THE ROLE OF PSYCHOPHYSIOLOGY IN FORENSIC ASSESSMENTS:
DECEPTION DETECTION, ERPS AND VIRTUAL REALITY MOCK CRIME
SCENARIOS

by
Ralf Mertens

A Dissertation Submitted to the Faculty of the
DEPARTMENT OF PSYCHOLOGY
In Partial Fulfillment of the Requirements
For the Degree of
DOCTOR OF PHILOSOPHY

In the Graduate College
THE UNIVERSITY OF ARIZONA

2006

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Ralf Mertens entitled **The Role of Psychophysiology in Forensic Assessments: Deception Detection, ERPs and Virtual Reality Mock Crime Scenarios** and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

John Allen, Ph.D.

Date: March 27th, 2006

Al Kazniak, Ph.D.

Date: March 27th, 2006

Randy Garland, Ph.D.

Date: March 27th, 2006

Shannah Biggan, Ph.D.

Date: March 27th, 2006

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director: John Allen, Ph.D.

Date: March 27th, 2006

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library. Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Ralf Mertens

Acknowledgments

We thank Lauren Crawford, Nicholas Culp and Mirijam Rupp for their assistance to bring 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 correspondence to Ralf Mertens or John JB Allen, Department of Psychology, PO Box 210068, Tucson, AZ 85721-0068. Email: rmertens@u.arizona.edu or jallen@u.arizona.edu

TABLE OF CONTENTS

ABSTRACT.....	Page 6
INTRODUCTION.....	Page 7
METHODS.....	Page 13
RESULTS.....	Page 23
DISCUSSION.....	Page 28
APPENDIX A.....	Page 37
APPENDIX B.....	Page 39
REFERENCES.....	Page 47

ABSTRACT

ERPs, specifically the P3, have been proposed as an alternative to traditional polygraphy, with one approach (i.e., Brain Fingerprinting) being promoted as infallible to justify its use on a commercial basis. Concerns have been voiced, however, that such techniques would have to undergo peer-reviewed studies to satisfy validity concerns. Rosenfeld et al. (2004) found, for example, that mental countermeasures were effective in reducing detection rates using an amplitude based, peak-to-peak measure. The present study attempted to replicate and extend Rosenfeld et al.'s study, and to test Brain Fingerprinting's vulnerability to participant manipulation 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. Participants reported a high degree of realism supporting the external validity of this study and suggesting future uses of virtual environments. Hit rates across statistical methods were significantly lower for standard guilty and innocent participants as compared to previous studies; countermeasures reduced the overall hit rates even further. Brain Fingerprinting was as vulnerable to countermeasures as other statistical measures, and produced a significant number of indeterminate outcomes. Nevertheless, innocent participants remained protected from being falsely accused across statistical methods, consistent with findings of prior studies. Reaction times were determined unsuitable in determining guilt or innocence in this study. Results suggested that ERP based deception detection measures might lack the level of validity required for use in an applied setting.

INTRODUCTION

Conventional polygraph examinations suffer from many limitations and have been widely criticized in the scientific literature (Lykken, 1987; Iacono 2000; Iacono & Lykken, 1997; Iacono & Patrick, 1997; Office of Technology Assessment, 1983; National Research Council, 2003). Psychophysiolgists have thus examined alternative methods of deception detection (Bashore & Rapp, 1993; Rosenfeld, 1995), including measures of cortical activity such as event-related potentials (ERPs). Unlike conventional polygraph approaches, which assess emotional arousal associated with lying, ERP-based alternatives most often assess instead memory for salient aspects of a situation that would only be known to a perpetrator and few others such as investigators or witnesses.

In the service of assessing memory, ERPs may be a promising approach. Multiple ERP components have been found sensitive in the recollection of past experiences, and the latency of these components is so soon following a stimulus that it may be difficult to manipulate them voluntarily, an important characteristic in the context of forensic assessments. The P3, a cognitive component of the ERP that appears relatively quickly after stimulus presentation (i.e., in the range of 300-500 msec.) 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 only known 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 ERPs 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 bona-fide 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, 2003; Lee et al. 2002; Wiederhold, Dong, Kim & Wiederhold, 2002). VEs allow convenient collection of data not easily possible in standard mock crimes, such as measuring reaction times, gait trajectory, or gaze direction. 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 possible the replication of experiments easily 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 utilized a VE crime scene as instead 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 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 (United States General Accounting Office, 2001).

