Psychophysiology, 45 (2008), 286–298. Blackwell Publishing Inc. Printed in the USA.

DOI: 10.1111/j.1469-8986.2007.00615.x

The role of psychophysiology in forensic assessments: Deception detection, ERPs, and virtual reality mock crime scenarios

RALF MERTENS AND JOHN J.B. ALLEN
Department of Psychology, University of Arizona, Tucson, Arizona, USA

Abstract

Few data are available to address whether the use of ERP-based deception detection alternatives have sufficient validity for applied use. The present study was designed to replicate and extend J. P. Rosenfeld, M. Soskins, G. Bosh, and A. Ryan's (2004) study by utilizing a virtual reality crime scenario to determine whether ERP-based procedures, including brain fingerprinting, can be rendered less effective by participant manipulation by employing a virtual reality crime scenario and multiple countermeasures. Bayesian and bootstrapping analytic approaches were used to classify individuals as guilty or innocent. Guilty subjects were detected significantly less frequently compared to previous studies; countermeasures further reduced the overall hit rates. Innocent participants remained protected from being falsely accused. Reaction times did not prove suitable for accurate classification. Results suggested that guilty verdicts from ERP-based deception detection approaches are likely to be accurate, but that innocent (or indeterminate) verdicts yield no useful interpretation in an applied setting.

Descriptors: ERP, Deception detection, Guilty knowledge, Virtual mock crime

Conventional field polygraph examinations, based on the control-question technique (CQT), suffer from many limitations and have been widely criticized in the scientific literature (Iacono, 2000; Iacono & Lykken, 1997; Iacono & Patrick, 1997; Lykken, 1987; National Research Council, 2003; U.S. Office of Technology Assessment, 1983). Psychophysiologists have thus examined alternative methods of deception detection, including an approach that assesses recognition for key facts, known as the guilty knowledge technique (GKT; Lykken, 1959), and the use of alternative physiological responses (Bashore & Rapp, 1993; Rosenfeld, 1995), including measures of cortical activity such as event-related potentials (ERPs). Unlike conventional polygraph approaches using the CQT that assess emotional arousal associated with lying, ERP-based alternatives most often utilize the GKT approach to assess memory for salient aspects of a situation that would only be known to a perpetrator and few others such as investigators or witnesses.

We thank Lauren Crawford, Nicholas Culp, and Mirijam Rupp for their assistance in bringing this study to completion. Special thanks go to Dean Klimchuk and Roman Mitura at Digital Media Works (www.dmw.ca) for their technical expertise in translating our ideas into a flexible and highly adaptable virtual environment. Part of this study was supported by a Homeland Security grant provided by the Office of The Vice President for Research at The University of Arizona. Portions of this study were presented at the annual meeting of the Society for Psychophysiological Research, October, 2003, Chicago, IL.

Address reprint requests to: John J.B. Allen, Department of Psychology, P.O. Box 210068, Tucson, AZ 85721-0068, USA. E-mail: jallen@u.arizona.edu

In the service of assessing memory, ERPs may be a promising approach. Multiple ERP components have been found sensitive to the recollection of past experiences. The P3, a cognitive component of the ERP that appears relatively quickly after stimulus presentation (i.e., in the range of 300–900 ms) has been especially interesting, as it has been associated with memory and context updating (Donchin & Coles, 1988) and can be elicited in the absence of an overt response by a participant as long as the stimulus presented is attended to (Mertens & Polich, 1997). The P3 has been successfully employed in a limited number of ERP-based deception detection studies (Allen, Iacono, & Danielson, 1992; Farwell & Donchin, 1991; Farwell & Smith, 2001; Rosenfeld, Angell, Johnson, & Qian, 1991). These studies correctly identified individuals in 89% (Rosenfeld et al., 1991), approximately 90% (Farwell & Donchin, 1991), and 95% (Allen et al., 1992) of cases, utilizing different statistical approaches (bootstrapping of peak-to-peak amplitude, bootstrapping of cross-correlations, and Bayesian analysis, respectively).

In a prototypical ERP-based deception detection paradigm, participants are presented with three types of items (i.e., probe, target, distracter) intermixed with each other. Probes and targets are presented infrequently (e.g., 10%–15% of the time) whereas distracter items are presented frequently (e.g., 70%–80% of trials). Probes refer to crime-relevant items that are known only to the perpetrator or others that have familiarity with the crime (e.g., witnesses, investigators) and that should elicit a large P3 if recognized as distinct, rare, and task relevant (Johnson, 1986). Target items are stimuli that are taught to everybody and that participants are required to respond to when presented. These

target items should also be recognized as distinct, rare, and task relevant and therefore elicit a large P3 in all participants. The use of target items serves two purposes: First, it ensures that participants attend to and process all information presented instead of ignoring stimuli. Poor responding to target items clearly suggests that participants do not attend to stimuli (excluding factors such as poor vision, etc.); second, it provides the investigator with a prototypical P3 component in response to a learned item—for each participant—to which the probe can be compared. The third type of item, distracter items, is not important to the task itself but provide a comparison condition that should elicit a small or no P3, and thus provide a template ERP in response to unfamiliar items.

Comparing the ERPs to probe, target, and distracter items then allows for an assessment of whether an individual has crime-relevant knowledge. Innocent individuals, for example, when confronted with a probe, should produce ERPs highly similar to the distracter items, as both stimuli should be completely unfamiliar to the innocent individual. For guilty participants, by contrast, probe items should produce ERPs highly similar to those in response to recognized target items.

Tasks are usually presented by requiring participants to lie about details of a mock crime (Farwell & Donchin, 1991), to deny autobiographical information (Miller & Rosenfeld, 2004; Rosenfeld et al., 1991; Rosenfeld, Rao, Soskins, & Miller, 2003), or to lie about materials acquired during a list-learning task (Allen et al., 1992). Controlled ERP investigations in conditions approximating field conditions have not been conducted, but more realistic environments emulating field settings may be helpful to further validate the use of ERPs in such settings. Recent advances in computer gaming technologies have created the possibility of developing highly realistic virtual environments (VEs), and such technology has been utilized to treat a variety of psychiatric conditions such as anxiety disorders (Kuntze, Störmer, Mager, Müller-Spahn, & Bullinger, 2003; Lee et al., 2002; Wiederhold, Jang, Kim, & Wiederhold, 2002). Furthermore, VEs would make it possible to measure brain function during the encoding of a crime scene, using fMRI, EEG, or other techniques. VEs thus make it possible to easily replicate experiments across multiple study sites and for crime scenes to be reconstructed realistically. With improving artificial intelligence, environments could eventually include artificial actors, allowing a broader range of hypothesis testing. The present study thus combined aspects of a VE with aspects of the more commonly used mock-crime procedure.

Commercial Use of ERP Deception Detection Procedures

Based on their study (Farwell & Donchin, 1991), Farwell patented an ERP-based deception detection procedure and apparatus (Farwell 1994, 1995a, 1995b) and commercially developed this deception detection procedure, which he termed "brain fingerprinting." Promoting this technique, Farwell and others have claimed that this approach is 100% accurate (Farwell & Smith, 2001; Feder, 2001; Scheeres, 2001) despite little empirical evidence (Rosenfeld, 2005) from scientific trials investigating the technique, despite strong contradictory evidence against one such claim of innocence (Zirinsky, 2002), and despite doubts raised by federal agencies about the validity of the method (U.S. General Accounting Office, 2001).

A potential weakness of ERP-based measures is their possible vulnerability to countermeasures. Because autonomic measures have been found susceptible to physical and mental countermeasures (Ben-Shakhar & Dolev, 1996; Elaad & Ben-Shakhar, 1991; Honts, Devitt, Winbush, & Kircher, 1996; Honts, Hodes, & Raskin, 1985; Honts, Raskin, & Kircher, 1987; Iacono, Cerri, Patrick, & Fleming, 1992; Kubis, 1962), it is possible that ERPs may be influenced using similar techniques. Although mental countermeasures (e.g., counting backward in sevens) did not affect P3 amplitude or latency (Sasaki, Hira, & Matsuda, 2001), a novel countermeasure developed by Rosenfeld, Soskins, Bosh, and Ryan (2004) using mental and physical countermeasures appeared to be effective. Instead of attempting to manipulate the P3 in response to probe or target items, Rosenfeld et al. taught participants to enhance the significance of distracter items only, which in turn increased P3 amplitude in response to these items. This manipulation thus decreased the relative P3 amplitude differences between probe and distracter items, making probes less detectable. Rosenfeld found that employing a consecutive set of simple behaviors each time a distracter item occurred was effective. Utilizing one of four behaviors to each distracter item in sequence (i.e., behavior one on one trial, behavior two on another trial, etc.) lowered detection rates dramatically. Using (a) an imperceptible physical response (e.g., applying pressure to a finger), (b) another imperceptible physical response (e.g., wiggling of toes), (c) imagining an emotion-evoking event (e.g., being slapped in the face), and (d) exercising no response for consecutively appearing distracter items reduced the detection rate to as low as 18%. By contrast, guilty participants not trained in this countermeasure were detected as high as 82% of the time in a typical mock-crime assessment and as much as 92% of the time when tested simply for recognition of their birthday.

