, as would an independent groups $t$ test with unequal numbers of subjects. It is further noted that a one-tailed 1.65 criterion yields a $p < .05$ confidence level 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 P300 s within individuals; see Rosenfeld et al., 1991.)

***Bootstrapped Correlation Analysis of Disparity (FIT)***

The other analysis method used to compare ERPs within individuals (FIT) determines if 90% (or 95%) or more of the 100 iterated, double-centered cross correlation coefficients between ERP responses to probe and target stimuli are greater than the corresponding cross correlations of responses to the probe and irrelevant stimuli. If so, the subject is found to be guilty (this is the Farwell & Donchin, 1991, criterion and method). For example, within each subject, the program starts with all 300 single sweeps to probe ($n = 50$), target ($n = 50$), and irrelevant ($n = 200$) stimuli, and, at each time point, determines the familiar within-subject ERP average over all stimuli. This series of points comprising the average ERP is called **A** (a vector). Then the computer randomly draws, with replacement, 50 probe sweeps from the probe sample of 50. These are averaged to yield a bootstrapped probe average. Similarly, a bootstrapped target average, and irrelevant average are obtained, except the latter is based on a 200-size draw. **A** is now subtracted from probe, target, and irrelevant. This is called "double-centering" and is performed because it enhances the differences among ERP responses. The first of 100 Pearson cross correlation coefficient pairs are now computed for the cross correlation of probe and target and for probe and irrelevant ($r$ [probe – target] and $r$ [probe – irrelevant], respectively). The difference between these two $r$ values, D1, is computed. The process is iterated 100 times yielding D2, D3,… D100. Using the Farwell and Donchin (1991) method, the number of D values in which $r$ [probe – target] $> r$ [probe – irrelevant] is then counted. If this number is greater than or equal to 90, a guilty decision is made. If this number is less than 90, then a guilty decision is not made. In the present studies, a mathematically similar criterion is used in which the confirmed normal distribution of D values is considered. If zero is more than 1.29 standard deviations (90% confidence) below the mean, then a guilty decision is made.

***Bootstrapped Analysis of Reaction Time (RT-BOOT)***

The bootstrapped analysis of reaction time (RT-BOOT) uses methodology identical to SIZE with the exception that instead of brainwaves, only reaction times are considered. It has been shown that reaction time may be a useful deception detector in that reaction time to probes is longer than reaction time to irrelevant stimuli (Seymour, Seifert, Shafto, & Mosmann, 2000). To pose and answer this question *within individuals*, RT-BOOT randomly samples, with replacement, average reaction times for the probes and irrelevants and subtracts the irrelevant average from the probe average. One hundred iterations of the above process yield a distribution of 100 differences between bootstrapped average reaction time to the probe and irrelevant stimuli. If the value of zero is more than 1.65 standard deviations (95% confidence) below the mean of the difference distribution, then the subject is considered guilty. (In general, we prefer to use a 95% confidence interval. We are here using 90% with ERPs because that is what Farwell and Donchin (1991) used with their FIT method for ERPs.)

## EXPERIMENT 1

This study was directed at developing a countermeasure to the Farwell and Donchin (1991) paradigm. These authors utilized a mock crime scenario with 6 details selected as probes, 24 details defined as irrelevants, and 6 other irrelevant details were utilized as targets. After shuffling, all items were repeatedly presented one at a time in random order. Responses to all probes, targets, and irrelevants were stored as single sweeps, though also separately averaged by category into separate P300 averages for display.

It is noted that the subjects used by Farwell and Donchin were paid volunteers, including associates of the experimenters. Our presently reported study uses introductory psychology students as subjects, more like the subjects one might find in the field in the sense of relative lack of motivation to cooperate with operators, and perhaps lower intelligence. The decision of Farwell and Donchin to use six probes, one may surmise, is based on the initial decision of guilty knowledge test developer Lykken (1959, 1981, p. 251) to use six multiple choice questions, each containing one probe among six multiple choice items. Lykken's notion was that a guilty subject should be shown to respond to each of a plurality (say 4/6) of stimuli, a result having a low chance probability. The appropriateness of this logic to ERP methods will be reconsidered in the discussion.

### Methods

***Participants***

The participants in this experiment, as approved by the Northwestern Institutional Review Board, were undergraduates at Northwestern participating in order to fulfill a course requirement. All had normal or corrected vision. Participants were randomly assigned to one of three groups; a simple guilty group, an innocent group, or a countermeasure group. There were a total of 33 participants (11 per group) after 6 participants were dropped due to high blink rate ($n = 4$) or failure to follow instructions (i.e., failure to press yes to the targets $> 10\%$ of the time; $n = 2$). The male–female ratio was either 5:6 or 6:5 in each group.

***Data Acquisition***

EEG was recorded with silver electrodes attached to sites Fz, Cz, and Pz. The scalp electrodes were referenced to linked mastoids. EOG was recorded with silver 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. The artifact rejection criterion was 80 μV. The EEG electrodes were referentially recorded but the EOG electrodes were differentially amplified. The forehead was grounded. Signals were passed through Grass P511 K amplifiers with a 30-Hz low-pass filter setting, and with 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 125 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 was measured in two ways: (1) Base to peak method (BASE–PEAK): The algorithm searches within a window from 400 to 900 ms for the maximally positive segment average of 104 ms. The prestimulus 104-ms average is also obtained and subtracted from the maximum positivity to define the BASE–PEAK measure. The midpoint of the maximum positivity segment defines P300 latency. (2) Peak to peak (PEAK–PEAK) method: After the algorithm finds the maximum positivity, it searches from P300 latency to 2,000 ms poststimulus onset for the maximum 104 ms negativity. The difference between the maximum positivity and negativity defines the PEAK–PEAK measure. We have repeatedly shown that using the SIZE method, PEAK–PEAK is a better index than BASE–PEAK for diagnosis of guilt versus innocence in deception detection (Soskins et al., 2001); it will be utilized here unless otherwise noted.

***Experimental Design and Procedures***

This study was an approximate replication and extension of the study by Farwell and Donchin (1991). Guilty and countermeasure group participants were trained and then performed one of two mock crime scenarios. Scenario assignment was at random. With each scenario were associated six specific details (later to be used as the probes), knowledge of which indicated the participation of the individual in that scenario. One scenario involved stealing a ring with a name tag out of a desk drawer in the laboratory. Probes included the item of jewelry stolen, the color of the paper lining the drawer, the item of furniture containing the ring, the name of the ring's owner, and so forth. The other scenario involved removing an official university grade list for a certain psychology course taught by a specific instructor mounted on a blue-colored construction paper, posted on a wall in a certain room. Probes included what was stolen, the color of the mounting paper, the name of the course, and so on.

To insure awareness of the relevant details, the training of a guilty participant (as in Farwell & Donchin, 1991) involved several repetitions of the instructions followed by tests that participants passed before beginning the ERP-based lie test. (We appreciate that such a procedure has very little ecological validity, but used it in the interest of replication.) Following successful completion of the instructional knowledge test and performance of the mock crime, participants underwent an ERP-based concealed information test (guilty knowledge test) for knowledge of the scenario executed. Innocent group participants, having participated in neither scenario, were given the same ERP tests, half with one scenario, half with the other. During the ERP-based concealed information test, stimuli consisting of single words were presented visually on a monitor 1.0 m in front of the participant for the duration of 304 ms. The interstimulus interval was 3,048 ms, of which 2,048 ms were used to record the ERP. (These timing parameters were chosen as they were used in the most recent embodiment of the Farwell & Donchin, 1991, paradigm as described by Farwell & Smith, 2001.) Participants were instructed to press one of two buttons in response to each stimulus. In response to stimuli designated as targets, participants were instructed to press a different button, the "yes" button signifying recognition, than in response to all other stimuli. The participants were not instructed regarding the fact that some of the nontarget stimuli were probes whereas others were irrelevants. It was nevertheless expected that probes would be recognized, though responded to in effect, dishonestly, via presses of the "no" button signifying nonrecognition. As noted above, our instructions explicitly stated that if guilty, they would be lying on probe trials. For all tests, participants completed at least 180 trials. On each trial, subjects were presented with one word that was a target, a probe, or an irrelevant, depending on the trial. These were presented in a Bernoulli sequence such that there were no cases in which one word was repeated twice in a row.

There were six *categories* of stimuli with six stimuli per category: As noted above, in the scenario involving removal of the ring from a drawer, one category was jewelry items; the ring was the stolen item and the five other stimuli were irrelevant items of jewelry (watch, bracelet, etc.), and one of these was designated the target for the category. The structure of stimulus sets was the same for the other categories, and the other scenario. For each scenario, the set of targets contained six words, each of which was repeated five times in each block so that the total number of target trials was at least 30 or 17% of the total number of trials. For each target, there were 4 irrelevants presented, for a total of 24 unique irrelevants. Each of the irrelevants was also repeated five times for a total of 120 or 67% of the total trials. The six probes for the six categories of each scenario were each also repeated five times per block yielding at least 30 or 17% of the total number of trials. It is noted that data rejected due to ocular artifact were not used and more trials were collected, usually yielding more than a total of 180 trials experienced.