A potential weakness of ERP based measures is their possible vulnerability to countermeasures. Since conventional polygraph examinations 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 backwards in sevens) did not affect P3 amplitude or latency (Sasaki, Hira, & Matsuda, 2001), a novel physical countermeasure developed by Rosenfeld, Soskins, Bosh & Ryan (2004) appeared to be effective. Instead of attempting to manipulate the P3 in response to probe or target items, Rosenfeld et al. (2004) taught participants to enhance the significance of distracter items only, which in turn increased P3 amplitude in response to these items. This manipulation artificially 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 lowered detection rates dramatically. Using (a) an imperceptible physical responses (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.

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 whether countermeasures in response to target items could also be effective. Since P3 responses to target items produce a “gold standard” to which responses to probe items will be compared, increasing amplitude in response to target items would reduce the significance of probe related increases, reducing the probability of being detected. 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 neglecting to consider 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, SD=3.09) and 41 female (mean age = 19.34, SD =4.43) undergraduate students of the University of Arizona, are represented in this dataset¹. 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 CNS, (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; 2,3,4) 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 3-D computer environment with crime-relevant items removed, but that were tested nevertheless for knowledge of crime relevant items.

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, citing feelings of distrust, uncertainty, and concern. Seventy-five percent of participants rated the study at 7 or above on a 10 point questionnaire, with 10 indicating the highest level of realism imaginable. As an attest to its realism, one forgetful student, scheduled to participate in the study, contacted police after receiving an email reminding him to carry out his “mission².” Level of motivation to perform to the best of their ability, was rated at 7 or above on the same ten-point scale by ninety-one percent of participants.

Mock crime procedure: At the end of an initial visit to the lab to determine eligibility, participants were informed that they would receive via email a set of instructions delineating their “mission.” Guilty participants received instructions to enter an unoccupied room usually off-limits to undergraduate students, to “break into” a password protected computer, and to retrieve several items located in a highly realistic VE, resembling a large, eight-room apartment (Figure 1). 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 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 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.

Guilty participants received instructions that 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 steal, and so forth). One more probe (i.e., broken bowl) was created while guilty participant carried out their crime. 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 arrive at a room, 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. Immediately after exiting the room, participants were intercepted, accused of having been in the room without authorization, “apprehended,” and escorted 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 re-emphasize probe items, by presenting them amongst a backdrop of items (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 have any special significance attached to them, 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 study¹.

Following the interrogation, participants learned a list of 12 target words to perfection and completed a recognition task approximately 20 minutes later to ensure that target words were as well learned as probe items for guilty participants.

Countermeasures: Participants in each of three countermeasures groups received specific instructions in advance of the laboratory visit to carry out tactics designed to interfere with ERP discrimination of familiar from unfamiliar items, by changing the significance of stimuli. Participants received detailed countermeasures instructions with their email detailing their “mission.” These instructions educated participants about the fact that there would be probe, target, and distracter stimuli, and how to identify them in the string of stimuli 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. The second group (CM2) was instructed to apply pressure to their toes whenever a target item was displayed. 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 until all stimuli were presented.

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 K Ω or less with inter-electrode differences at 1 K Ω 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 eye-blink correction, epochs with EOG deflections exceeding $\pm 75 \mu$ volts from baseline were rejected as a subsequent precaution against including ocular artifacts such as eye blinks. Files were epoched from -250 ms prestimulus onset to 1750 ms poststimulus, linearly detrended, baseline corrected to a prestimulus baseline of -250 to 0 ms and finally re-referenced offline to linked mastoids. Epochs free of artifacts or incorrect responses were then averaged by condition to create ERPs for each stimulus type.