Rosenfeld et al.'s (2004) findings are important and raise serious reservations about the applicability of ERP deception detection procedures for field use. On the other hand, a replication utilizing a highly realistic crime scenario and expanding the range of countermeasures employed could address the generalizability of these findings. Rosenfeld et al.'s creative and novel approach, however, left unexplored the question of whether countermeasures in response to target items could also be effective. Redirecting attention to target items can affect P3 amplitudes to other stimuli as well, as it changes the relative salience of the various item types. Moreover, increasing amplitude in response to target items would reduce the significance of probe-related amplitude increases, reducing the probability of being detected using some analysis approaches (e.g., Farwell & Donchin, 1991). Furthermore, a simple countermeasure in response to target items appears less complicated compared to the complex behavioral sequence in response to distracter items. Lastly, analysis performed by Rosenfeld et al. concentrated on bootstrapping techniques but did not implement the Bayesian analysis as a comparison despite its excellent specificity and sensitivity (Allen et al., 1992). The present study therefore attempted to replicate and extend Rosenfeld et al.'s study by employing a highly realistic virtual reality crime scenario, multiple countermeasures, and Bayesian and bootstrapping analytic approaches to classify individuals as being guilty or innocent.

Methods

Participants

A total of 79 participants, 38 male (mean age = 19.00 years, $SD = 3.09$) and 41 female (mean age = 19.34 years, $SD = 4.43$) undergraduate students of the University of Arizona, are

represented in this data set.[1] Participants were native English speakers, not regular substance users, currently not under the influence of drugs, alcohol, or psychotropic medications, and free of psychological disorders and/or disorders known to affect the central nervous system (e.g., previous head injury resulting in the loss of consciousness). Participants were able to navigate a virtual environment with the aid of a computer keyboard and mouse and had normal or corrected-to-normal vision. Participants received $10 for each hour of participation with the possibility of earning an additional $100 bonus if successful in their respective tasks.

Procedures and Apparatus

Qualified participants were randomly enrolled in one of five experimental conditions consisting of (1) a standard guilty group that engaged in the virtual reality mock crime and was tested without special instructions for countermeasures, (2, 3, and 4) three countermeasures groups (defined below) that engaged in the mock crime and that were taught strategies that might impair the ability of the ERP procedure to detect their knowledge of crime relevant information, and (5) an innocent group that learned to navigate the three-dimensional computer environment with crime-relevant items removed, but that were tested nevertheless for knowledge of crime-relevant items. This innocent control group has good external validity as oftentimes innocent suspects may be familiar with aspects of a crime scene but have no knowledge of specific crime-relevant information. Suspects are often identified based on circumstantial evidence, such as being in the proximity of a crime or being familiar with a business or home that was involved in the crime.

All efforts were made to maintain maximal realism within a laboratory setting while heeding guidelines set forth by the Behavioral Science Institutional Review Board. Participants rated the study as highly realistic, with 75% of participants rating the study at 7 or above on a 10-point questionnaire, with 10 indicating the highest level of realism students could envision. As an testiment to its realism, one forgetful student, scheduled to participate in the study, contacted police after receiving an e-mail reminding him to carry out his "mission."[2] Level of motivation to perform to the best of their ability was rated at 7 or above on the same 10-point scale by 91% of participants.

Mock Crime Procedure

Guilty and innocent participants eligible for the study received details about their respective task in the form of a mission plan the same day as enrollment. Innocent participants were instructed to enter an unoccupied office and log on to a computer and explore the VE for a period of 6 min. Guilty participants were instructed to enter an unoccupied office, usually off-limits to undergraduate students, and to access a password-protected computer using a numeric code. Following their log-on, guilty participants then navigated a highly realistic VE, resembling a large, eight-room apartment (Figure 1), and within 7 min they were to retrieve several items using a computer mouse and keyboard. The VE was developed by a Canadian firm (www.dmw.ca) for the first author, based on the commercially available Quake III gaming engine (www.idsoftware.com) and it included several features that added to the realism. For example, an inconspicuously placed timer reminded participants to complete their task within a time limit, which, considering the size of the virtual apartment, was not necessarily easy. Similarly, participants were likely surprised when the theft of a gun was accompanied by the loud, unexpected noise of a bowl that was broken in the process.

Figure 1. Sample screenshot of an eight-room virtual apartment created based on a modification of a commercially available gaming engine (www.idsoftware.com) created by Scaled Environments at Digital Media Works (www.dmw.ca).

Instructions for guilty participants included 11 critical items (i.e., probes) they had to learn verbatim prior to executing their task (e.g., the numeric code to break in, items to retrieve, etc.). One more probe (i.e., broken bowl) was created while guilty participants carried out their mission. All 12 items used for guilty participants are listed in the Appendix. Instructions for innocent participants were modified to exclude crime-relevant items; instead, participants were told to enter an office, log on to a computer, and simply navigate the virtual environment, which did not contain the crime-relevant stimuli (e.g., they used a different numeric code, no item to steal was present, etc.). Participants were monitored via a hidden camera to ensure that they carried out their respective tasks as instructed, with the camera revealing their entrance to the office, their use of the computer, and their exit. Immediately after exiting the office, participants were intercepted by a confederate who accused them of having been in the office without authorization and escorted them in silence up three flights of stairs to a laboratory.

Participants were interrogated using a scripted format by investigators blind to the participant's condition. The interrogation was designed to reemphasize probe items, by presenting them among a backdrop of items so that subjects would not be able to know which item was, in fact, the crime-relevant alternative (e.g., "Assuming your innocence, you wouldn't know that the perpetrator received the mission plan in a blue envelope; how about a green envelope? A red envelope maybe? Does a yellow envelope seem familiar? Was it maybe received in a black envelope?"). Although probe items were presented at this point among all the possible choices, they were in no way identified, so that innocent participants would not attach any special significance to them

[1] A total of 7 participants were excluded at various stages of the experiment and were not included in the present data set. Four participants broke their respective blinds, 1 participant was observed using notes carrying out their "mission," and 2 participants were unable to enter the virtual environment.

[2] After learning about the incident in the campus newspaper (Halperin, 2002), the first author contacted the student and campus police to clarify the purpose of the e-mail.

and guilty participants would only find them significant by virtue of having participated in the mock crime. Guilty and innocent participants knew to deny any involvement, and anybody who broke the blind at any point in time during the experiment was excluded from the study (see footnote 1).

Following the interrogation, participants learned a list of 12 target words to a criterion of perfect recall and completed a recognition task following the sensor placement approximately 20 min later to ensure that target words were as well learned as probe items for guilty participants.

Electrophysiological Recording

Electrophysiological data were recorded using a digital Synamps amplifier (Compumedics Neuroscan, El Paso, TX) with signals amplified by a factor of 500 and digitized continuously at 250 Hz. Online filtering ranged from a high pass of 0.1 Hz to a low pass of 100 Hz. Midline electrodes (Fz, Pz) and mastoid electrodes (A1, A2) were referenced online to Cz with a forehead ground. Electrode impedances were maintained at 5 KΩ or less, with interelectrode differences at 1 KΩ or less. Electrooculographic (EOG) activity was monitored by placing a bipolar montage at the inferior and superior orbit of the right eye and a second bipolar montage at the outer canthi of each eye. Files were digitally filtered with a 12.5-Hz, 96-dB-per-octave low-pass filter and then corrected for ocular artifact (Semlitsch, Anderer, Schuster, & Presslich, 1986) using the bipolar vertical and horizontal EOG channels. Following eyeblink correction, epochs with EOG deflections exceeding ± 75 μV from baseline were rejected as a subsequent precaution against including ocular artifacts such as eyeblinks. Files were epoched from -250 ms before stimulus onset to 1750 ms after stimulus, linearly detrended, baseline corrected to a prestimulus baseline of -250 to 0 ms, and finally re-referenced off-line to linked mastoids. Epochs free of artifacts or incorrect responses were then averaged by condition to create ERPs for each stimulus type.

Countermeasures

Participants in each of three countermeasures groups received specific instructions with their mission plan to carry out countermeasures designed to interfere with ERP discrimination of familiar from unfamiliar items. The intent of these countermeasures was to alter the subjective significance of various stimuli. These instructions educated participants about the fact that there would be probe, target, and distracter stimuli and told them how to identify them in the string of stimuli so as to utilize the countermeasure. The first countermeasures group (CM1) was instructed to think about being slapped by a bully whenever a target item was displayed, thus increasing target significance. The second group (CM2) was instructed to apply pressure to their toes whenever a target item was displayed, again increasing target significance. The third group (CM3) was instructed to enhance the significance of distracter items by responding in alternating ways to them by thinking about being slapped by a bully after the first distracter, tightening of the sphincter muscle after the second distracter, applying pressure to toes after the third distracter, and not responding at all to the fourth distracter (cf. Rosenfeld et al., 2004). The entire CM3 sequence was repeated after each sequence of four distracters until all stimuli were presented.

Stimuli

Seventy-two word stimuli were selected for this study and consisted of 12 crime-relevant items (i.e., probes; see the Appendix), 12 items associated with the crime scene but not relevant for the crime (i.e., targets), and 48 distracter items (i.e., unlearned), 2 for each probe and target. Probes were semantically matched to target and distracter items and were similar in word length. A pilot study revealed that probes were equally salient to target and distracter items when a group of 20 participants were asked to select items they thought were associated with an imaginary crime that was recently committed. Statistical comparison of probes, targets, and four lists of distracters clearly indicated that participants had no preference for a particular stimulus type, $F(5,50) = 0.02$, n.s.). Furthermore, word lists did not differ in word frequency, $F(5,50) = 0.153$, n.s.) for items that were able to be identified using the Krucera–Francis written word frequency database available on-line at http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm.