Prior to the test block, participants were shown a list of the target stimuli for that block and asked to repeat and memorize the list to a 100% criterion. They were then instructed to press one button (labeled "yes") in response to those target stimuli and another button (labeled "no") in response to all other stimuli during the task as soon as the stimuli appeared on the screen. The time from stimulus onset to button press was the RT.

Participants were randomly assigned either to the *innocent* group, *countermeasure* group, or the no-countermeasure, simple *guilty* group. The only difference between the latter two groups is that those in the countermeasure group were informed about a countermeasure. They were instructed to perform covert acts in response to the irrelevant stimuli. More specifically, all participants in the countermeasure group were instructed to perform the acts in response to irrelevant stimuli based on stimulus groups or categories, rather than to individual stimuli. Each member of the group performed the same acts. For example, for the name-of-operation category, animal names were used. For the probe "cow" and target "pig," the irrelevants were: horse, goat, sheep, and mule. In this case, the same covert countermeasure behavior was performed every time the participant saw any of these irrelevants, as they are all in the same irrelevant group or animal name category. The covert acts were

(1) pressing the left *forefinger* imperceptibly greater on the leg where the left hand rested, (2) pressing the left *middle* finger imperceptibly greater on the leg where the left hand rested (the right *forefinger* was being used to press *overt* response buttons), (3) imperceptibly wiggling the big toe in the *left* shoe, (4) imperceptibly wiggling the big toe in the *right* shoe, and (5) imagining the experimenter slapping the participant in the face. There was no act for the sixth irrelevant stimulus category, all members of which would thus become oddballs as the only irrelevant stimuli *not* requiring a particular covert response. Thus the essential countermeasure strategy was to make presumed irrelevant stimuli relevant.

***Analyses***

To determine the success rate of the countermeasure, SIZE with base–peak and peak–peak P300 and FIT were performed. The use of these analysis methods also allowed comparisons with regard to efficacy and resistance to countermeasures. It is noted that because Farwell and Donchin (1991) utilized a 90% confidence interval with the FIT method they introduced, we also utilize a 90% confidence level for SIZE analyses in this study for purposes of comparison of methods. Additionally, for the first time, analysis with RT-BOOT was performed. The innocent group served to provide estimates of false positive rates for each analysis method. For each analysis method, one is diagnosed as not guilty if one's behavioral data show that one paid attention to the stimuli ($>90\%$ of the targets correctly identified with the unique button press), and if the ERP analysis method did not yield a guilty result. This was the case with each analysis method employed.

## Results

***Behavioral***

In the 33 retained subjects, all followed instructions as indicated by the fact that proportions of erroneous responses to the three categories of stimuli were well under 10%, as seen in Table 1, which shows error rates to the three stimulus types for the three groups. Independent groups ANOVAs were done separately for each stimulus type to assess task effects in the three groups. For the probes, $F(2,30) = 5.7$, $p < .008$; for the irrelevants, $F(2,30) = 3.6$, $p < .04$. There was no effect with targets ($p > .5$). Table 1 indeed suggests that these effects are due to the greater task demand in the countermeasure group where the error rates to probes and irrelevants are greater in the countermeasure than in the other groups. Nevertheless, with these rates all $<7\%$, it is clear that all groups cooperated with the task, and that excessive errors do not help much in identifying individual countermeasure users. RT data will be considered later.

***ERPs: Qualitative Analysis***

In all descriptions of P300 amplitude to follow, the results at site Pz only are noted because Pz is the site where P300 is usually reported to be maximal, and because the analytic diagnostic procedures (below) are performed on Pz data only. Figure 1, left column, shows grand averages in the guilty group for superimposed probe, target, and irrelevant responses. It is as expected that a moderately larger P300 response is seen to the probe than to the irrelevant at Pz. Figure 1 also shows that although the target is larger than the probe, the morphology of probe and target are similar. It is worth noting that although the analytic tests, which are performed only at Pz, will show $>80\%$ detection of guilty subjects, Figure 1 suggests that the P300 to the probe at other sites is not different than that to the irrelevant.

**Table 1.** *Error Rates in the First Countermeasure Study*

| | Targets (%) | Probes (%) | Irrelevants (%) |
|---|---|---|---|
| Guilty group | 4.2 | 1.5 | 0.1 |
| Innocent group | 5.2 | 0.1 | 0.1 |
| Countermeasure group | 6.6 | 2.6 | 0.5 |

Figure 1, middle column, shows superimposed ERPs in the innocent group, and it is clear that there is little difference between the probe and irrelevant P300 s, as expected, as, for the innocent, the probe is just another irrelevant. However, the target response towers over the probe response, as expected. This is the prototypical innocent picture. The expected effect of the countermeasure is shown in the right column of Figure 1: Probe and irrelevant are virtually identical in the countermeasure group. Of course, they superimposed in the innocent group also (so in this sense, the countermeasure users appear innocent), although there appears to be more of a P300 for *both* probe and irrelevant in the countermeasure group. This is probably because in the innocent group, the probe is just another irrelevant, but in the countermeasure group, the probe is relevant because the subject is guilty, yet the irrelevants have also been made task-relevant by the covert responses. Finally in the right column of Figure 1, it is clear that the countermeasure has produced the desired effect in that the target P300 at Pz clearly exceeds (by 2.25 μV, BASE–PEAK or PEAK–PEAK) the probe P300, which is about the same as the irrelevant. *This is the innocent look; that is, probe = irrelevant, target > probe. When a guilty subject shows this innocent look, we will refer to the effect as a "classical defeat" of the test*. Because we will repeatedly refer to this effect later in this article, we will present one quantitative result here: A paired $t$ test on the probe versus target P300 yielded $t(10) = 2.48$, $p < .04$ BASE–PEAK and $t(10) = 1.66$, $p = .12$ PEAK–PEAK. It is further noted that in the BASE–PEAK P300, 10/11 target responses were substantially (>1.5 μV) larger than probe responses. For PEAK–PEAK, the proportion was 9/11. There were no effects of probe versus irrelevant.

***ERPs, Quantitative Analysis***

Table 2 gives the proportions of guilty decisions as a function of group and analysis method (SIZE vs. FIT). Also presented are results with RT-BOOT, an analysis of the differences in RTs, probe minus irrelevant, as it is expected that the RT to the irrelevant stimuli should increase due to performance of the covert acts in the countermeasure group. Table 3 gives the results of Experiment 1 in terms of the signal detection theoretical parameter, $A$', based on Grier (1971). This is a function of the distance between a receiver operating characteristic curve and the main diagonal of a receiver operating characteristic plot of hits and false alarms. It makes no assumptions about the shape or variances of the distributions of the key variables (such as probe-irrelevant P300 amplitude differences). $A' = 1/2 + ((y - x)(1 + y - x)/4y(1 - x))$, where $y$ is the hit rate and $x$ is the false alarm rate.

The following results are to be highlighted: Most members of the guilty group (82%) are detected with SIZE on PEAK–PEAK values, and as usual (e.g., Soskins et al., 2001), PEAK–PEAK outperforms BASE–PEAK (73%). The false alarm rate using