Stimuli

Seventy-two word stimuli were selected for this study and consisted of twelve crime-relevant items (i.e., probes; see appendix), twelve items associated with the crime scene but not relevant for the crime (i.e., targets) and 48 distracter items (i.e., unlearned),

two 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 ($F[5,50]=.02$; *ns*). Furthermore, word lists did not differ in word frequency ($F[5,20]=.153$; *ns*) 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 degrees, and 2.06 degrees, 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 1000 ms. Participants indicated recognition of targets by pressing a button with the thumb of their dominant hand and pressed with a thumb of their non-dominant hand for all other items (i.e., probes and distracters). Double responses were excluded from the analysis because of the concomitant distortion of P3 amplitude and latency.

Stimuli were presented in four blocks with serial position randomized within and across blocks. Each block consisted of 3 probes, 3 targets, and 12 distracters (two

matched to each probe and target presented). After presentation of the 18 items within a block, a new block of randomized items was presented. Four blocks presented all stimuli, yielding 72 presentations. The blocks were repeated once before a self-paced break, after which the entire procedure was randomized and repeated again, yielding 288 presentations or 48 possible ERP trials per stimulus type (i.e., probe, target, distracter 1 through distracter 4). 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. 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 (Rosenfeld, Soskins, Bosh, & Ryan, 2004), each of which is detailed below. The statistical outcomes were also compared to each other utilizing receiver operating characteristic (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, 1st, 2nd derivative of P3 amplitude) that have 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 amongst indicators would decrease this probability. Various indicators (Figure 2) were converted to within-participant z-scores to reduce irrelevant individual differences, and enhance the pattern of response for each participant. Cutpoints that maximally differentiated learned from unlearned items in the validation study of Allen et al. (1992) were then used to differentiate recognized from unfamiliar items. 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. Bootstrap statistics were thus computed by creating ERP averages through repeated sampling with replacement from the pool of all available sweeps for each item type (i.e., probes, targets, distracters). These averages are then used to create a probe-target and probe-distracter double-centered cross-correlation during each of 100 iterations. Each pair of cross-correlations is then compared to each other and a count taken for each probe-distracter correlation that is larger than a probe-target correlation. Farwell and Donchin (1991) found that counts below 10% for cross-correlations indicating familiarity (larger probe-target correlation), and tallies above 70% for cross-correlations indicating unfamiliarity (larger probe-distracter correlation), allowed them to identify 90% of guilty participants and 85% of innocent participants, respectively. Values above 10% but below 70% are deemed indeterminate resulting in 10% of guilty and 15% of innocent participants being classified

as indeterminate in the Farwell and Donchin (1991) study.

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 less than -1.65 standard deviations taken to indicate with 95 percent 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 5th 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 distracters, such a participant would not be deemed guilty. Rosenfeld (2004) 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 individual 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 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 itemtype, $F(2.148)=94.9$, $p<.001$) was qualified by an Itemtype by 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 did not transfer

into 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 Itemtype, $F(2,148)=186.2$, $p<.001$), an effect that was not qualified by group (Itemtype by 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, while 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 increase of P3 amplitude from frontal to parietal sites. Moreover, innocent participants (left column) produced the predicted waveform pattern of large P3 amplitude in response to target items, while 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 Itemtype (Probe, Target, Distracter) and Site (Fz, Cz, Pz) as repeated measures factors and Group (INN, CM0, CM1, CM2, CM3) as a between subjects factor was conducted. Main effects of Itemtype ($F[2,140] = 171.87$, $p< 0.01$) and Site ($F[2,140] = 170.61$, $p< 0.01$) were observed, in addition to several

significant interactions: Group by Itemtype ($F[8,140] = 2.82, p < 0.01$), Itemtype by Site ($F[4,280] = 14.49, p < 0.01$) and Group by Itemtype by Site ($F[16,280] = 3.031, p < 0.01$). To decompose the interactions, separate Itemtype by Site ANOVAs were run for each group separately. Details of the results are summarized in Table 2. Because the rationale of the P3-passed deception detection is based on a maximal effect at site Pz, post hoc analyses for each group involving P3 amplitudes at site Pz revealed that for all groups, targets produced significantly larger P3 amplitudes than distracter items (all p 's $< .001$) and significantly larger P3 amplitudes than probe items (all p 's $< .001$). 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 $p < .35$).