Stimuli were presented in a sound-dampened chamber and displayed centrally on a CRT monitor at a viewing distance of 150 cm. Maximal vertical and horizontal visual angles were at 0.38° and 2.06°, respectively. All word stimuli and instructions were presented using DMDX software (available at http://www.u.arizona.edu/~kforster/dmdx/dmdx.htm). Stimuli were presented at a rate of one word every 3000 ms, with each word present on the screen for 1000 ms. Participants indicated recognition of targets by pressing a button with the thumb of their dominant hand, and they pressed with the thumb of their nondominant hand for all other items (i.e., probes and distracters). Double responses (i.e., both buttons pressed) were excluded from the analysis because of the concomitant distortion of P3 amplitude and latency.

Stimuli were presented in blocks with the serial position randomized within and across blocks. Each block consisted of three probes, three targets, and 12 distracters (2 matched to each probe and target presented). As previously mentioned, probes within each block were semantically matched to targets and distracters, and all items were similar in word length. After presentation of the 18 items within a block, a new block of randomized items was presented. Four blocks presented all stimuli (i.e., 12 probes, 12 targets, 48 distracters), yielding a total of 72 stimulus presentations, with each stimulus being presented twice. Serial position of the four blocks and 72 items within each block was randomized and presented again. Following a self-paced break, the aforementioned sequence was repeated again. The total procedure yielded 288 presentations or 48 possible ERP trials per stimulus type (i.e., probe, target, distracters 1–4), with each individual stimulus being repeated four times during the entire assessment. Trials were omitted from analysis for artifacts (other than blinks) and for incorrect responses. Across participants, 92% of probe trials, 73% of target trials, and 91% of distracter trials were included in the analysis.

Analysis

Group analyses were employed to determine whether the procedure was successful in producing a prototypical P3 amplitude and latency pattern. P3 amplitude was taken as the most positive deflection in a 350–950-ms search window. Classification of individual participants was accomplished using three approaches: bootstrapped correlations (Farwell & Donchin, 1991), Bayesian analysis (Allen et al., 1992), and bootstrapped peak-to-peak amplitude difference (Soskins, Rosenfeld, & Niendam, 2001), each of which is detailed below. The statistical outcomes were also compared to each other utilizing receiver operating charac-

teristic (ROC) curves to assess whether one particular approach proved superior.

Bayesian analysis. This analytic procedure, derived from that originally developed by Thomas Bayes (1763) involves the combination of several indicators, each known to differentiate two conditions from each other, in order to enhance classification accuracy. Allen et al. (1992) adopted this approach to aid in the identification of ERP waveforms in response to learned and unlearned word lists by combining different features of waveforms (e.g., P3 amplitude and first and second derivatives of P3 amplitude) that have been shown to be effective in differentiating familiar from unfamiliar items based on information contained in these waveforms. In short, classification is achieved by computing the probability that a given ERP is in response to recognized items, given a pattern of indicators. The computed probability is high if all indicators suggest that the ERP was in response to recognized items, whereas disagreement among indicators would decrease this probability. Various indicators that were identified as useful in the Allen et al. study were converted to within-participant z scores to reduce irrelevant individual differences and enhance the pattern of response for each participant. As an example, the z score representation of P3 amplitudes from the current study is presented in Figure 2, which summarizes the expected finding that familiar items produce ERPs with larger P3 amplitudes. Z score cutpoints that maximally differentiated learned from unlearned items in the validation study of Allen et al. were then applied to each indicator in the present study, with the result being that each indicator now suggested that a list of items was familiar or not familiar to the participant. Based on the sensitivity and specificity of these indicators as established in the validation study of Allen et al., the indicators were then combined in a simple Bayesian fashion to derive the probability that a given list of items appeared familiar given a particular constellation of indicators (e.g.,Indicators 1, 2, and 5 suggested familiarity, but indicators 3 and 4 did not). For a detailed explanation of the computational approach, the reader is referred to Allen et al. (1992).

Bootstrapped correlation between item types. Bootstrapping (see Wasserman & Bockenholt, 1989) provides a method for generating a distribution of values for any measure for a given participant, thereby allowing for a statistical estimation of the replicability of a given finding even when multiple replications have not been obtained. In this application, repeated sampling—with replacement—from the raw epochs is used to create averaged waveforms. Upon each iteration, the relevant measures are obtained from the averaged waveforms, and then the process is repeated 100 times. After all iterations, there will be a distribution of values of the relevant measures, and one can determine whether the predicted outcome is robust across replications.

The rationale for the bootstrapping approach taken by Farwell and Donchin (1991) assumes that infrequently displayed, learned items (i.e., probe and target) produce larger P3 responses as compared to more frequently displayed unlearned distracter items. Cross-correlations for probe and targets in guilty participants should be larger when compared to the cross-correlations of probe and distracter items, whereas the reverse pattern should emerge in innocent participants. By using "double-centered" correlations, first subtracting the grand mean ERP across conditions from each ERP, the prediction is strengthened in that for guilty participants the probe–target correlation should be positive and the probe–distracter correlation should be negative. The bootstrapping procedure thus created ERP averages for each iteration, using repeated sampling with replacement from the pool of all available sweeps for each item type (i.e., probes, targets, distracters). For each of 100 iterations, a probe–target and probe–distracter double-centered cross-correlation was computed. The decision rules of Farwell and Donchin were used: (1) if the probe–target correlation exceeded the probe–distracter correlation (suggesting guilty knowledge for probe items) for more than 90% of the iterations, the subject was classified as guilty;

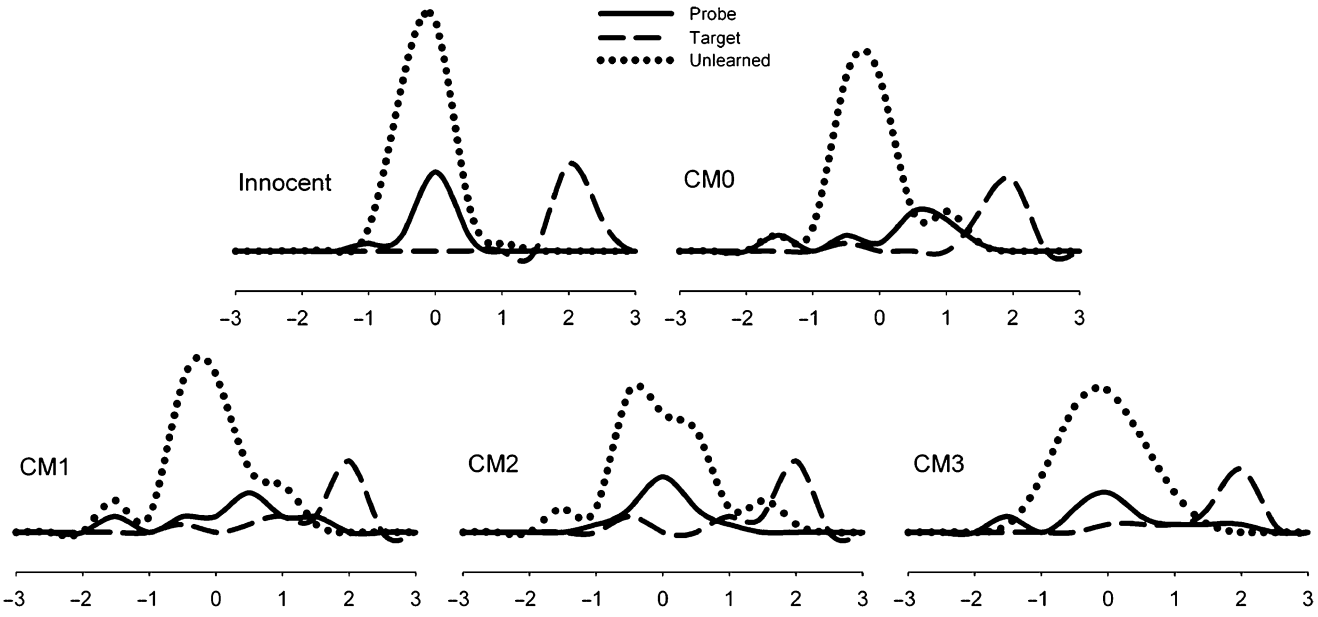

Figure 2. Frequency distribution of z scores for P3 amplitude across five experimental conditions for the 79 participants. Targets were commonly associated with larger z scores as compared to other items. Probes and targets appeared dissimilar in the standard guilty condition (CM0), however. Probes were similar to distracter items in the three countermeasures groups (CM1, CM2, CM3).