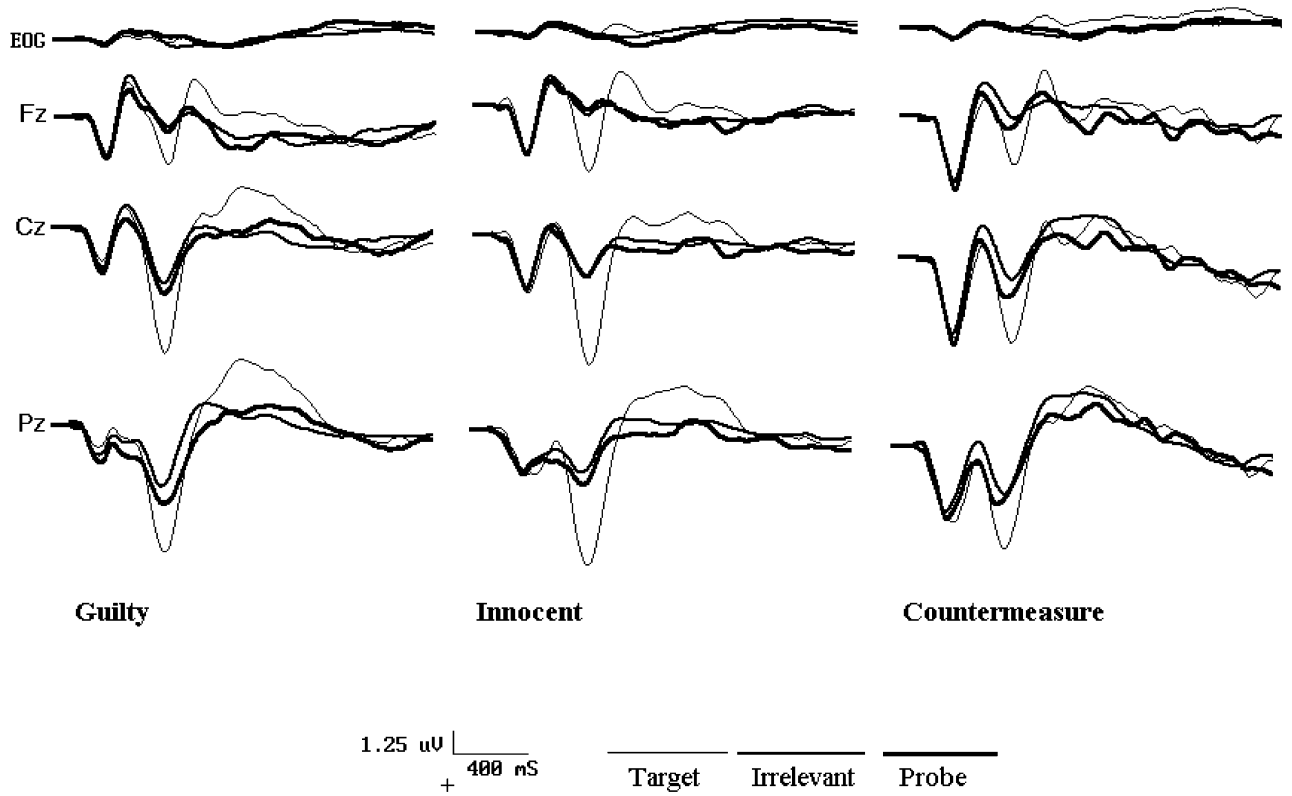


**Figure 1.** Left column: Grand average ERPs in the guilty group; probe, irrelevant, target at four sites as indicated. Probe > irrelevant is clear at Pz where analyses are usually performed and P300 is largest. Positive is down in all ERP figures. Middle column: Innocent group, superimposed grand averages to probe, irrelevant, target. Note that probe is similar to irrelevant, but target towers over probe and irrelevant. Right column: countermeasure group. Superimposition of probe, target, and irrelevant. Note target > irrelevant and probe, which are similar.

SIZE on PEAK–PEAK and BASE–PEAK data is 9%. Thus the manipulations appear to be working, lending credibility to the effect of the countermeasure which (with SIZE, PEAK–PEAK) reduces the 82% hit rate in guilty subjects to 18% in guilty subjects using the countermeasure, $p = .08$, Fisher exact test. *A*' is also reduced from .92 (SIZE, PEAK–PEAK) to .65. Secondly, it is clear that in terms of detection of guilty subjects, FIT performs poorly (54%) in these guilty subjects. SIZE (PEAK–PEAK) outperformed FIT at $Z = 2.45$, $p < .05$ on McNemar's test of differences between correlated proportions. In terms of signal detection methodology, the *A*' parameter of Grier (1971) is less (.89) for FIT than for the SIZE (PEAK–PEAK) value (.92), but the difference is not large; indeed, overall detection efficiency as indexed by *A*' is similar across all four methods. The high *A*' for FIT is likely due to the 0.0% (0 of 11) false positive rate with FIT versus .09 (1/11) with the other indices. If the false alarm rate for FIT had been 1/11, the *A*' would have been .82, compared to .92 for SIZE (PEAK–PEAK). With an *N* of only 11, the 0.0 value for false alarm rate may not be reliable. It is indeed unremarkable that such a rate is obtained with a test having such low sensitivity (54% detection), which would be unacceptable in field situations.[1] The poor hit rate with FIT is not simply a matter of too stringent a criterion in the FIT test, as we redid the FIT analyses with a criterion of .8 and got the same 54% hit rate (and 0% false alarm rate) as we did using the .9 criterion.

[1] We agree with the National Research Council report (2003, p. 50) that signal detection indices (like Grier's *A*') are profitably used in comparisons of diagnostic test results from multiple sources and/or involving differing methodologies. Such indices integrate both hit and false alarm data and thus provide overall detection efficiency estimates in one number; hence, we provided Table 3 as well as Table 2. [Of course, the data from Table 3 are very obviously directly derived from Table 2 and it is easily verified that, as is intuitively obvious, a plot of *A*' as a *f*(hit rate) at constant false alarm rate is a monotonically increasing function, just as a plot of *A*' as a *f*(false alarm rate) at constant hit rate is a monotonically decreasing function.] We believe, however, that in determination of countermeasure effects, hit rates in guilty subjects are the preferable dependent measures in the appropriate subject group. Although Honts et al. (2001) observed that 45.8% of innocent subjects in their laboratory analogs report using countermeasures, the National Research Council report (2003, p. 145) responded to this report by stating "it is unwise to conclude that countermeasures are equally prevalent in high stakes field situations." It is also noted that the Honts et al. study was based on the comparison question test protocol—not the concealed information test protocol used here. There is no information on countermeasure use by innocents against a concealed information test. In this matter, the National Research Council report ultimately concluded (2003, p. 146): "Of course, the most serious concern about countermeasures is that *guilty* [our italics] individuals may use them effectively to cover their guilt." Indeed the Honts et al. study reported that countermeasure use was attempted by 67.7% of the guilty subjects (vs. 45.8% in innocents). Observing this more expected type of countermeasure use in the present study was thus primarily facilitated by observing countermeasure effects on detection rates in *guilty* subjects, not with the overall efficiency indices from signal detection.

**Table 2.** *Outcomes, on Four Tests, of First Countermeasure Study in Terms of Percentages of Diagnosed Guilty Subjects*

| Group | SIZE (b–p) | SIZE (p–p) | FIT | RT-BOOT |
|---|---|---|---|---|
| Guilty | 8/11 (73%) | 9/11 (82%) | 6/11 (54%) | 10/11 (91%) |
| Innocent | 1/11 (9%) | 1/11 (9%) | 0/11 (0%) | 1/11 (9%) |
| Countermeasure | 2/11 (18%) | 2/11 (18%) | 6/11 (54%) | 5/11 (45%) |

b–p: BASE–PEAK, p–p: PEAK–PEAK.

**Table 3.** *Outcomes of First Countermeasure Study in Terms of Grier's (1971) A' Index*

| Group | SIZE (b–p) | SIZE (p–p) | FIT | RT-BOOT |
|---|---|---|---|---|
| Guilty | .90 | .92 | .89 | .95 |
| Guilty/countermeasure | .65 | .65 | .89 | .80 |

b–p: BASE–PEAK, p–p: PEAK–PEAK.

**Table 4.** *Outcomes (Detection Percentages) of Second Countermeasure Study*

| Week | Condition | SIZE | FIT | RT-BOOT |
|---|---|---|---|---|
| 1 | Naïve, no countermeasure | 12/13 (92%) | 9/13 (69%) | 8/13 (62%) |
| 2 | Explicit countermeasure | 6/12 (50%) | 3/12 (25%) | 0/12 (0%) |
| 3 | No countermeasure | 7/12 (58%) | 3/12 (25%) | 5/12 (42%) |

PEAK–PEAK index of P300 used.

We believe the poor performance of FIT here (vs. 87.5% hit rate in Farwell and Donchin, 1991) is in part attributable to the greater P300 latency variance one might expect to see in the unmotivated naïve subjects run in the present study, versus the motivated, paid subjects of Farwell and Donchin, some of which were colleagues of the authors. In particular, differences in latency between target and probe stimuli could lead to out-of-phase ERPs to target and probe; FIT, which looks at the simple cross correlation of probe and target, could find low cross correlation coefficients between such out-of-phase responses. This situation, which leads FIT to a miss decision, is illustrated in Figure 2. Figure 2A shows superimposed probe and irrelevant responses in this guilty subject; with probe $\gg$ irrelevant, he is clearly guilty and that is the outcome of the SIZE test. However, Figure 2B shows the target and probe waves as well out of phase, and thus FIT failed to detect this subject. At least three other subjects appeared to show this pattern.

Indeed, even the respectable 82% hit rate seen here with SIZE (PEAK–PEAK) in guilty subjects is about 5–10% lower than we generally report (Rosenfeld & Ellwanger, 1999) using the single probe paradigms described in the next study. It could also be the case that the six-probe paradigm is more complex than the one-probe paradigm, producing more task demand, which depresses P300. It is incidentally noted that Allen and Iacono (1997) reported that the FIT method was slightly *more* accurate that the SIZE method. This is not really contradictory, as Allen and Iacono were using paid subjects (unlike our Introductory Psychology Pool participants), and their SIZE analysis was on BASE–PEAK amplitudes, which we know to be up to 30% less accurate than PEAK–PEAK amplitudes in ERP-based concealed information tests (Soskins et al., 2001).

To get some evidence on these matters, we ran a separate (six-probe) study exactly like the present guilty group, except that the 14 subjects were advanced, likely more motivated, sophisticated subjects in an elective, upper level laboratory course. In this study, both FIT and SIZE (at a 90% level of confidence) detected 10/11 (91%) subjects. (Three subjects were dropped for having target error rates $>10\%$.) Clearly, motivation is not systematically manipulated in comparing SIZE versus FIT between two groups who differed in class standing, but who could have also differed in intelligence, proportion of psychology majors, and so on. The results suggest a more systematic study in the future.