Individual Classification

Table 3 presents the percent 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. By 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 studies³.

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 p 's $> .12$). Because this analysis treated each countermeasure group (CM1, CM2, CM3) as a distinct group, the analysis may have lacked power. The analysis was rerun collapsing all countermeasure groups into a single group, thus assessing the Chi-square for Group (No countermeasure Vs Countermeasure) by Verdict (Guilty vs. Innocent or Guilty vs. Innocent vs. Indeterminate for bootstrapping of correlations). A significant effect of countermeasures on hitrate 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 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. As seen in Figure 4, countermeasures were generally effective at reducing the discrimination between guilty and innocent participants. For the Bayesian method, the Area Under the Curve (AUC) for the three countermeasure groups (CM1, CM2, and CM3) were all significantly smaller ($p < .01$) than that for the standard guilty group (CM0). The CM2 group additionally had significantly ($p < .01$) smaller AUC than the other two CM groups, while the latter did not significantly differ from one another. For the bootstrapping of peak-to-peak amplitudes, the three countermeasure groups again had significantly ($p < .01$) smaller AUCs than the standard guilty group, but the three countermeasure 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 countermeasure and the standard guilty group, while these latter three groups did differing significantly from one another.

DISCUSSION

The present study was conducted to expand 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 traditional approaches was used to assess if physical, mental, or a combination of these countermeasures would be sufficient to elude detection. Statistical approaches were expanded 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. The three approaches to analysis – Bayesian and the two bootstrapping approaches – 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,” additionally produced many indeterminate verdicts. 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 were 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.

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, 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. Although a pilot study found no statistical differences in how frequently naïve participants selected probes over distracter items, it is possible that items chosen had stereotypical characteristics that slowed reaction times but failed to emerge on an electrocortical level. The behavioral findings are informative, however, as they exemplify the difficulty field studies would have to overcome to produce a stimulus set beyond reproach. 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, bona fide innocent participants would have been found guilty at unacceptably high rates. The implications of this finding are 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 the capacity (e.g., Allen et al., 1992; Seymour, Seifert, Shafto, & Mosmann, 2000). The other 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 be 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. The call for a simplified stimulus set (Rosenfeld et al., 2004) comprised of fewer items might address some of the problems documented in this study, but runs the additional risk that with limited number of probe items, idiosyncratic responses may lead innocent participants to appear guilty at higher rates than with larger stimulus sets.

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 limitation 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 to ERPs in naturalistic settings.

Furthermore, the successful use of VE in combat training (Lampton, Clark, & Knerr, 2003; Pleban, Matthews, Salter, & Eakin, 2002), pilot training and assessment (Bennett, Schreiber, & Andrews, 2002) or treatment of psychiatric disorders (Klinger et al., 2005; Krijn, Emmelkamp, Olafsson, & Biemond, 2004) (Lee et al., 2004; North,

North, & Coble, 1998; Rothbaum et al., 1999; Rothbaum, Hodges, Ready, Graap, & Alarcon, 2001) attest to the effectiveness of VE to mimic naturalistic environments. Lastly, 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. Thus, it seems unlikely that the low hit rates were solely attributable by partial use of virtual technology.