Table 1. *Mean Values (SD in Parentheses) for Reaction Times and Incorrect Response Rates for Three Stimulus Types across Five Experimental Conditions*

	Reaction Time (ms)			Incorrect Responses (%)		
Condition	Probe	Target	Unlearned	Probe	Target	Unlearned
CM0	824 (216)	833 (194)	727 (211)	1.5 (1.6)	9.5 (5.3)	0.3 (0.6)
CM1	777 (138)	855 (237)	702 (147)	1.7 (1.9)	11.2 (5.1)	0.3 (0.6)
CM2	739 (127)	853 (158)	617 (100)	1.7 (2.3)	13.5 (8.9)	0.4 (0.7)
CM3	832 (202)	840 (174)	743 (186)	2.1 (2.2)	12.9 (9.0)	0.6 (1.3)
INN	722 (123)	770 (94)	678 (102)	0.6 (0.6)	7.8 (5.7)	0.3 (0.8)

(2) if the probe–target correlation exceeded the probe–distracter correlation for fewer than 30% of the iterations (suggesting lack of guilty knowledge), the subject was classified as innocent; (3) if the probe–target correlation exceeded the probe–distracter correlation on 30% to 90% of the iterations, the subject was classified as indeterminate. Using these decision rules, Farwell and Donchin were able to correctly classify 90% of guilty participants and 85% of innocent participants, leaving 12.5% of the sample classified as indeterminate, but with no false positives and no false negatives.

Bootstrapped amplitude difference. This is a variant of the aforementioned bootstrapped correlation technique, but one that compares P3 amplitudes instead of cross-correlations. As with Farwell and Donchin's (1991) approach, each iteration involves random sampling with replacement a set of accepted single sweeps from probe and distracter items, respectively. However, instead of computing cross-correlations, this process computes the average P3 amplitude of probe and distracter items during each of 100 iterations. Then, the P3 of the distracter average is subtracted from the P3 of the probe average for each of the iterations to create a distribution of P3 amplitude differences. Using the mean and standard deviation of this distribution, a z score can be computed, with a z score greater than 1.65 standard deviations taken to indicate with 95% confidence that the probe produces a significantly larger P3 than the distracter items for that participant, thus resulting in a guilty verdict. A z score of 1.65 corresponds to the 95th percentile of a distribution, which results in a 95% directional confidence interval because of the a priori hypothesis that probes produce significantly *larger* P3 values than distracters; if distracters produced significantly larger P3 amplitudes than probes, such a participant would not be deemed guilty. Soskins (Soskins et al., 2001) has reported that the peak-to-peak method, subtracting the negative peak subsequent to the P3 from the P3 amplitude, enhances detection of guilty participants. The peak-to-peak method was thus used in the present study.

Results

Although classification of individual participants as guilty or innocent on the basis of the ERP-based assessment was of primary interest, group level comparisons are presented first to provide an overview of the effects of interest. In cases where repeated measures factors in an ANOVA had more than two levels, Greenhouse–Geisser correction for violations of the sphericity assumption were used. In all cases, original degrees of freedom are presented, along with the epsilon-corrected p values.

Group-Level Analysis

Behavioral data. Table 1 presents reaction times (RTs) for the 79 participants, separately by experimental group. It was expected that probe and target RTs would be significantly longer than those for distracter items for guilty participants. The main effect of item type, $F(2,148) = 94.9$, $p < .001$, was qualified by an Item Type × Group interaction, $F(8,148) = 4.22$, $p < .01$. Although for participants in all groups, probes and targets had significantly ($p < .05$ in simple contrasts) longer RTs than did distracter items, targets had significantly longer RTs ($p < .05$ in simple contrasts) than probes only for participants in CM1, CM2, and innocent conditions. Probe and target RTs did not differ for CM0 and CM3 participants. Especially surprising was the finding for innocent participants that probe RTs were longer than RTs to distracter items, despite matching probe and distracter items on a variety of factors such as word frequency, word length, and semantic category, and piloting probe, target, and distracter items to ensure that they were equally salient. As detailed below, however, this differential response pattern was not observed in the electrophysiological domain (i.e., amplitude or latency). Incorrect responses for the five groups are presented in Table 1 (right column). Similar to the reaction time data, participants made significantly more errors to targets and probes than they did to distracter items: main effect of item type, $F(2,148) = 186.2$, $p < .001$, an effect that was not qualified by group: Item Type × Group interaction, $F(8,148) = 1.6$, $p < .17$. Further breakdown for all groups and conditions revealed higher error rates for targets as compared to probe items, whereas error rates for target and probe items were significantly higher than for distracter items.

P3 amplitudes. Grand average waveforms for midline sites are depicted for each group and item type in Figure 3. Visual inspection revealed the prototypical scalp topography of P3 amplitude when recorded with linked mastoids reference, that is, an increase from frontal to parietal sites. Moreover, innocent participants (left column) produced the predicted waveform pattern of large P3 amplitude in response to target items, whereas distracter and probe items produced highly similar waveforms, indicating unfamiliarity with these items. Among guilty participants, targets generally produced larger P3 amplitudes than did distracter items, with the crucial probes possessing P3 amplitudes somewhat smaller than those to the target and in some cases larger than those to the distracter items.

A mixed-model ANOVA with item type (probe, target, distracter) and site (Fz, Cz, Pz) as repeated-measures factors and group (INN, CM0, CM1, CM2, CM3) as a between-subjects

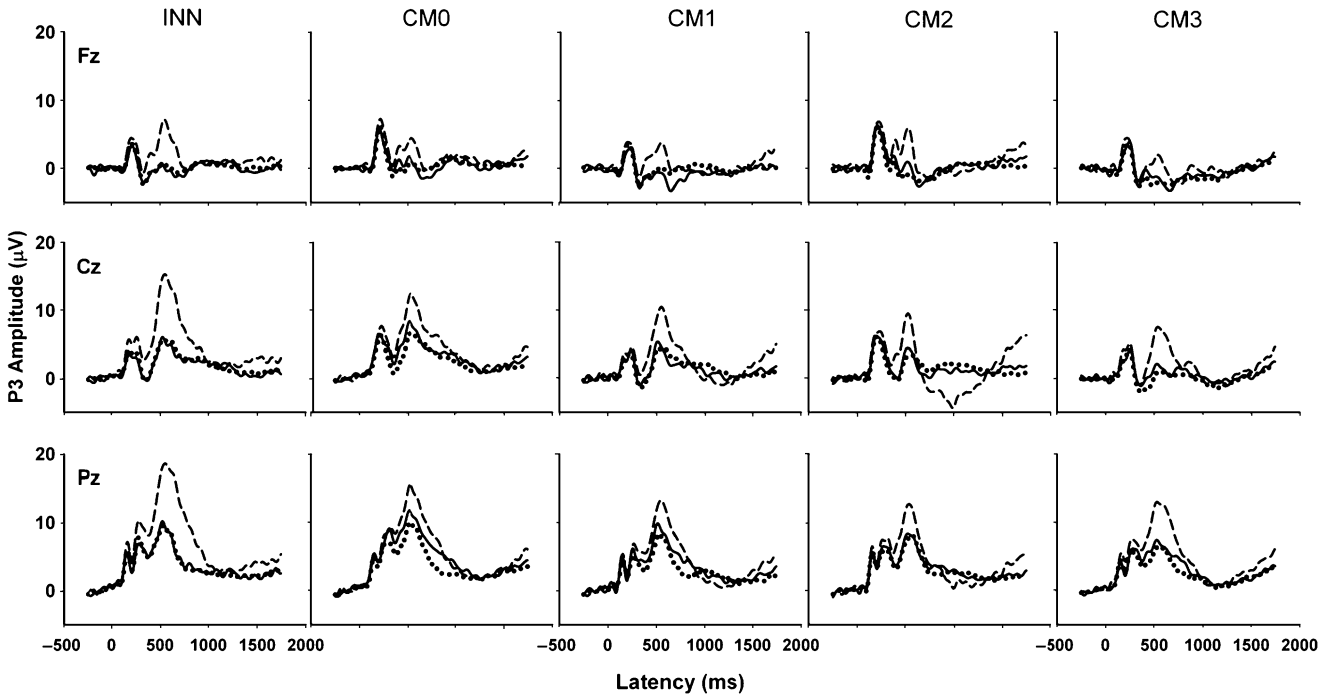

Figure 3. Grand average waveforms across three midlines sites for probe (dotted), target (dashed), and distracter (solid) items across innocent (INN), standard guilty (CM0), and three countermeasures groups (CM1, CM2, CM3).

factor was conducted. Main effects of item type, $F(2,140) = 171.87$, $p < .01$, and site, $F(2,140) = 170.61$, $p < .01$, were observed, in addition to several significant interactions: Group × Item Type, $F(8,140) = 2.82$, $p < .01$, Item Type × Site, $F(4,280) = 14.49$, $p < .01$, and Group × Item Type × Site, $F(16,280) = 3.031$, $p < .01$. To decompose the interactions, separate Item Type × Site ANOVAs were run for each group separately. Details of the results are summarized in Table 2. Post hoc analyses of P3 amplitude for each group at size Pz revealed that, for all groups, targets produced significantly larger P3 amplitudes than distracter items (all $ps < .001$) and significantly larger P3 amplitudes than probe items (all $ps < .001$). Probe items were significantly larger than distracter items for CM0 participants ($p < .02$), but did not significantly differ for any other groups (CM1, $p < .08$, CM2, $p < .12$, CM3, $p < .24$, INN, $p < .35$), although, given the clear directional hypothesis that probes should produce larger P3 amplitudes than distracters, one might also accept that probes were significantly larger than distracters for CM1 (one-tailed $p < .04$).