Finally, regarding Table 2, it is noted that the reaction time data indicate that although RT correctly classifies 91% of the guilty subjects, that figure is halved to 45% (and *A*' is reduced from .95 to .8) when the countermeasure is used. These data are based on the probe-irrelevant difference. It might be suggested

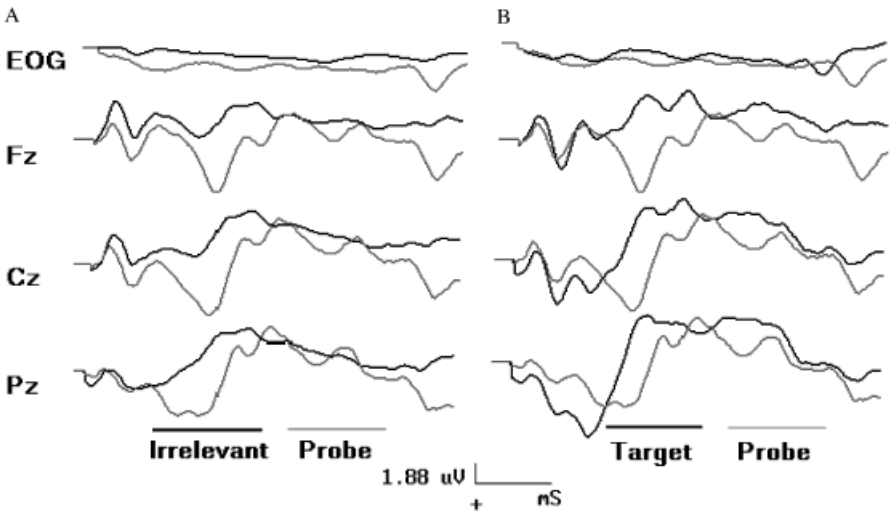


**Figure 2.** Data from 1 participant in the guilty group. A: Superimposed averages of probe and irrelevant suggest clear guilt. B: Superimposed target and probe average responses from the same guilty participant; note striking P300 phase shifts evident at Cz and Pz. Positive is down.

that the expectedly increased *absolute* value of RT for either probe or, especially, irrelevant in the countermeasure group could alert an examiner that countermeasures are being used. As Figure 3A will make clear, however, although this expectation is borne out by group analysis, it is *not* in individual analysis.

The RT variance in Figure 3A (top) is clearly greater in the countermeasure group than it is in the guilty group, as is the mean RT ($t = 3.1$, $df = 20$, $p < .008$); however, it is evident that at least 5 members of the countermeasure group fall into the distribution of the guilty group, and all 5 of these subjects were successful in defeating the test (i.e., they were undetected by SIZE, PEAK–PEAK). Indeed the 2 slowest members of the countermeasure group—in the RT sense—failed to beat the test (slow responders seem to be slow countermeasure learners). Figure 3A (bottom) shows the similar pattern of results with respect to the irrelevant stimulus: Again, the difference between means is significant ($t = 4.5$, $vdf = 20$, $p < .001$), however, it is clear that for any given individual undetected by SIZE, one can never be certain from RT whether or not a countermeasure is being used. One would like to think that certain very high RT values are typical of countermeasure users; however, as we have seen with probe stimuli, with irrelevant stimuli also, the 3 fastest responders all defeated the test, and the 2 slowest responders did not escape detection with the P300 test. It is emphasized that these data are from unmotivated, unsophisticated subjects. It is likely that well-trained, highly motivated subjects (extremists) could be practiced and fast, even on their countermeasure trials.

We also utilized Grier's (1971) *A*' measure to examine the ability of RT to detect countermeasure use in guilty subjects. We defined hit rate here as the proportion of countermeasure users correctly detected as countermeasure users, and false alarm rate as the proportion of simple guilty subjects (not using a countermeasure) classified falsely as using a countermeasure. Classification was based on drawing lines through superimposed countermeasure/guilty distributions as in Figure 3A at points designed to provide maximum separation of the subjects in the two groups. We also generated the distributions (not shown here) of probe-minus-irrelevant RT differences, because this difference is used by the RT-BOOT analysis, which was able to detect 91% of the guilty subjects with an *A*' value of .95 (based on guilty and innocent groups; see Table 2). The *A*' value for guilty versus countermeasure groups was a respectable .82 using the RT difference measure: 7/11 of the countermeasure users were correctly classified (poor sensitivity) and 2/11 of the simple guilty subjects were wrongly classified as countermeasure users. For the countermeasure group, the difference scores ranged from −40 to 271 ms in comparison to 51 to 289 ms in the guilty group, that is, there was much overlap: Except for 3 of 22 subjects, all the others in both groups were within an overlapping range. Thus, despite the significant mean difference between difference scores, countermeasure versus guilty group, $t(10) = 2.185$, $p < .05$, and *A*' score of .82, RT difference is not much help in detecting *individual* countermeasure users.

Doing the same analysis as described in the preceding paragraph on the RTs to the irrelevant stimuli (i.e., deriving *A*' from the distributions of Figure 3A, bottom, yielded a high *A*' = .92 (corresponding to the high *t* value at $p < .001$ for the difference between the distributions in Figure 3B described earlier). But again, as discussed above, the overlapping distributions discourage confident decisions about countermeasure use in given *individuals*.

In Figure 3B are the RT distributions comparing innocent and countermeasure groups. The figure looks extremely similar to Figure 3A, and again, the fastest countermeasure users who

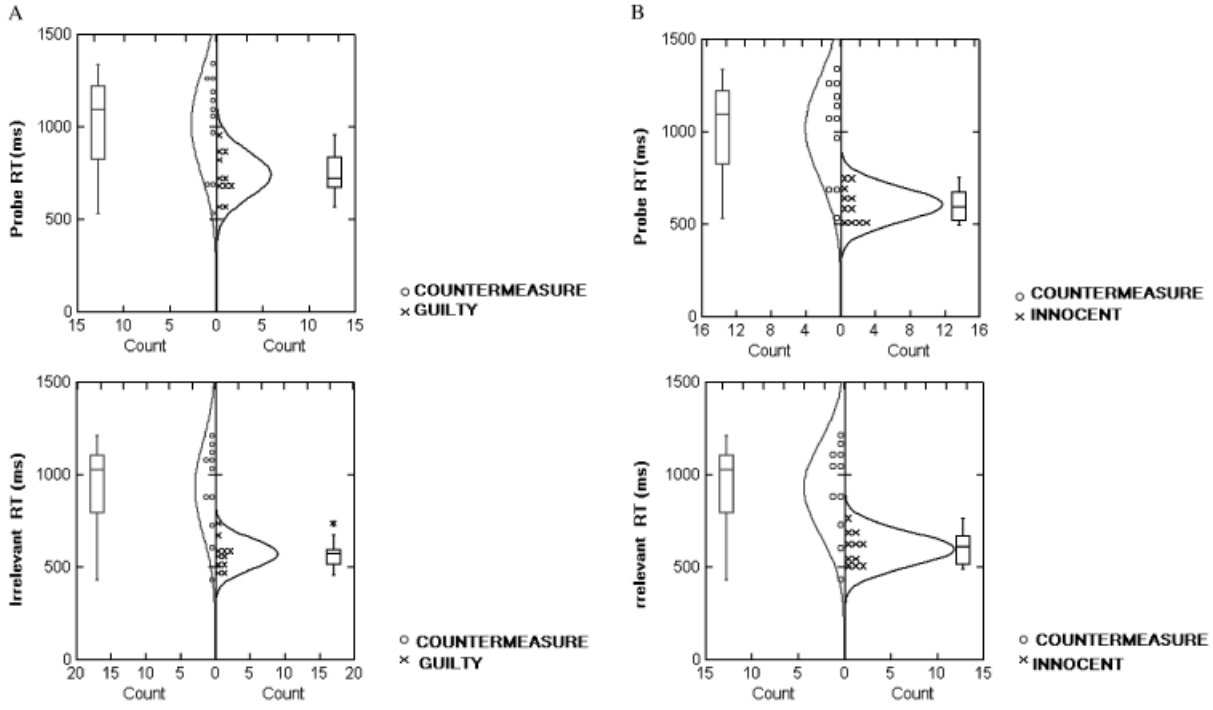


**Figure 3.** A: top: Reaction times and fitted distributions of dishonest probe "no" responses for both countermeasure (left) and guilty (right) groups. Bottom: RTs to irrelevant/"no" stimuli in countermeasure (left) and guilty (right) groups. B: top: Reaction times and fitted distributions of dishonest probe "no" responses for both countermeasure (left) and innocent (right) groups. Bottom: RTs to irrelevant/"no" stimuli in countermeasure (left) and innocent (right) groups.

beat the test have RTs within the distribution of innocent subjects.