Another possible explanation for the low hit rates concerns the target stimuli. Because participants learned probe items to perfection, targets were learned at a similar level of proficiency by way of a multiple-trial list-learning task. Since targets were most recently learned, and considerable emphasis was placed on correct encoding, it is possible that the salience of targets relative to probes was increased, altering amplitude differences between these items, 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. It is worth noting that Farwell (personal communication, January 2001) does not extensively review targets, but rather relies on very brief presentation, thus leading guilty participants to have some confusion over the probes versus targets. 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 over-embellishing 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 email reminder of his mission, most participants described the study as realistic and anxiety provoking, particularly since 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 sizeable reward. Carmel, Dayan, Naveh, Raveh, and Ben-Shakhar (2003) noted that a more realistic mock crime was associated with lower recall rate and weaker detection efficiency than the standard mock crime involving theft of jewelry or money from an unoccupied office. Carmel et al. argued that standard mock crime scenarios inherently maximize recall of crime relevant information as participants are exposed to a limited amount of information, learn all crime relevant items to perfection, and are tested immediately afterwards. Unfortunately, true crime does not afford most of such luxuries as participants may not encode items the way they are presented subsequently (e.g., crime being committed in a dark room voiding color or other information, anxiety interferes with encoding), or long passes between the crime and examination, or participants are intoxicated. However, realistic mock crime scenarios such as the one produced by Ginton, Daie, Elaad & Ben-Shakhar (1982) are unlikely to pass current moral and ethical standards, forcing investigators to come up with less troublesome compromises.

Results of this paradigm offer suggestions regarding the use of countermeasures. Although physical countermeasures are easy to implement, they should be relatively easy

to detect in an applied setting through the use of pressure sensors (Lafayette, 2002) or baseline EMG to detect muscle activity. Results also suggested that the complex countermeasure modeled after Rosenfeld et al. (2004) appeared too difficult to implement for most participants without specialized training. Exit interview data revealed that only 53% of participants executed CM3 more or less as instructed, while 47% relied on impromptu strategies that may or may have not been successful. Such heterogeneous behavior reduced the applicability of this measure in a controlled setting. Nevertheless, the difficulty level of this task may be its inherent strength. Participants may assign a disproportionate amount of resources on how to perform the task, which may reduce the salience of the probe items. Further work to produce high-load measures that can be reliably and imperceptibly executed is needed.

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 (Bureau of Justice Statistics, 2002), 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 ERP-based deception detection procedure may be vulnerable to relative 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.

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 polygraphy, changes in facial expression (Ekman & Friesen, 1974; Ekman, Friesen, & O'Sullivan, 1988) (Cheng & Broadhurst, 2005), examination of demeanor (Granhag & Strömwall, 2002; Pollina &

Squires, 1998; Vrij, Edward, & Bull, 2001; Vrij, Semin, & Bull, 1996), thermography (Pavlidis, Eberhardt, & Levine, 2002), and voice stress analysis (Meyerhoff, Saviolakis, Koenig, & Youric, 2001). 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; Rosenfeld et al., 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. 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 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 since they almost exclusively occur among the guilty participants. This stands in stark contrast to standard 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. By 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. Such double-dissociation suggests the possibility of combining the two approaches – standard control-question technique polygraphy with ERP-based guilty knowledge approach – thereby overcoming the limitations of each. If both tests were administered, consistent guilt on both tests of consistent innocence on both tests would in fact be conclusive, but inconsistent verdicts across tests would be inconclusive. A guilty verdict, although relatively uninformative from the CQT would in fact only be likely to have come from a guilty participant using the ERP-GKT. Conversely, an innocent verdict, although uninformative from the ERP-GKT, would only be likely to have come from an innocent participant with the standard CQT. Whether or not these two approaches could be combined, for example administering one after the other, without altering the accuracy of either test, remains an empirical question.

APPENDIX A

Summary of 12 probe, 12 targets, and 48 unlearned distracter items.

Category	Item Type					
	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

Footnotes

¹ A total of seven participants were excluded at various stages of the experiment and were not included in the present dataset. Four participants broke their respective blinds, one participant was observed using notes carrying out their “mission”, and two participants were unable to enter the virtual environment.

²After 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 email.

³Farwell (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%.

APPENDIX B

Table 1

Mean reaction times (ms) and hit rates (%) for three stimulus types across five experimental conditions and three stimulus types.

Condition	Mean Values (SD in parenthesis)					
	Reaction Time			Incorrect Responses		
	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)

Table 2

Summary of main and interaction effects of P3 amplitude for 79 participants across five experimental conditions for stimulus type and electrode site.