Table 2. *Summary of Main and Interaction Effects of P3 Amplitude for 79 Participants across Five Experimental Conditions for Stimulus Type and Electrode Site*

	Type		Site		Type × Site	
Condition	*df*	*F*	*df*	*F*	*df*	*F*
CM0	2,28	25.50*	2,28	24.48*	2,56	3.71**
CM1	2,34	30.74*	2,34	61.23*	2,56	4.61**
CM2	2,28	26.34*	2,28	27.32*	2,56	n.s.
CM3	2,28	21.30*	2,28	21.55*	2,56	6.19*
INN	2,30	102.32*	2,30	99.80*	2,60	16.10*

*$p \leq .01$, **$p \leq .5$.

Individual Classification

Table 3 presents the percentage of participants in each group classified as guilty, innocent, or indeterminate using each of the three classification methods. Overall, classification accuracy for guilty participants was rather low, and considerably lower than that reported in previous studies. This low classification rate was further reduced in the countermeasures groups. In contrast, innocent participants were almost never classified as guilty. The Bayesian and the bootstrapping of amplitude methods revealed high accuracy for innocent participants, comparable to those seen in past studies. The bootstrapping of cross-correlations method, however, correctly exonerated only 44% of the innocent participants, with an indeterminate rate considerably higher than that seen in earlier studies.[3]

To statistically assess the impact of countermeasures on detecting guilty participants, a chi-square analysis for guilty participants was conducted. If countermeasures altered the hit rate, this would be reflected in a significant chi-square statistic. A chi-square test was conducted involving group (CM0, CM1, CM2, CM3) and verdict (guilty vs. innocent for Bayes and bootstrapping of amplitudes, or guilty vs. innocent vs. indeterminate for bootstrapping of cross-correlations). For none of the three measures was there a significant chi-square (all $ps > .12$). Because this analysis treated each countermeasures group (CM1, CM2, CM3) as a distinct group, the analysis may have lacked power. The analysis was rerun collapsing all countermeasures groups into a

[3]Farwell (personal communication, October 2002) noted that he uses all trials, regardless of correctness of response in his analysis. Bootstrapping of cross correlations using all responses, instead of only correct response trials, changed results slightly for guilty, innocent, and indeterminate verdicts to the following values for each group: CM0 27%, 7%, 67%; CM1 22%, 6%, 72%; CM2 13%, 7%, 80%; CM3 33%, 0%, 67%; and innocent 0%, 50%, 50%.

Table 3. *Hit Rates (Percentage) of Participants Determined Guilty, Innocent, or Indeterminate across Three Classification Procedures*

	Bootstrapping					Bayesian					Peak-Peak				
Verdict	Inn	CM0	CM1	CM2	CM3	Inn	CM0	CM1	CM2	CM3	Inn	CM0	CM1	CM2	CM3
Guilty	0	27	11	13	7	6	47	17	20	13	0	47	11	20	27
Innocent	44	13	11	13	0	94	53	83	80	87	100	53	89	80	73
Indeterminate	56	60	78	73	93	0	0	0	0	0	0	0	0	0	0

single group, thus assessing the chi-square for Group (no countermeasure vs. countermeasure) × Verdict (guilty vs. innocent or guilty vs. innocent vs. indeterminate for bootstrapping of correlations). A significant effect of countermeasures on hit rate was observed for the Bayesian method ($p < .02$) and for the bootstrapping of amplitudes method ($p < .03$), but not for the bootstrapping of cross-correlations ($p < .22$).

To investigate whether the specific thresholds for determining guilt were optimal and to compare the utility of the different classification methods (Bayes and the two bootstrapping methods), receiver operator characteristic (ROC) analyses were conducted, with ROC curves for each method and guilty group plotted in Figure 4. Each curve represents the performance of a method when differentiating between a given guilty group (i.e., CM0, CM1, CM2, or CM3) and the innocent group. Input for each method was the test statistic used to determine guilt or innocence: Bayesian probability for Bayes, bootstrap statistic for bootstrapping of cross-correlations, and z score for bootstrapping of peak-to-peak amplitude differences. For the Bayesian method, the area under the curve (AUC) for the three countermeasures groups (CM1, CM2, and CM3) were all significantly smaller ($p < .01$) than that for the standard guilty group (CM0). The CM1 group additionally had significantly ($p < .01$) smaller AUC than the other two CM groups, whereas the latter did not significantly differ from one another. For the bootstrapping of peak-to-peak amplitudes, the three countermeasures groups again had significantly ($p < .01$) smaller AUCs than the standard guilty group, but the three countermeasures groups did not differ significantly from one another. The pattern of results for the bootstrapping of cross-correlations stood in contrast to these other metrics. The CM3 group produced a significantly ($p < .01$) larger AUC compared to the other countermeasures and the standard guilty group, whereas these latter three groups did not differ significantly from one another.

To directly compare the utility of the three approaches, the AUCs for the three approaches were compared, using the approach of Metz, Wang, and Kronman (1984) for testing the significance of differences between ROC curves measured from correlated data, AUCs were first computed comparing all guilty subjects combined (CM0, CM1, CM2, CM3) to innocent subjects. The AUC of .88 for the bootstrapping of cross-correlations method significantly exceeded ($p < .05$) the AUC of .74 for the Bayesian method, which in turn significantly exceeded ($p < .05$) the AUC of .59 for the bootstrapping of peak-to-peak amplitude differences. The AUC of the bootstrapping of cross-correlations method also significantly exceeded ($p < .001$) that of the bootstrapping of peak-to-peak amplitude differences. Next each countermeasures condition was considered separately (AUC for that condition compared to innocent, as presented in Figure 4), comparing the three methods. The bootstrapping of cross-correlations had significantly better AUC than the Bayesian method for the CM1 and CM3 conditions, but not for the CM0 and CM2 conditions. The bootstrapping of cross-correlations had significantly better AUC than the bootstrapping of peak-to-peak amplitude differences for all four conditions (CM0, CM1, CM1, CM2). The Bayesian method had significantly better AUC than the bootstrapping of peak-to-peak amplitude differences for CM0 and CM2, but not CM1 or CM3.

Discussion

The present study was conducted to extend findings (Rosenfeld et al., 2004) demonstrating the P3 oddball deception-detection paradigm's vulnerability to physical and mental countermeasures. Specifically, a highly realistic mock-crime scenario combining virtual reality with an ERP-based guilty knowledge test was used to assess if physical, mental, or a combination of these countermeasures would be sufficient to elude detection. Statistical approaches were extended beyond those of Rosenfeld et al. (2004) by using Bayesian classification in addition to bootstrap statistics of cross-correlations and bootstrap differences of peak-to-peak amplitude differences. Two approaches to analysis—Bayesian and peak-to-peak bootstrapping—shared in common their low hit rates and susceptibility to countermeasures. The bootstrapping of cross-correlations first utilized by Farwell and Donchin (1991), and still a key component of the commercial "brain fingerprinting," showed promise, as it was relatively resistant to countermeasures, although many indeterminate verdicts limited the interpretability of these result. The main findings reveal that overall classification of guilty participants was rather poor. Guilty participants who were not instructed in the use of countermeasures were correctly classified between 27% and 47% of the time, depending on the analysis approach used. These rates were further lowered among participants instructed in the use of countermeasures, with classification accuracy ranging from 7% to 27% depending on countermeasure and analysis approach used. Although overall hit rates where generally lower compared to prior ERP-based deception detection studies (Allen et al., 1992; Farwell & Donchin, 1991; Rosenfeld et al., 2004), data regarding innocent participants were consistent with prior findings, as innocent participants were almost never incorrectly classified as guilty based on ERP data.

Data presented above for bootstrapping of cross-correlations (i.e., brain fingerprinting) and Bayesian analysis used previously established cutpoints. Because these cutpoints were originally established using relatively small samples, these cutpoints may not be optimal. Although it would be tempting to adjust respective cutpoints to maximize present hit rates, such a post hoc approach would be problematic, as it would not reflect adequately how these methods would function in an applied situation and would run the risk of inflating the accuracy of these methods. Ideally, a large validation study with verifiable ground truth from field data would be needed to establish highly sensitive and generalizable cutpoints to create an acceptable balance between correct and incorrect classifications.