### Discussion

The previous study showed that the six-probe paradigm of Farwell and Donchin (1991) may be significantly impacted by the countermeasure of making irrelevants secretly relevant. The result was that the probe and irrelevant responses became largely indistinguishable in the guilty subject employing a countermeasure successfully (as in Figure 1, right column, a grand average figure which well represented most individuals). Both stimuli evoked reduced P300 responses of about the same small size. One could argue that an investigator could become suspicious in such a case because theoretically, one would not expect *any* P300 response to irrelevant stimuli. However, the fact of the subjects' cooperation would be supported by the accurate response rates ($>90\%$) to the target stimuli. Also, it turned out that the target responses in the countermeasure group (Figure 1) were larger than the probe responses, which would make it very difficult to press the case that the subject was guilty, but using a countermeasure. The large target response would indicate a normal P300 to a sole oddball and a cooperative subject. One would perhaps conclude that the subject was aberrant in the sense of having a small but distinct P300 to irrelevants and probes, but one could not conclude guilt. (Indeed, small, as opposed to no, P300 s to frequent stimuli are common.) Clearly, however, an ideal countermeasure would make the subject's responses look like those in Figure 1, middle column, above, the responses of an innocent subject, in which the target response towers over the probe response, which contains no or a very small P300 response, comparable to the irrelevant response: probe = irrelevant and target $\gg$ probe. Again, we refer to such a pattern as a *classical defeat* of the test.

## EXPERIMENT 2

As just shown, countermeasures are effective against the six-probe protocol. In the present study, we examine effects of countermeasures on the one-probe protocol, and also explore the possibility that once learned, countermeasures may exert a disruptive influence in subsequent tests even though the subject does not explicitly use them. This empirical hypothesis was based on a pilot finding in which 13 subjects took part in 3 weeks of experiments. In the first week, there were no countermeasures, in the second, countermeasures were used, and the third week repeated the first week. Surprisingly, we found that in the third week, the target responses were "released" to Week 1 size while probes and irrelevants remained reduced even though countermeasures were not in explicit use. Classic defeat patterns were seen in 7 subjects in Week 3. The present study was a more formal attempt at replication of the pilot study (which did not include reaction time data or controls for habituation).

### Methods

#### *Participants*

The participants in the experimental group were initially 14 members (5 female) of a junior–senior level advanced laboratory class in psychophysiology. All had taken and received B+ to A grades in two previous quarters of a neurobiology class. All had normal or corrected vision. Attrition of 1 participant in the first week and another thereafter is described below. A control group (no countermeasure) of 10 paid volunteers (5 graduate students, 6 senior psychology majors, 5 females) was also run; the class members of the experimental group were no longer available for this procedure, nor, due to their experience as experimental subjects, would they have been appropriate participants. Moreover, the controls were paid volunteers who were associates of and were recruited by the advanced students in the laboratory class. These controls were advanced undergraduate or graduate students doing independent studies in other laboratories, and thus represented virtually the same population as represented by the experimental participants.

#### *Procedure*

All participants in this study were guilty in the sense of having concealed birth date information. The experimental group was run through the one-stimulus birthday paradigm described above in 3 successive weeks. In the first week, they were completely naïve about the countermeasure, and were told that the first experiment was simply to demonstrate the ability of the P300-based concealed information test to detect behaviorally denied autobiographical information. In the second week, they were instructed in the countermeasure. They were specifically told to execute the covert finger press upon encountering the first nontarget, nonprobe, that is, irrelevant stimulus, the covert toe wiggle upon encountering the second irrelevant, and the mental visualization of being slapped by the instructor upon encountering the third irrelevant. They were told explicitly to do nothing upon seeing the fourth irrelevant, which would take care of itself by being the only irrelevant stimulus requiring no response—an oddball in that sense. In the third week, the participants were told to perform without the countermeasure, as they had done the first week.

The *control* group completed the protocol as in the first week for the experimental participants, and they proceeded that way for all 3 weeks. We told them, truthfully, that we were interested in possible changes in the patterns of responses over a 3-week period. The timing parameters of stimulus presentation and duration were just as in the first experiment. For all participants, each run consisted of a minimum of 180 trials with each of the four irrelevant, one probe, and one target stimuli, each repeated a minimum of 30 times, yielding average ERPs of at least 30 sweeps each.

#### *EEG and Data Analysis Methods*

These were exactly the same as in the previous experiment.

### Results

#### *Behavioral*

RT data will be presented later. That the experimental subjects followed instructions is evidenced by the fact that only 1 subject in the first week had a target error (a "no" response) rate $>10\%$. His ERP data were not used. Corrupted ERP files were later found in 1 other subject for Weeks 2 and 3. His RT data were used. Thus for ERP analysis, $n = 13$, 12, and 12 for the 3 weeks, respectively. The average target error rates for Weeks 1–3 on all remaining subjects were 6.8%, 2.0%, and 6.1%, respectively; these differences failed to reach significance in a $1 \times 3$ ANOVA. Errors to probes would be *truthful* ("yes") responses; the proportions of these were low also in weeks 1–3: 0.8%, 1.3%,

and 0.5%, respectively (no significant difference). Errors to irrelevants would be "yes" responses also. The rates in Weeks 1–3 were 0.3%, 0.1%, and 0.1%, respectively (no significant difference).

*ERPs. Qualitative*. Figure 4, first column, has the grand averages from the run of the first week in experimental subjects. This is the expected and usual result of running the one-probe paradigm in guilty subjects (as reported in Soskins et al., 2001, and papers cited there): The P300 s of probe and target are similarly large and tower over the response to irrelevant at all sites. This is in contrast to the results with the six-probe paradigm (Figure 1), where there was a smaller difference between probe and irrelevant, mostly restricted to Pz. The next column of Figure 4 shows the grand average ERPs for the second week of the experimental run, that is, when the countermeasure was explicitly in effect. The P300 s to probe and irrelevant, particularly at Pz, the site where P300 is usually largest, are small and of similar size. The probe P300 is slightly larger than the irrelevant P300 because not all of the subjects contributing to these averages successfully defeated the test. Figure 4, column 2, also indicates that the targets were greatly reduced also, and are not much larger than the probes. That is, all three stimulus types generated a small P300, probably because all were meaningful. The reduced size is probably due to the loss of unique oddball probability for probe and target.

The next column of Figure 4 shows the superimposed grand averages during the third run of the experimental subjects (Week 3), when they were instructed to *not* use the countermeasure anymore. It is the case that the probe response is only slightly larger than the irrelevant (compare the first column showing the countermeasure-naïve subject), which is because not all subjects contributing to the grand average defeated the test in the absence of explicit use of the countermeasure. However, it is clear that in the week after the explicit use of the countermeasure, the *target* response is again (i.e., as in the pilot study) "released" to its normally large size, relative to the probe, in all or most subjects in the third week. (Figure 5B, a line graph of computed values, makes this more obvious.) As a group, the experimental grand averages only tend toward a classical defeat here (target $\gg$ probe, but probe still $>$ irrelevant), but for at least 5 of 12 subjects, the true classical defeat pattern was indeed obtained, as is clear from the last column of Figure 4, showing the grand averages of these 5 test beaters in Week 3. None of the five cases in the last column of Figure 4 were called guilty by either FIT or SIZE (90% confidence, PEAK–PEAK). Not shown are the individuals contributing to these averages. Each and every one of them shows the classical defeat pattern. In another 4 subjects, although the probe is somewhat larger than the irrelevant, the target, again, towers over both probe and irrelevant (as in the third column of Figure 4). As one might surmise, SIZE which simply looks at probe-irrelevant did successfully detect these subjects, although FIT did not.

Results for the control group are shown in Figure 6; their response patterns for all 3 weeks are similar and strongly resemble those of Figure 1, leftmost column, above, except that