Condition	Type		Site		Type x Site	
	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	ns
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*

* $p. \leq 0.01$

** $p. \leq 0.5$

Table 3

Hit rates (%) of participants determined guilty, innocent, or indeterminate across three classification procedures.

[illegible]

Figure Captions

Figure 1. Sample screenshot of an 8-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).

Figure 2. Frequency distribution of Z-scores 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 countermeasure groups (CM1, CM2, CM3).

Figure 3. Grand average waveforms across three midlines sites for Probe, Target, and Distracter items across innocent (INN), standard guilty (CM0), and three countermeasure groups (CM1, CM2, CM3).

Figure 4. Receiver operating characteristic (ROC) curves to compare detection efficiency of three classification methods for four guilty conditions. Legend displays Area Under the Curve (AUC) and associated standard deviation.

Figure 1

Figure 2

Figure 3

Figure 4

REFERENCES

- 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(5), 504-522.
- Bashore, T. R., & Rapp, P. E. (1993). Are there alternatives to traditional polygraph procedures? Psychological Bulletin, 113(1), 3-22.
- Bayes, T. (1763). Essay towards solving a problem in the doctrine of chances. Philosophical Transactions, 53, Royal Society: London (Reprinted in 1958 in *Biometrika*, 45, 293-315), 370-418
- Bennett, W. J., Schreiber, B. T., & Andrews, D. H. (2002). Developing competency-based methods for near-real-time air combat problem solving assessment. Computers in Human Behavior, 18(6), 773-782.
- Ben-Shakhar, G., & Dolev, K. (1996). Psychophysiological detection through the guilty knowledge technique: Effect of mental countermeasures. Journal of Applied Psychology, 81(3), 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): 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(3), 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(4), 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(1), 107-118.

Donchin, E., & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? Behavioral and Brain Sciences, 11(3), 357-427.

Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. Journal of Personality and Social Psychology, Vol. 29(3), 288-298.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. Journal of Personality and Social Psychology, 54(3), 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(2), 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(2), 683-691.

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ('lie detection') with event-related brain potentials. Psychophysiology, 28(5), 531-547.

Farwell, L. A., Inventor. (1994). Method and Apparatus for Multifaceted Electroencephalographic Response Analysis (MERA), US patent 5,363,858. 1994 Nov.15.

Farwell, L. A., Inventor. (1995a). Method and Apparatus for Truth Detection, US Patent 5,406,956. 1995 Apr. 18.

Farwell, L. A., Inventor. (1995b). Method for Electroencephalographic Information Detection, US patent 5,467,777. 1995 Nov.21.

Farwell, L. A., & Smith, S. S. (2001). Using brain MERMER testing to detect knowledge despite efforts to conceal. Journal of Forensic Sciences, 46(1), 135-143.

Feder, B. J. (2001, October 9th). Truth and Justice, by the Blip of a Brain Wave. The 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(3), 243-257.

Ginton, Daie, Elaad & Ben-Shakhar (1982). A Method for Evaluating the Use of the Polygraph in a Real-Life Situation. Journal of Applied Psychology, 67 (2), 131-137.

Halperin, D (2002, December 6). Police Beat: Suspicious Email Sent. The Daily Wildcat, [On-line} Available at 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(1), 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(1), 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(3), 241-247.

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(1), 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(3), 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): Guilford Press.

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, NY: Cambridge University Press.

Johnson, R. (1986). A Triarchic Model of P300 amplitude. Psychophysiology, 23(4), 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(3), 217-253.

Johnson, R. J., Barnhardt, J., & Zhu, J. (2004). The contribution of executive processes to deceptive responding. Neuropsychologia, 42(7), 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(3), 386-404.

Klinger, E., Bouchard, S., Légeron, P., Roy, S., Lauer, F., Chemin, I., & Nugues, P. (2005). Virtual Reality Therapy Versus Cognitive Behavior Therapy for Social Phobia: A Preliminary Controlled Study. CyberPsychology & Behavior, 8(1), 76-88.