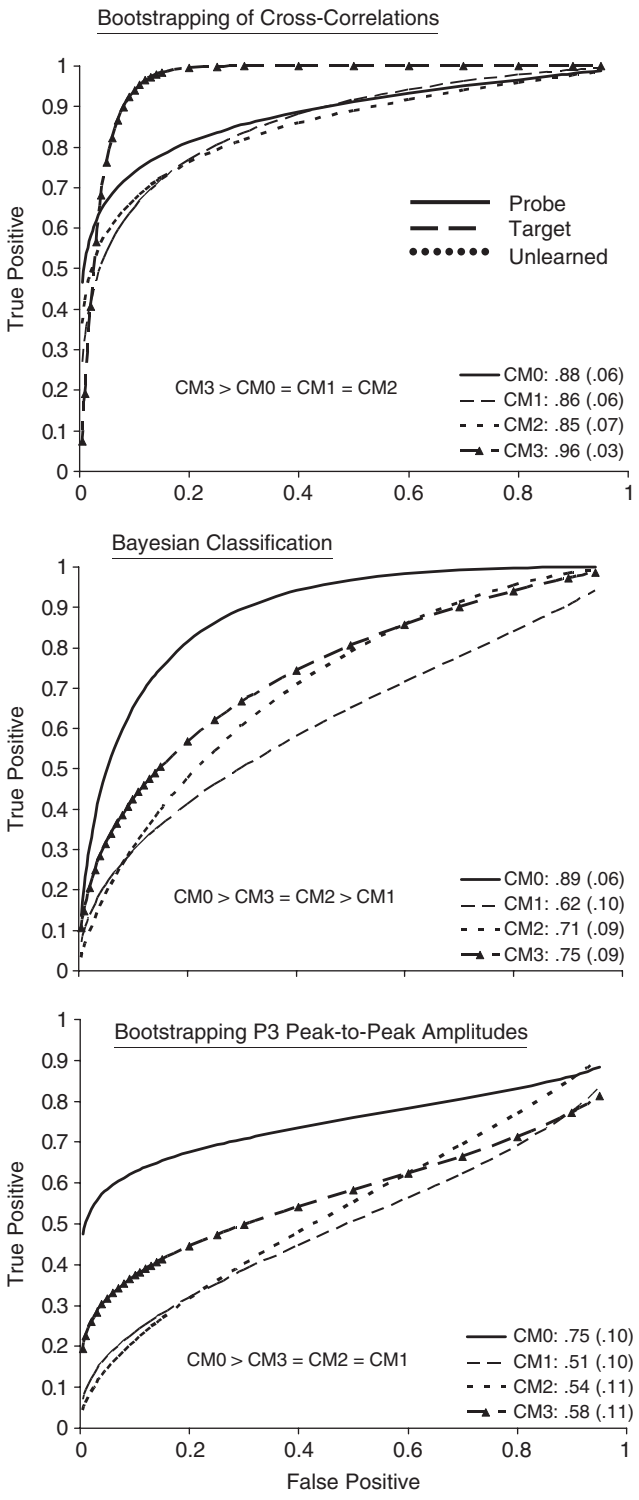

Figure 4. Receiver operating characteristic (ROC) curves to compare detection efficiency of three classification methods for four guilty conditions, each compared to the innocent condition. Legend displays area under the curve (AUC) and associated standard deviation.

The present ROC results suggest, however, that altering the cutpoint may significantly improve the bootstrapping of cross-correlations classification. This method provided significantly better discrimination of innocent from guilty subjects, as indexed by larger AUC values, than the Bayesian or the bootstrapping of peak-to-peak methods. Because the preestablished cutpoints for this bootstrapping method are devised to provide classification in only nonambiguous cases, the high number of indeterminates reduces the current applicability of this approach. Adjustment of these established cutpoints could improve classification by assigning verdicts to the current indeterminate cases. On the other hand, the AUC values obtained with the bootstrapping of cross-correlations in the present study (AUCs ranging from .85 to .96 depending on CM condition) are nonetheless smaller than those obtained in either the study of Farwell and Donchin (1991; AUC = .994) or that of Allen and Iacono (1997;AUC = .99), suggesting that the ability of this approach to adequately classify subjects cannot be assumed to generalize beyond these contrived laboratory situations.

Classifications based on behavioral responses such as reaction time were less consistent, however. Despite considerable care to create a carefully matched stimulus pool similar in word frequency and physical and semantic characteristics, innocent participants responded to probe items with longer response times compared to distracter items, even though probes should have seemed as unfamiliar as distracters to this group of participants. Because a pilot study found no statistical differences in how frequently naïve participants selected probes over distracter items, results from that pilot study might have constituted a Type 1 error. The behavioral findings are informative, however, as they exemplify the difficulty field studies would have to overcome to produce a stimulus without such problems. If participants in this study had actually been involved in a crime, with ground truth unknown and reaction times used as the only measure to determine guilt, legitimately innocent participants would have been found guilty at unacceptably high rates. One of the implications of this finding is that response time may not necessarily be well suited as a predictor of guilt or innocence, even though many studies have found that response time often functions well in that capacity (e.g., Allen et al., 1992; Seymour, Seifert, Shafto, & Mosmann, 2000), although other investigators raise concern about the suitability of response times (Gronau, Ben-Shakhar, & Cohen, 2005). Another implication is that it could be difficult to match crime-relevant items in field work and that stimulus sets may require close scrutiny to avoid false positive classifications. Because crime-relevant probes are dictated by the nature of the crime, selection of targets will likewise be constrained by crime-related parameters, making high-quality matches across item types more challenging, particularly with large stimulus sets like those used in this study.

In selecting items, care must be taken to avoid creating a heterogeneous set consisting of "peripheral" and "central" items (Carmel, Dayan, Naveh, Raveh, & Ben-Shakhar, 2003). The former referred to items related to the crime but not as well encoded as "central" items, which are items vividly remembered long after completion of the crime. Thus, a hypothetical test constructed solely from "peripheral" items may lead to a high rate of false negatives, just because perpetrators may have failed to encode stimuli at the time of the crime. A large stimulus set may also heighten response requirements, which, in the present study, may have resulted in lower hit rates, as participants may have been simply overwhelmed by the large amount of information presented to them. The use of a simplified stimulus set (Rosenfeld et al., 2004) comprised of fewer items might address some of the problems documented in this study. However, such an approach runs the additional risk that, with a limited number of probe items, idiosyncratic responses may lead

innocent participants to appear guilty at higher rates than with larger stimulus sets or to greater false negative results for guilty participants if item selection included too many items that were peripheral to the crime or otherwise poorly encoded or remembered.

Findings of the present investigation contrast with previous studies, many of which involved only list learning or simple mock crimes and suggest that ERP-based deception detection procedures may have limitations in more naturalistic environments. Despite careful selection of crime-relevant items and immediate test following the mock crime, hit rates were quite low. Because integration of a virtual reality (VR) component into ERP research is a relatively novel approach, concerns that the use of this technology may be related to the aberrant findings warrant consideration. Although research on the effects of virtual environments (VE) is scant (Mager, Bullinger, Mueller-Spahn, Kuntze, & Stoermer, 2001; Mager, Bullinger, Roessler, Mueller-Spahn, & Stoermer, 2000), data suggest that ERPs in response to VE are not different from ERPs in naturalistic settings. Moreover, the present study utilized a mixed design, with the VE only being used to encode a minority of probe items; most probes were learned via a standard reading task. We recognize that the present study is limited because of its lack of parametric data comparing the potential difference of ERPs during a standard mock crime compared to the VE. Although such comparison was beyond the scope of this study, future comparisons of the more traditional mock crime environment and VE are clearly necessary.

Another possible explanation for the low hit rates concerns the extent to which the present protocol emphasized the target stimuli. Because participants had memorized all probe items in order to carry out their ''mission,'' it was decided that targets should be learned to a similar level of proficiency. Because targets were learned just prior to the ERP task and because considerable emphasis was placed on correct encoding, they may have been very easy to identify compared to the probe items, thus increasing the P3 amplitude for these target items above what would usually be seen, thereby reducing the hit rate. This effect appeared visible on the P3 group average level where targets had significantly larger P3 amplitudes than probes or distracters. After commencing this study, it was learned that Farwell (personal communication, January 2001) does not extensively review targets, but rather relies on very brief presentation of targets that have been selected from the crime environment, thus leading guilty participants to have some confusion over the probes versus targets. This would reduce the extent to which targets appear as especially distinct and avoid a problem of creating exceptionally large P3 amplitude. Although previous studies (e.g., Allen et al. 1992) relied on list learning to a criterion of perfect recall, these studies also provided a similar study phase for probes. It is hypothesized the key factor in the success of the ERP-based approach is matching the distinctiveness of probes and targets, and having similar encoding procedures may indeed be important. In the previous studies with list learning, both probes and targets were learned explicitly and to a recognition criterion. In Farwell's approach, both probes and targets are learned incidentally. If this hypothesis is correct, it appears that overemphasis of targets could be utilized as a generalized countermeasure to reduce overall detection rates. Although unintended in the present study, this finding adds to concerns that ERP-based measures may be susceptible to participant intervention, especially if countermeasures could be easily constructed by simply rehearsing or emphasizing target items.

Lastly, the low detection rates might be attributable to the realism of the present study. Not only did one forgetful student contact police upon receiving his e-mail reminder of his mission (Halperin, 2002), most participants described the study as realistic and anxiety provoking, particularly because participants were under time pressure to complete part of their study and had to remember a substantial amount of information in order to obtain a rather sizable reward. Carmel et al. (2003) argued that standard mock crime scenarios make the recall of crime relevant information easier, as participants are only exposed to a limited amount of information, learn all crime-relevant items to perfection, and are tested immediately afterward. Ecologically less valid approaches, such as the use of highly salient stimuli like autobiographical information (e.g., birthdates) has been shown to result in higher hit rates (Miller & Rosenfeld, 2004), perhaps because of the lower difficulty level remembering such information. In contrast, in mock-crime investigations, there may need to be some reliance on the peripheral instead of central items, which would lower hit rates even further (Carmel et al., 2003). Although the design of the present study did not allow for a reanalysis to differentiate peripheral from central items, at least 11 of the 12 probes were highly studied and thus were more like central items, which in turn yield reliable and robust hit rates, even under conditions that make encoding and retrieval of information more difficult (Carmel et al., 2003). As previously noted, however, the large stimulus set may have contributed to the lower hit rates, as having to remember and respond to 12 probes and targets may have been taxing and reduced the significance of individual items, thus reducing the significance of these stimuli, leading to lower hit rates.