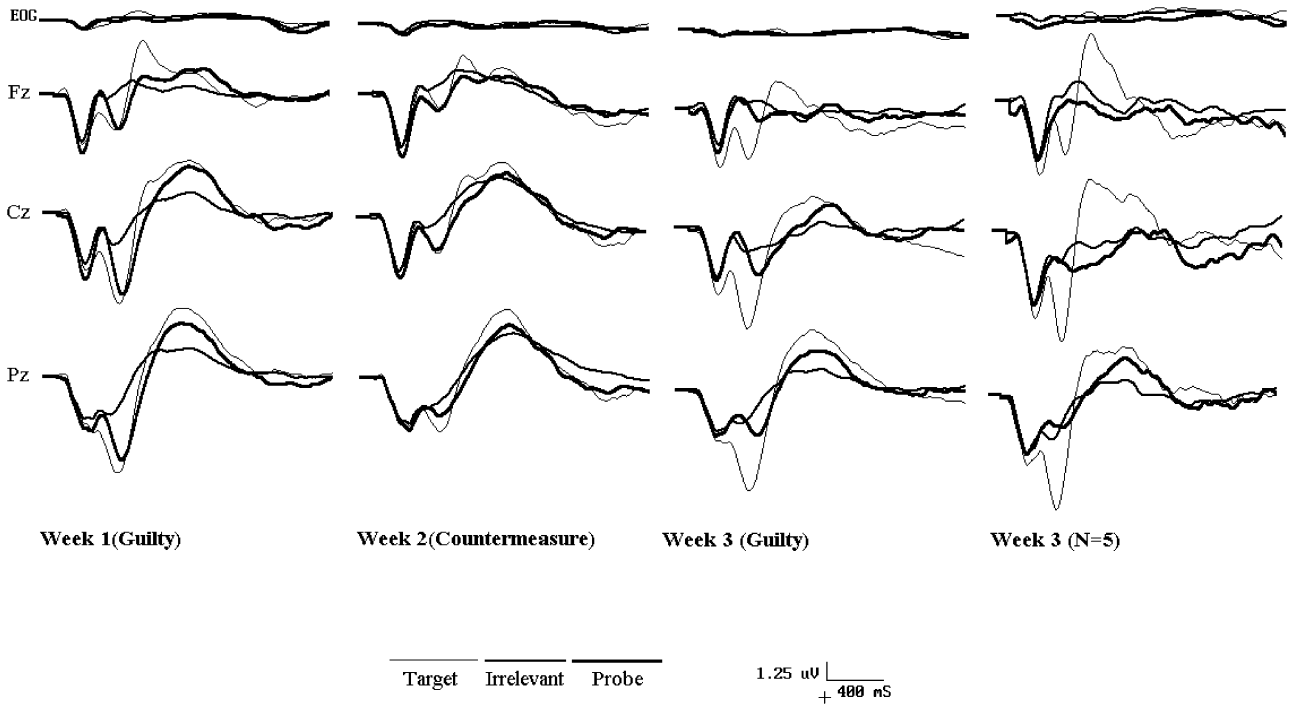


**Figure 4.** Leftmost column: These are the superimposed probe, target, and irrelevant grand averages from the first week of the one-probe experiment. Note probe = target $\gg$ irrelevant. Next column: Superimposed grand averages to probe, target, and irrelevant during the explicit use of the countermeasure, experimental group, Week 2. Probe and target P300 s are reduced (compare with Week 1 data at left). Next column: Superimposed probe, target, and irrelevant responses from third (Week 3) run of *all* experimental subjects. Rightmost column: Same as previous column, except these grand averages in Week 3 are over *only 5* experimental participants whose concealed information was undetected; the P300 for the probe is actually less positive at Pz than that to the irrelevant. Note that target clearly towers over probe. This figure in the countermeasure group illustrates the classical defeat pattern. Positive is down.

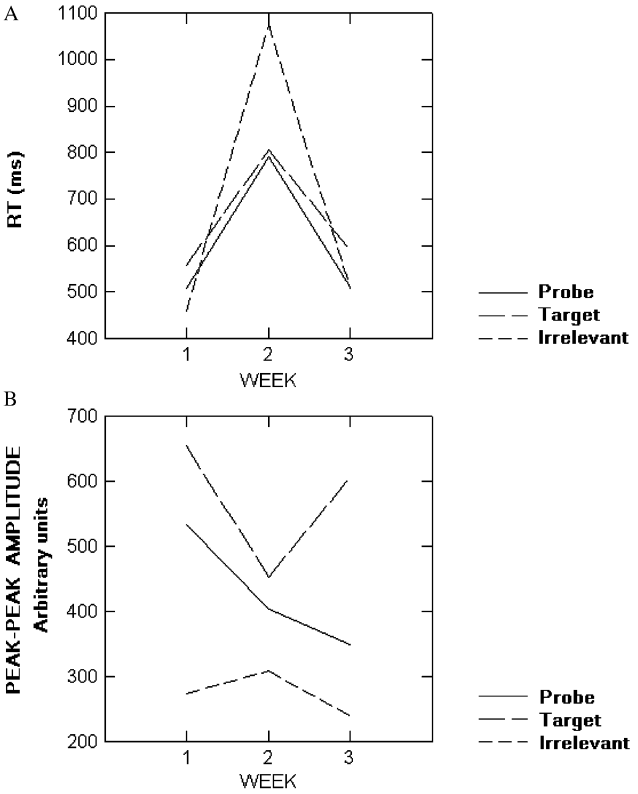


**Figure 5.** A: Reaction times (in ms) for three stimulus types across 3 weeks of Experiment 2. B: Computed mean PEAK–PEAK amplitudes, all participants, in computer units (10 μV = 409.6 units) for three stimuli over 3 weeks.

probe but not target responses slightly declined in the third week, though not nearly enough to defeat SIZE: Probe responses (PEAK–PEAK) were significantly greater than irrelevant responses in the third week, $p < .001$, and 9/10 of the controls were detected by SIZE in Week 3, as in Week 1. There was no significant difference between the probe-irrelevant differences (PEAK–PEAK) from Week 1 to Week 3, $p = .2$, nor between the PEAK–PEAK probe sizes, $p > .2$. Results with BASE–PEAK responses were virtually identical.

*ERP data. Quantitative*. Table 4 shows the detection rates for the experimental subjects using the bootstrap tests, FIT (90% confidence), SIZE (90% confidence, PEAK–PEAK), and RT-BOOT (95% confidence) over the 3 weeks of testing. It was noted that in the first week, 1 subject (of 14) was dropped due to a target error rate >10%; in the second and third weeks, another subject's data files for the bootstrap tests were irretrievably corrupted. Thus the final *N*s used for bootstrap tests for the three weeks are 13, 12, and 12, respectively. (RT data from all subjects were available for all weeks.) The major findings in the table are:

1. SIZE detects 3 of the subjects in week 1 which FIT misses.
2. Using the more sensitive SIZE test, explicit use of the countermeasure in Week 2 drops the hit rate from 92% to 50% ($p < .08$, McNemar), and from 69% to 25% with FIT ($p < .05$).
3. In view of the control data just presented, it is notable that in the third week, with the countermeasure not used (confirmed below with RT data, and by postexperiment interviews), the hit rate is still poor with SIZE (58%), and as we saw above in the qualitative ERP data, the 5 of 12 subjects who defeat the test do so with classical defeats, appearing like innocent subjects. Indeed, the FIT test in the third week detected only 25% of the subjects, the same number as when the explicit countermeasure was in use. It is reasonable to speculate that more intensive practice might result in a higher proportion of such defeats. Future research on the mechanism of these classical defeats could yield more effective countermeasure training methods.
4. The RT measure, RT-BOOT, which looks at the difference between probe and irrelevant RT, performs poorly throughout. This is not consistent with its performance in the first experiment in which at least in the simple guilty condition it performed very well ($A' = .95$. We could not compute $A'$ on guilty vs. innocent subjects in the second experiment, as there were no innocent subjects.). Because one difference between the first week of the second study and the guilty group in the first study involved the type of subject participating, it may be speculated that subject type is the source of discrepancy. However, type of concealed information (mock crime details vs. autobiographical data) also differed between the two experiments; the discrepancy may also be due to this variable (or to both variables). RT-BOOT is, of course, worthless in the second week of the second experiment when, as we will shortly see, the RTs for irrelevants (IRT in Figure 7) are more than doubled in most cases (see also Figure 5A) as the subject must recall which countermeasure to execute following each irrelevant stimulus.

Concerning *absolute* reaction times, Figure 5A shows the reaction times to the three stimulus types over the course of the 3 weeks in the experimental group.

Two points are implied by the RT data in Figure 5A: First, it is clear that after the dramatic increase in RTs during the explicit use of the countermeasure, the RTs drop down in the third week to the level of the first week, *providing clear evidence that the subjects followed instructions and did not use the countermeasure in the third week*. A 3 × 3 repeated-measures ANOVA on these scores yielded Greenhouse–Geiser corrected, significant effects for all variables: Week: $F(2,24) = 60.8$, $p < .001$; Stimulus type: $F(2,24) = 9.47$, $p < .003$; Interaction: $F(4,48) = 36.5$, $p < .001$. The interaction appears due to the greater increase in irrelevant than to other RTs in Week 2.

Second, a post hoc ANOVA comparing RTs just in Weeks 1 and 3 showed only one significant effect, that of stimulus type, $F(2,24) = 15.7$, $p < .001$, due to the expected effect of higher RTs to target than to other stimuli, probably related to the need to switch response buttons for this stimulus. The major implication here is that in the third week, when 42% of the subjects are *un*detected by the SIZE test and 75% *un*detected with the FIT test, RT would be no help in identifying the countermeasure-using test beater. Indeed, the 5 subjects who showed classical defeats of the ERP-based concealed information test had mean RTs to all stimuli in the *lower* half of the RT distributions of all subjects.

One would think, however, that RTs could at least be used to identify countermeasure users during the explicit use of the countermeasure. Indeed, post hoc *t* tests comparing RTs on Weeks 1 and 2 yield $t > 5$, $p < .001$ for all three stimulus types. Figure 7A shows the RT distributions for probe stimuli in Weeks 1 and 2, and it is seen that there is some slight overlap: Using the

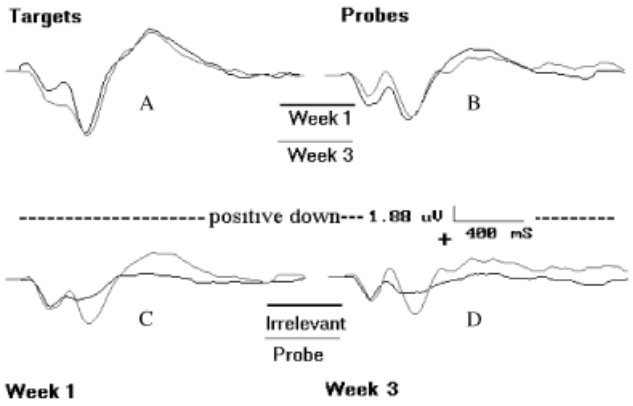


**Figure 6.** Grand averages, all Pz, are from the control experiment; data are shown from Weeks 1 and 3 only. In the top half, A and B are to be compared. A contains targets for Weeks 1 and 3 superimposed; no change is seen. In B, the probes from Weeks 1 and 3 are superimposed; there is a slight decrement in PEAK–PEAK P300 over time. In the bottom half of the figure, probes and irrelevants are superimposed in C, which is Week 1, and D, which is Week 3. The decrement in the probe over time is seen again, but the probe-irrelevant difference is still clear, unlike the classical defeats of Figure 4, column 4, Pz trace, from Week 3 of the experimental group. Positive is down.

same methods as in Experiment 1, we found that for probe stimuli, $A' = .98$ with 100% of the Week 1 subjects identified correctly as guilty with no countermeasure (0% false alarms) and 13 of 14 countermeasure users correctly identified (hit rate = 93%). The picture is similar for the target distributions (not shown) where two subjects from each week were misclassified.