Krijn, M., Emmelkamp, P. M. G., Olafsson, R. P., & Biemond, R. (2004). Virtual reality exposure therapy of anxiety disorders: A review. Clinical Psychology Review, 24(3), 259-281.

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(5), 428-435.

Lampton, D. R., Clark, B. R., & Knerr, B. W. (2003). Urban Combat: The Ultimate Extreme Environment. Journal of Human Performance in Extreme Environments, 7(2), 57-62.

Lafayette, I. (2002). LX3000 Activity Sensor. [On-line] Available at <http://www.lafayetteinstrument.com/polyactivitysensor.htm>.

Lee, J., Lim, Y., Graham, S. J., Kim, G., Wiederhold, B. K., Wiederhold, M. D., Kim, I. Y., & Kim, S. I. (2004). Nicotine Craving and Cue Exposure Therapy by Using Virtual Environments. CyberPsychology & Behavior, 7(6), 705-713.

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(3), 195-204.

Lee, J. M., Ku, J. H., Jang, D. P., Kim, D. H., Choi, Y. H., Kim, I. Y., & Kim, S.

I. (2002). Virtual reality system for treatment of the fear of public speaking using image-based rendering and moving pictures. *CyberPsychology & Behavior*, 5(3), 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): Sage Publications, Inc.

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): Jessica Kingsley Publishers, Ltd.

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(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(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(6), 488-497.

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(1), 13-26.

Moore, B. A., & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury*, 18(10), 975-984.

North, M. M., North, S. M., & Coble, J. R. (1998). Virtual reality therapy: An effective treatment for phobias. In G. Riva & B. K. Wiederhold & E. Molinari (Eds.),

Virtual environments in clinical psychology and neuroscience: Methods and techniques in advanced patient-therapist interaction. (pp. 112-119): IOS Press.

Office of Technology Assessment (1983). Scientific Validity of Polygraph Testing: A Research Review and Evaluation - A Technical Memorandum (Publication OTA-TM-H-15). Washington, D.C: US, Library of Congress.

Pavlidis, I., Eberhardt, N. L., & Levine, J. A. (2002). Seeing through the face of deception. Nature, 415(6867), 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(9), 598-603.

Pleban, R. J., Matthews, M. D., Salter, M. S., & Eakin, D. E. (2002). Training and assessing complex decision making in a virtual environment. Perceptual and Motor Skills, 94(3), 871-882.

Pollina, D. A., & Squires, N. K. (1998). Many-valued logic and event-related potentials. Brain and Language, 63(3), 321-345.

Rosenfeld, J. P. (1995). Alternative views of Bashore and Rapp's (1993) alternatives to traditional polygraphy: A critique. Psychological Bulletin, 117(1), 159-166.

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(3), 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(1), 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(1), 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(2), 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(2), 163-179.

Rothbaum, B. O., Hodges, L., Alarcon, R., Ready, D., Shahr, F., Graap, K., Pair, J., Hebert, P., Gotz, D., Wills, B., & Baltzell, D. (1999). Virtual reality exposure therapy for PTSD Vietnam veterans: A case study. Journal of Traumatic Stress, 12(2), 263-271.

Rothbaum, B. O., Hodges, L. F., Ready, D., Graap, K., & Alarcon, R. D. (2001). Virtual reality exposure therapy for Vietnam veterans with posttraumatic stress disorder. Journal of Clinical Psychiatry, 62(8), 617-622.

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(4), 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(6), 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(1), 30-37.

Scheeres, J. (2001). Thought Police Peek into Brains [On-line]. Available at www.wirednews.com. Available: www.wirednews.com October 5].

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), 1-9.

Vrij, A., Edward, K., & Bull, R. (2001). People's insight into their own behaviour and speech content while lying. British Journal of Psychology, 92(2), 373-389.

Vrij, A., Semin, G. n. R., & Bull, R. (1996). Insight into behavior displayed

during deception. Human Communication Research, 22(4), 544-562.

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. Psychophysiology, 26(2), 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(1), 77-82.

Zirinsky, S. (2002, June 14). "It's All In Your Head", 48 Hours. New York, NY: Columbia Broadcasting Services