Results suggested that the complex countermeasure modeled after Rosenfeld et al. (2004) appeared too difficult to implement for most participants left to their own training. Because participants used written instructions to train on their own to implement their respective countermeasures, there is concern that there was poor countermeasure compliance for CM3. Exit-interview data revealed that 47% of CM3 participants relied on impromptu strategies that may or may have not been successful, whereas only 53% of these participants executed CM3 more or less as instructed. Almost no participants using CM1 or CM2 reported any compliance problems or difficulties with training and implementation of their respective countermeasure at the time of test. Although data suggested that more complex countermeasures may require specific and verifiable training, simple measures such as thinking about being slapped or wiggling a toe in response to specific stimuli do not appear to require such preparation. Idiosyncratic responding of participants in CM3 may have inadvertently aided classification in the bootstrapping of cross-correlation approach, however. As subjects deviated from their experimental instructions, their impromptu strategies may have maximized the cross-correlation classification system, yielding greater accuracy for the CM3 as compared to all other experimental groups as depicted in Figure 4.

The present results, as well as those of other studies (Carmel et al., 2003; Rosenfeld et al., 2004; Sasaki et al., 2001), suggest that further research is needed to more clearly ascertain the limits of the P3 and of ERPs more generally as a means in deception detection, in addition to limitations associated with mock-crime scenarios. For example, considering that 50% of jailed inmates were under the influence of drugs during their crime (U.S. Bureau of Justice Statistics, 2005), further investigations would be helpful to determine how Bayesian classification, bootstrapping of

cross-correlations, or bootstrapping of peak-to-peak amplitude would perform differently in scenarios in which perpetrators were intoxicated or when conditions during encoding are significantly different from conditions during testing. Furthermore, results of the present study indicate that the ERP-based deception detection procedure may be vulnerable to relatively minor deviation from protocol (e.g., degree of emphasis on targets) and could benefit from standardization. Lastly, because the ability to create realistic VE is continuously advancing, such technology has great potential to evolve into a cost-effective and efficient alternative to emulate naturalistic settings with unprecedented experimental control in a laboratory setting. fMRI-based research has utilized this type of technology for a variety of topics, such as smoking (Lee, Lim, Wiederhold, & Graham, 2005), alcohol intoxication (Calhoun, Carvalho, Astur, & Pearlson, 2005), and social behaviors (Pelphrey, Viola, & McCarthy, 2004). Rapidly advancing software and hardware make integration of biometric variables such as a response time, gaze direction, or gait tracking relatively easy. Artificial intelligence and emotional expression of virtual actors could raise the level of interactions of study participants with computer systems to a new level. Further development of VEs is therefore encouraged in addition to detailed parametric comparisons beyond the practical investigation in this study.

Although new technologies may improve or overcome some of the limitations of extant ones, it is important to consider that all approaches will likely use one of three approaches to the detection of deception. These approaches involve detecting (1) arousal or emotion associated with lying, (2) recognition of information only known to those associated with the crime, or (3) alterations in cognitive processes that may occur when individuals lie. Examples of the first approach include traditional field polygraphy, changes in facial expression (Cheng & Broadhurst, 2005; Ekman & Friesen, 1974; Ekman, Friesen, & O'Sullivan, 1988), examination of demeanor (Granhag & Strömwall, 2002; Pollina & Squires, 1998; Vrij, Edward, & Bull, 2001; Vrij, Semin, & Bull, 1996), or thermograpy (Pavlidis, Eberhardt, & Levine, 2002). Examples of the memory/recognition approach include, of course, the ERP deception detection paradigm used in the present study and others (Allen et al., 1992; Farwell & Donchin, 1991; Miller & Rosenfeld, 2004; Rosenfeld et al., 1999, 2004), as well as standard skin-conductance-based guilty knowledge tests (Ben-Shakhar & Elaad, 2002; Elaad & Ben-Shakhar, 1991; Engelhard, Merckelbach, & van den Hout, 2003; Lykken, 1959, 1960, 1991) and performance on forced-choice tests (Moore & Donders, 2004; Rosenfeld, Sweet, Chuang, & Ellwanger, 1996; Tardif, Barry, Fox, & Johnstone, 2000). Examples of other cognitive correlates included processing of aspects of deception in various cortical areas (Johnson, Barnhardt, & Zhu, 2003, 2004, 2005). Each approach will have inherent limitations, with arousal-based procedures prone to false positive verdicts and the other approaches somewhat vulnerable to false negative verdicts.[4] Ultimately the best classification accuracy may involve combining different approaches to overcome the limitations of each.

The practical implication is that no matter what method of analysis is used with this ERP-based guilty knowledge paradigm, an innocent verdict is not necessarily informative, occurring among both guilty and innocent participants quite often, yet guilty verdicts may in fact be quite useful, because they almost exclusively occur among the guilty participants. This stands in stark contrast to standard field polygraphy using autonomic measures and the control-question technique, where innocent verdicts are informative (occurring almost exclusively among the innocent) but where guilty verdicts are not informative due to the high false positive rate of this test. The different pattern of performance of the ERP guilty knowledge test and the standard polygraph control-question test derive from the assumptions about how guilty participants will appear deceptive. In the control-question test, it is assumed that guilty participants will show greater anxiety or arousal to the relevant questions, but because innocent participants may also show such arousal, guilty verdicts will also occur for innocent participants. In contrast, for the ERP-based GKT, it is assumed that guilty participants will recognize relevant crime details and innocent participants rarely would recognize such items and therefore are rarely classified as guilty.

REFERENCES

Allen, J. 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. 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.

Bashore, T. R., & Rapp, P. E. (1993). Are there alternatives to traditional polygraph procedures? *Psychological Bulletin*, *113*, 3–22.

Bayes, T. (1763). Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society, London 53*, 370–418 (reprinted in 1958 in *Biometrika*, *45*, 293–315).

Ben-Shakhar, G., & Dolev, K. (1996). Psychophysiological detection through the guilty knowledge technique: Effect of mental countermeasures. *Journal of Applied Psychology*, *81*, 273–281.

Ben-Shakhar, G., & Elaad, E. (2002). The Guilty Knowledge Test (GKT) as an application of psychophysiology: Future prospects and obstacles. In M. Kleiner (Ed.), *Handbook of polygraph testing* (pp. 87–102). New York: Academic Press.

Calhoun, V. D., Carvalho, K., Astur, R., & Pearlson, G. D. (2005). Using virtual reality to study alcohol intoxication effects on the neural correlates of simulated driving. *Applied Psychophysiology and Biofeedback*, *30*, 285–306.

Carmel, D., Dayan, E., Naveh, A., Raveh, O., & Ben-Shakhar, G. (2003). Estimating the validity of the guilty knowledge test from simulated experiments: The external validity of mock crime studies. *Journal of Experimental Psychology: Applied*, *9*, 261–269.

Cheng, K. H. W., & Broadhurst, R. (2005). The detection of deception: The effects of first and second language on lie detection ability. *Psychiatry, Psychology and Law*, *12*, 107–118.

Donchin, E., & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*, 357–427.

Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology*, *29*, 288–298.

[4]To some extent the tendency toward false positive versus false negative outcomes may reflect the particular cutpoints selected to differentiate innocent from guilty subjects, as a change in cutpoint can increase or decrease the false positive or false negative rate. Additionally, however, the rationale for the arousal-based tests versus the recognition-based tests will mean that there is a greater propensity for false positives using arousal-based procedures like the CQT and a greater chance for false negatives using a recognition procedure like the GKT.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, *54*, 414–420.

Elaad, E., & Ben-Shakhar, G. (1991). Effects of mental countermeasures on psychophysiological detection in the guilty knowledge test. *International Journal of Psychophysiology*, *11*, 99–108.

Engelhard, I. M., Merckelbach, H., & van den Hout, M. A. (2003). The Guilty Knowledge Test and the modified Stroop task in detection of deception: An exploratory study. *Psychological Reports*, *92*, 683–691.

Farwell, L. A. (1994). Method and apparatus for multifaceted electroencephalographic response analysis (MERA). *U.S. patent*, *5*, 363–858.

Farwell, L. A. (1995a). Method and apparatus for truth detection. *U.S. Patent*, *5*, 406–956.

Farwell, L. A. (1995b). Method for eletroencephalographic information detection. *U.S. patent*, *5*, 467–777.

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ('lie detection') with event-related brain 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.

Feder, B. J. (2001, October 9). Truth and justice, by the blip of a brain wave. *New York Times*, p. F3.

Granhag, P. A., & Strömwall, L. A. (2002). Repeated interrogations: Verbal and non-verbal cues to deception. *Applied Cognitive Psychology*, *16*, 243–257.

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.

Halperin, D. (2002, December 6). Police beat: Suspicious email sent. *The Daily Wildcat*. Retrieved from http://wildcat.arizona.edu/papers/96/71/01_50.html.

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

Honts, C. R., Hodes, R. L., & Raskin, D. C. (1985). Effects of physical countermeasures on the physiological detection of deception. *Journal of Applied Psychology*, *70*, 177–187.