In the irrelevant RT distributions from Weeks 1 to 2, however, there was in the present sample**,** *no* overlap, of actual reaction times from the 2 weeks (Figure 7B). Such data yield an $A' = 1.0$. Using the probe-minus-irrelevant RT as in RT-BOOT, only 1 of 28 subjects would have been misclassified, yielding $A' = .98$. This suggests that an absolute value of RT to the irrelevant stimulus or to the probe-irrelevant difference might be determinable, which could aid in identifying *explicit* use of the present countermeasure. However, if all the RTs (to probe and target also) in a particular subject were long, that would make the identification of a countermeasure user all the more difficult. Indeed in a follow-up study not reported further here, we found that it was simple for these subjects to voluntarily raise RTs to *all* stimuli such that there were no significant differences between RTs to any stimuli during explicit performance of the countermeasure. Moreover, it is possible that practice might reduce even the irrelevant RTs.

One does not, in the waveforms, see a classical defeat of the ERP test in Week 2 anyway, although the ERP analysis methods for detection are largely defeated in Week 2. The classic defeats of the test (in 5/11 subjects) are seen in the third week, when the RT distribution to the irrelevant stimulus is as shown in Figure 7C, along with the RT distribution in the countermeasure-naïve subject's first week.

The gross overlap, expectedly yields $t(12) = 1.15$, $p > .25$. (The corresponding $A'$ was close to .5.) The results were very similar to the other stimuli, as predictable from Figure 5A.

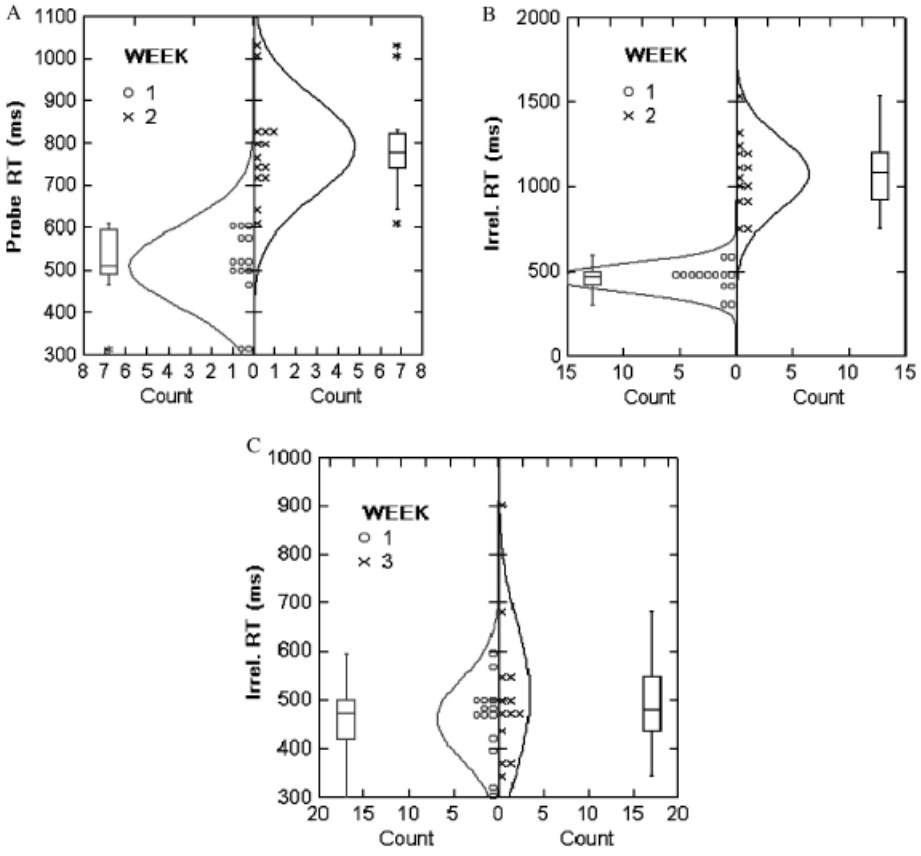


**Figure 7.** A: Reaction time distributions for probe stimuli in Weeks 1 (left) and 2 (right). B: RT distributions in Weeks 1 and 2 for irrelevant stimuli. C: RT distributions in Weeks 1 (left) and 3 (right) to irrelevant stimuli.

Figure 5B plots the computer-calculated PEAK–PEAK means to the three stimuli during the 3 weeks for all subjects.

A 3 × 3 repeated-measures ANOVA on these means found significant effects of week, $F(2,24) = 39.6$, $p < .001$, Greenhouse–Geiser, and the interaction of stimulus and week, $F(4,48) = 25.9$, $p < .001$, Greenhouse–Geiser, but the effect of stimulus type, surprisingly, was only a trend, $F(2,24) = 2.48$, $p = .11$, Greenhouse–Geiser. It should be noted that these data contain both detected and nondetected subjects in the second and third weeks, and are shown simply to more clearly convey the general trend of what was suggested in the ERP figures above. It is clear that the probe response is decreased from Weeks 1 to 2, and *stays reduced in Week 3*, unlike what happened in the control group. The irrelevant response is slightly increased from Weeks 1 to 2, then slightly declines in Week 3. The target response is depressed in Week 2 and, as noted above, is "released" in Week 3. With BASE–PEAK values, all three effects were $p < .001$, Greenhouse–Geiser, and the graphed data would look similar to Figure 5B. Post hoc $t$ tests comparing probes for just Weeks 1 versus 3 yielded $t(12) = 4.0$, $p < .003$ (BASE–PEAK) and $t(12) = 4.37$, $p < .002$ (PEAK–PEAK). For targets, $t(12) = 0.66$, $p > .5$ (BASE–PEAK) and $t(12) = 1.08$, $p > .3$ (PEAK–PEAK). In the control group, as already noted, probe remained greater than irrelevant, $p < .001$, in the third week in which SIZE still correctly diagnosed guilt in 9 of 10 cases.

## General Discussion

Referring to P300 ERPs used in guilty knowledge tests, Lykken opined: "Because such potentials are derived from brain signals that occur only a few hundred milliseconds after the GKT [*sic*; guilty knowledge test] alternatives are presented, and because as yet, no one has shown that humans can alter these brain potentials at will, it is unlikely that countermeasures could be used successfully to defeat a GKT [*sic*] derived from the recording of cerebral signals" (Lykken, 1998, p. 293).

Superficially considered, this expectation seems intuitively reasonable, although it clearly conveys a lack of awareness of the now sizeable literature on voluntary control of ERPs (Elbert, Rockstroh, Lutzenberger, & Birbaumer, 1984). Nevertheless, our major novel findings here are: (1) The *six-probe*, P300-based concealed information test paradigm *can* be defeated, and RT analysis cannot help with identification of any particular individual using explicit countermeasures, which can be made covert and undetectable, mental, or subtly physical. (2) The one-probe protocol can also be explicitly defeated; however, it remains possible that RT could be used to detect explicit countermeasure users in this protocol. This is based on our finding that there was no overlap of RT distributions to irrelevant stimuli from countermeasure and innocent conditions. However, as will be discussed further below, the one-probe protocol is subject to residual effects of countermeasure training on future occasions without explicit countermeasure use. Indeed, in view of the results of the third week of the *one-probe* study, we speculate that defeat of the six-probe paradigm—like the one-probe paradigm—might also not even require explicit countermeasures after having had a practice session with explicit countermeasures. This is an as yet untested empirical question.

The mechanism of the successful *explicit* countermeasure in both protocols is to covertly destroy their intended oddball paradigms whereby probes and targets are the sole rare, meaningful stimuli and thus lead to large P300 responses in comparison to the frequent, meaningless irrelevant stimuli. By executing covert responses to the irrelevant stimuli (one-probe paradigm) or stimulus categories (six-probe paradigm), the subject puts all stimuli on a more equivalent footing regarding probability and meaningfulness, and thus, the P300 s tend toward similar amplitude to all stimuli.