Honts, C. R., Raskin, D. C., & Kircher, J. C. (1987). Effects of physical countermeasures and their electromyographic detection during polygraph tests for deception. *Journal of Psychophysiology*, *1*, 241–247.

Iacono, W. G. (2000). The detection of deception. In J. T. Cacioppo & L. G. Tassinary (Eds.), *Handbook of psychophysiology* (2nd ed, pp. 772–793). New York: Cambridge University Press.

Iacono, W. G., Cerri, A. M., Patrick, C. J., & Fleming, J. A. (1992). Use of antianxiety drugs as countermeasures in the detection of guilty knowledge. *Journal of Applied Psychology*, *77*, 60–64.

Iacono, W. G., & Lykken, D. T. (1997). The validity of the lie detector: Two surveys of scientific opinion. *Journal of Applied Psychology*, *82*, 426–433.

Iacono, W. G., & Patrick, C. J. (1997). Polygraphy and integrity testing. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed, pp. 252–281). New York: Guilford Press.

Johnson, R. (1986). A triarchic model of P300 amplitude. *Psychophysiology*, *23*, 367–384.

Johnson, R. J., Barnhardt, J., & Zhu, J. (2003). The deceptive response: Effects of response conflict and strategic monitoring on the late positive component and episodic memory-related brain activity. *Biological Psychology*, *64*, 217–253.

Johnson, R. J., Barnhardt, J., & Zhu, J. (2004). The contribution of executive processes to deceptive responding. *Neuropsychologia*, *42*, 878–901.

Johnson, R. J., Barnhardt, J., & Zhu, J. (2005). Differential effects of practice on the executive processes used for truthful and deceptive responses: An event-related brain potential study. *Cognitive Brain Research*, *24*, 386–404.

Kubis, J. F. (1962). *Studies in Lie Detection: Computer Feasibility Considerations*.

Kuntze, M. F., Störmer, R., Mager, R., Müller-Spahn, F., & Bullinger, A. (2003). Die Behandlung der Höhenangst in einer virtuellen Umgebung. *Nervenarzt*, *74*, 428–435.

Lee, J.-H., Lim, Y., Wiederhold, B. K., & Graham, S. J. (2005). A functional magnetic resonance imaging (fMRI) study of cue-induced smoking craving in virtual environments. *Applied Psychophysiology and Biofeedback*, *30*, 195–204.

Lee, J. M., Ku, J. H., Jang, D. P., Kim, D. H., Choi, Y. H., Kim, I. Y., et al. (2002). Virtual reality system for treatment of the fear of public speaking using image-based rendering and moving pictures. *CyberPsychology & Behavior*, *5*, 191–195.

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

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

Lykken, D. T. (1987). The detection of deception. In L. S. Wrightsman, C. E. Willis, & S. M. Kassin (Eds.), *On the witness stand* (pp. 37–47). Thousand Oaks, CA: Sage.

Lykken, D. T. (1991). The lie detector controversy: An alternative solution. In J. R. Jennings, P. K. Ackles, & M. G. H. Coles (Eds.), *Advances in psychophysiology: A research annual* (Vol. 4, pp. 209–214). London: Kingsley.

Mager, R., Bullinger, A. H., Mueller-Spahn, F., Kuntze, M. F., & Stoermer, R. (2001). Real-time monitoring of brain activity in patients with specific phobia during exposure therapy, employing a stereoscopic virtual environment. *CyberPsychology & Behavior*, *4*, 465–469.

Mager, R., Bullinger, A. H., Roessler, A., Mueller-Spahn, F., & Stoermer, R. (2000). Monitoring brain activity during use of stereoscopic virtual environments. *CyberPsychology & Behavior*, *3*, 407–413.

Mertens, R., & Polich, J. (1997). P300 from a single-stimulus paradigm: Passive versus active tasks and stimulus modality. *Electroencephalography & Clinical Neurophysiology: Evoked Potentials*, *104*, 488–497.

Metz, C. E., Wang, P., & Kronman, H. B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In F. Deconinck (Ed.), *Information processing in medical imaging*. The Hague: Martinus Nijhoff.

Miller, A. R., & Rosenfeld, J. P. (2004). Response-specific scalp distributions in deception detection and ERP correlates of psychopathic personality traits. *Journal of Psychophysiology*, *18*, 13–26.

Moore, B. A., & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury*, *18*, 975–984.

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

Pavlidis, I., Eberhardt, N. L., & Levine, J. A. (2002). Seeing through the face of deception. *Nature*, *415*, 35.

Pelphrey, K. A., Viola, R. J., & McCarthy, G. (2004). When strangers pass: Processing of mutual and averted social gaze in the superior temporal sulcus. *Psychological Science*, *15*, 598–603.

Pollina, D. A., & Squires, N. K. (1998). Many-valued logic and event-related potentials. *Brain and Language*, *63*, 321–345.

Rosenfeld, J. P. (1995). Alternative views of Bashore and Rapp's (1993) alternatives to traditional polygraphy: A critique. *Psychological Bulletin*, *117*, 159–166.

Rosenfeld, J. P. (2005). Brain fingerprinting: A critical analysis. *Scientific Review of mental health practic*, *4*, 20–37.

Rosenfeld, J. P., Angell, A., Johnson, M., & Qian, J.-h. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average waveforms. *Psychophysiology*, *28*, 319–335.

Rosenfeld, J. P., Ellwanger, J. W., Nolan, K., Wu, S., Bermann, R. G., & Sweet, J. (1999). P300 scalp amplitude distribution as an index of deception in a simulated cognitive deficit model. *International Journal of Psychophysiology*, *33*, 3–19.

Rosenfeld, J. P., Rao, A., Soskins, M., & Miller, A. (2003). Scaled P300 scalp distribution correlates of verbal deception in an autobiographical oddball paradigm: Control for task demand. *Journal of Psychophysiology*, *17*, 14–22.

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

Rosenfeld, J. P., Sweet, J. J., Chuang, J., & Ellwanger, J. (1996). Detection of simulated malingering using forced choice recognition enhanced with event-related potential recording. *Clinical Neuropsychologist*, *10*, 163–179.

Sasaki, M., Hira, S., & Matsuda, T. (2001). Effects of a mental countermeasure on the physiological detection of deception using the event-related brain potentials. *Japanese Journal of Psychology*, *72*, 322–328.

Semlitsch, H. V., Anderer, P., Schuster, P., & Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology*, *23*, 695–703.

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.

Scheeres, J. (2001). Thought police peek into brains. Retrieved from www.wirednews.com.

Soskins, M., Rosenfeld, J. P., & Niendam, T. (2001). Peak-to-peak measurement of P300 recorded at 0.3 Hz high pass filter settings in intraindividual diagnosis: Complex vs. simple paradigms. *International Journal of Psychophysiology*, *40*, 173–180.

Tardif, H. P., Barry, R. J., Fox, A. M., & Johnstone, S. J. (2000). Detection of feigned recognition memory impairment using the old/new effect of the event-related potential. *International Journal of Psychophysiology*, *36*, 1–9.

U.S. Bureau of Justice Statistics (2005). *Substance dependence, abuse, and treatment of jail inmates, 2002*. NCJ publication no. 209588. Washington, DC: Author.

U.S. General Accounting Office (2001). *Investigative techniques: Federal agency views on the potential application of "brain fingerprinting."* Publication No. 440010. Washington, DC: Author.

U.S. Office of Technology Assessment (1983). *Scientific validity of polygraph testing: A research review and evaluation—A technical memorandum (Publication OTA-TM-H-15)*. Washington, DC: U.S. Library of Congress.

Vrij, A., Edward, K., & Bull, R. (2001). People's insight into their own behaviour and speech content while lying. *British Journal of Psychology*, *92*, 373–389.

Vrij, A., Semin, G. n. R., & Bull, R. (1996). Insight into behavior displayed during deception. *Human Communication Research*, *22*, 544–562.

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

Wiederhold, B. K., Jang, D. P., Kim, S. I., & Wiederhold, M. D. (2002). Physiological monitoring as an objective tool in virtual reality therapy. *CyberPsychology & Behavior*, *5*, 77–82.

Zirinsky, S., (Writer). (2002, June 14). *"It's all in your head" [Television series episode] 48 Hours*. New York: Columbia Broadcasting System.

(Received July 4, 2007; Accepted July 25, 2007)

APPENDIX

Table A.1 is a summary of 12 probe, 12 targets, and 48 unlearned distracter items.

Table A.1. *Summary of Stimuli*

	Item type					
Category	Probe	Target	D1	D2	D3	D4
ID of intruder	G47B	J37C	Z29Y	K65L	M93S	V27X
Object broken during mission	bowl	plate	cup	window	bottle	monitor
Object taken from safe	note	book	chain	magazine	coin	ring
Door combination	5676	4958	8901	4621	4576	5920
Time inside apartment	6	7	4	2	9	3
Name of contact person	Glen Plat	Ray Snell	Tim Howe	Gene Falk	Phil Jenks	Neil Rant
Code name of object in safe	rain	snow	hail	wind	ice	fog
Location of safe	picture	wall	sofa	mirror	chair	bed
Name of mission plan	op cow	op pig	op horse	op goat	op sheep	op mule
Color of envelope	red	white	blue	yellow	green	black
Location of 2nd object	closet	cabinet	basket	box	purse	table
2nd object taken	pistol	camera	knife	television	watch	radio