Regarding countermeasures in tests of deception based on autonomic responses, the National Research Council report stated, "A series of studies by Honts and his colleagues suggests that training subjects in…a combination of physical and mental countermeasures can substantially decrease the likelihood that deceptive subjects will be detected…" (National Research Council, 2003, p. 143; see also Honts et al., 1996). A particular concern according to the National Research Council report has involved the difficulty in detecting mental countermeasures in concealed information tests using autonomic responses. The degree of accuracy reduction of the present countermeasures in P300-based concealed information tests is similar to what one sees with the use of (somewhat different) countermeasures in autonomic response-based tests; however, at least with the P300-based one-probe protocol, *explicit* countermeasure use—mental or physical (both were used here)—may be detectable with RT observations, as discussed above.

In these studies, we were secondarily interested in comparing one- and six-probe protocols. However, from a theoretical perspective, the six-probe paradigm has theoretical difficulties not discussed prior to this report: One surmises that Farwell and Donchin (1991) chose to use six probes because the developer of the guilty knowledge test (Lykken, 1981) used six items in his original study of the polygraph-based guilty knowledge test. The point of using multiple items was as follows. If for one item there is a choice of five evaluated alternatives, then the probability of a chance hit on that item is $1/5 = 0.2$. The use of more orthogonal items reduces the multiplied fractional chance hit probabilities to, for example, 0.000064, with six items (0.2 to the 6th power). With a six-item test, even hitting on just three items yields $p = .08$ chance hit probability. The point is that in the format of a standard polygraph guilty knowledge test, one has *separate* responses to *each, individual probe*. This is *not* the case with the six-probe paradigms of Farwell and Donchin (1991) or Farwell and Smith (2001), which average all probe P300 s together. Let us suppose that an innocent subject produces a consistent P300 to just one and only one of the probes in a six-probe test—for whatever reason, such as actually recognizing this one guilty knowledge item through press leakage. The resulting average ERP to all probes should contain a small P300, as it is an average of five actual irrelevants and one probe. The target will reliably produce a large P300. The FIT method, as Farwell and Donchin (1991) used it, looks at cross correlations, which will, in calculating correlation coefficients based on standard scores, scale the amplitude differences between averaged probe and target away and likely declare guilt, not able to determine which or how many probe items were really recognized. The SIZE method might also find the probe greater than the irrelevant and also produce a false positive.

We would suggest that the use of repeated blocks of the one-probe paradigm, with a new probe on each block (and perhaps new sets of targets and irrelevants) is more likely to avoid the problems just described in the six-probe paradigm. Moreover, in the one-probe paradigm (unlike in the six-probe paradigm), at least in our sample, there was no actual overlap of RTs to

irrelevant stimuli between the naïve guilty and explicit countermeasure runs. One would have to run many more subjects to confirm the lack of overlap, but it is certainly conceivable that even if there is slight overlap, a cut-off RT could be determined to identify *explicit* countermeasure users. Moreover, in the six-probe paradigm, the target P300 is substantially larger than the probe and irrelevant responses during explicit countermeasure use, making the pattern of the three ERPs closer to that of the innocent pattern—the *classical defeat* profile. This was not true in the one-probe paradigm where all three P300 s were similarly small. Finally, it seems intuitively compelling that the six-probe protocol is a more demanding task than the one-probe protocol. This should reduce amplitudes in the latter protocol (Kramer et al., 1987). Consistent with this view, Table 4 shows higher detection rates in the one-probe paradigm (compare Table 2). However, subject type and type of concealed information (autobiographical vs. mock crime information) were confounded here with protocol, so that the question remains open.

A remaining serious challenge for P300 amplitude recognition indices in deception detection is suggested by the results of the Week 3 run with the one-probe paradigm. There it was found that 5 of 12 subjects still tested (SIZE) as innocent without using an explicit countermeasure (4 more were undetected with FIT). RTs for all subjects in Week 3 returned to Week 1 levels, confirming the nonuse of countermeasures and also making it impossible to use RTs to help identify former countermeasure trainees. Moreover, the test beaters in this third week produced classical defeats of the concealed information test, by presenting ERP averages indistinguishable from those of innocent subjects. Although we do not yet understand the mechanism of this effect, it strongly suggests that the P300 amplitude method may be utterly defeated by a good proportion of those that receive prior countermeasure training. Of course, further research could reveal the mechanism of this effect and allow for more targeted and explicit countermeasure training. Even if this mechanism were simply due to some kind of habituation effect—which it is not, as our control group failed to show the same changes over similar times—it would still pose a problem, whatever its basis, for field use; determined test beaters could practice often with an explicit countermeasure. How, in practice, would this be done?

To practice a countermeasure with a P300-based concealed information test, a potential or actual wrongdoer would have to have an idea of what probes would be used during the actual test. This might seem difficult in a forensic situation, but, in fact, common sense suggests that the criminal would recall more real, salient—for him—details of the crime scene in which he was involved than would anyone else. He may easily be able to generate more probes than authorities, who can only guess at what *should* be remembered; the criminal *knows* what he remembers. Now in practice, such a criminal would likely have to consult with an informed professional who has the expertise to train him. It is hopefully not likely that such professionals in the United States (members of the Society for Psychophysiological Research) would become involved in such marginal activity. So assuming that the domestic scientific community is reasonably free of criminals, the domestic forensic situation may be safe for criminals lacking the intelligence and resourcefulness to make use of published papers such as the present one.

The counterterrorism scenario is a much different matter. As already noted, Farwell and Smith (2001) and Farwell's web site have strongly promoted the use of the P300 concealed information test as a counterterrorist tool. These sources reported that various U.S. security professionals were shown to possess concealed information using the P300 paradigm. The generalization implied is that a member of a foreign terrorist organization also has concealed information ("guilty knowledge details") about his organization: frequently used acronyms, names of lower level leaders, training camp layouts, and so on. Assuming our security agencies also have some of these details, a concealed information test could be composed for would-be or actual terrorists. In this situation, it is clear that intelligent terrorists certainly can guess well ahead of the test what probes may be used, and so practice the countermeasure technique. Because these individuals will likely come from a different culture and society whose professional members could be sympathetic with the goals of the foot soldiers, or who could be coerced into cooperating, obtaining professional training might not prove to be difficult.

Finally, we have shown that the method of analysis appears to interact with subject type. Subjects cooperative with experimenters are detectable just as well with SIZE as with FIT, but truly naïve subjects, a category which would likely include those encountered in the field, are not well detected with FIT. This difference was related to the possibility of phase differences between the ERP responses to targets and probes. It might be countered that a latency adjustment procedure could be readily used on all probe, target, and irrelevant waveforms, and that as long as the same algorithm is utilized on all suspects, the FIT procedure may be shown to work well. This is, of course, an empirical question, and there has been no research on the matter. In fact, research may reveal that such latency adjustments could improve the probe-irrelevant correlation more than the probe-target correlation, thus leading to false negatives. Moreover, the FIT procedure was initially based on the notion of greater probe-target resemblance overall than probe-irrelevant resemblance in guilty persons. It would seem that latency adjustment procedures must distort the genuine appearance of the basic data set, which gets away from the notion of appearance comparison. As noted, *practically*, this might not make a difference in detecting deception. In the absence of more study, the data available thus far are based on the FIT method as in Farwell and Donchin (1991), and that method is outperformed here by the SIZE method.

## REFERENCES

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

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.

Coonts, S. (2003). *Liberty*. New York: St. Martin's Press.

Elbert, T., Rockstroh, B., Lutzenberger, W., & Birbaumer, N. (Eds.). (1984). *Self-regulation of the brain and behavior*. Berlin: Springer-Verlag.

Ellwanger, J., Rosenfeld, J. P., Sweet, J. J., & Bhatt, M. (1996). Detecting simulated amnesia for autobiographical and recently learned information using the P300 event-related potential. *International Journal of Psychophysiology*, *23*, 9–23.

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. *Psychological Bulletin*, *75*, 424–429.

Honts, C. R., Amato, S. L., & Gordon, A. K. (2001). Effects of spontaneous countermeasures used against the comparison question test. *Polygraph*, *30*, 1–10.

Honts, C. R., Devitt, M. K., Winbush, M., & Kircher, J. C. (1996). Mental and physical countermeasures reduce the accuracy of the concealed information test. *Psychophysiology*, *33*, 84–92.

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

Kramer, A. F., Sirevaag, E. J., & Braune, R. (1987). A psychological assessment of operator workload during simulated flight missions. *Human Factors*, *29*, 145–160.

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

Lykken, D. T. (1981). *A tremor in the blood*. New York: McGraw-Hill.

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

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

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

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

Rosenfeld, J. P., 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., 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., Ellwanger, J. W., Nolan, K., Wu, S., Bermann, & Sweet, J. J. (1999). P300 scalp amplitude distribution as an index of deception in a simulated cognitive deficit model. *International Journal of Psychophysiology*, *33*, 3–20.

Rosenfeld, J. P., & Ellwanger, J. W. (1999). Cognitive psychophysiology in detection of malingered cognitive deficit. In J. J. Sweet (Ed.), *Forensic neuropsychology: Fundamentals and practice*. Lisse, Netherlands: Swets and Zerlanger, Publishers.

Sasaki, M., Hira, H., & Matsuda, T. (2002). Effects of a mental countermeasure on the physiological detection of deception using P3. *Studies in the Humanities and Sciences*, *42*, 73–84.

Seymour, T. L., 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*, 173–180.

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

(Received January 17, 2003; Accepted August 27, 2003)