PSYCHOPHYSIOLOGY
Copyright © 1991 by The Society for Psychophysiological Research, Inc.

Vol. 28, No. 3
Printed in U.S.A.

An ERP-Based, Control-Question Lie Detector Analog: Algorithms for Discriminating Effects Within Individuals' Average Waveforms

J. PETER ROSENFELD, ANDREA ANGELL, MARY JOHNSON,
Department of Psychology

AND JIA-HE QIAN
Department of Statistics, Northwestern University

ABSTRACT

Two experimental, P3-based analog control question tests were run. In both, guilty subjects were presented with a set of seven phrases describing antisocial acts of which they were innocent, plus one phrase describing a guilty act (the analog *relevant* question), and one act to which a "yes" response (yes-target stimulus) was required to assure attention. Innocent subjects (run only in Experiment 1) saw all innocent acts plus the yes-target act. Thus nine acts were seen by guilty and innocent subjects. In both experiments, all subjects had to *selectively* review their guilty acts privately. Also in both experiments, all subjects were especially questioned about four acts of which guilty subjects were known to be innocent of all but one, and of which innocent subjects were known to be innocent of all. (These falsely accused acts were regarded as *control* question analogs.) In Experiment 1, the private review and rehearsal took place on the same day as the main test. In Experiment 2, one subgroup (delay-only) of guilty subjects was run as in Experiment 1, except that the private review-rehearsal was separated from the main run by 7–14 days. Another subgroup (delay-rehearsal) of guilty subjects was run just as was the subgroup delay-only, except that the delay-rehearsal subgroup additionally received a *non-selective* additional interrogation/rehearsal on the delayed main run day. Parietally maximal P3 responses were obtained to yes-target items in all groups. In Experiment 1, only in the guilty group was the relevant-minus-control P3 amplitude difference significant. In Experiment 2, the difference was significant only in the delay-rehearsal subgroup. A four-step algorithm (involving relevant-control amplitude differences and relevant target vs. control-target cross-correlations) was used to assess effects within individuals. In Experiment 1, 12 of 13 guilty subjects and 13 of 15 innocent subjects were correctly diagnosed. In Experiment 2, 3 of 8 delay-only subjects and 7 of 8 delay-rehearsal subjects were correctly diagnosed. In Experiment 2, the relevant-minus-control group P3 amplitude difference was significant in the delay-rehearsal but not in the delay-only subgroup. The results suggest that temporally proximal, non-selective rehearsal procedures are sufficient to activate personal knowledge of a salient (oddball), P3-generating stimulus phrase, and that even selective rehearsal of guilty acts is not sufficient without temporal proximity.

DESCRIPTORS: Event-related potentials, P3, Lie detection, Control Question Test.

The two basic procedures currently used for physiological detection of deception are the guilty knowledge and control question tests (Lykken, 1959; Reid & Inbau, 1977; Raskin, 1986). In the former, a subject (usually male) is presented with a series of test items of which one or more are related to a crime, e.g., the murder weapon, stolen item, etc. If the subject has specific guilty knowledge, he is expected to react differentially to the key items, whereas an innocent subject, ignorant of the crime's details, is expected to react no differently to the key items than to the other members of the item series. The term "react" refers to emotional

A preliminary report of some of these data was presented at the 29th annual meeting of the Society for Psychophysiological Research, New Orleans, October 1989.

We thank Albert Erlebacher and Robert Boruch for statistical consultation, and David Paik for computer processing and programming assistance.

Address requests for reprints to: J. Peter Rosenfeld, Northwestern University, Department of Psychology, 2029 Sheridan Road, Evanston, Illinois 60201.

reactions and their putative physiological correlates.

Because there are situations in which the guilty knowledge test is not feasibly done (e.g., the details of a crime are published so that innocent persons could acquire guilty knowledge and show guilty reactions), the control question test is more widely used. In this procedure, suspects are asked *relevant*, *control*, and *irrelevant* questions. A *relevant* question asks about specific acts; e.g., "Did you steal the jewels from Smith's store on January 20, 1987?" A *control* question asks about general antisocial acts that all persons have probably done or seriously considered at one time or another in their lives; e.g., "Did you ever steal anything?" An *irrelevant* question is neutral and intends to establish response baselines; e.g., "Are you in Chicago today?" In standard practice it is assumed that a *guilty* person will tend to respond more vigorously to relevant than to control questions because he is expected to be more concerned about detection for the specific crime that he knows he has committed and about which he is being interrogated. Control questions are assumed not to concern a guilty person whose attention remains focused on his actual crime. In contrast, even though he may be subjected to the same interrogation, an *innocent* subject is expected to be more concerned about and thus more reactive to control than to relevant questions. This is because the questioner has presumably persuaded him that his responses to both control and relevant questions are important. The guilty subject is believed to be mainly worried about his actual crime but the innocent subject knows that he is innocent of that crime and remains concerned about the control question because he has probably had some experience in the control question area at some time in his life.

The usual dependent measures in both the guilty knowledge and control question tests are respiratory, electrodermal, and cardiovascular responses that are assumed to be correlates of the emotional state triggered by confrontation with one's guilt. Frequent criticism of these procedures (Kleinmuntz & Szucko, 1984; Saxe, Dougherty, & Cross, 1985; Furedy, 1986; Ekman, 1985) prompted development of a new guilty knowledge test, which utilized late positivity in the event-related brain potential (ERP) as the response index (Rosenfeld, Nasman, Whalen, Cantwell, & Mazzeri, 1987). As discussed in that report, the ERP response is believed to index cognitive in addition to or instead of emotional activity and may be less subject to some of the criticism that has been levied against current polygraphic methodology. ERP-based guilty knowledge tests have also been described by Farwell and Donchin (1986, 1988) and Forth, Hart, Hare, and Harpur (1988), and we extended our 1987 report in a subsequent study (Rosenfeld et al., 1988).

In our earlier studies, subjects chose one item from a box of nine items and were asked to pretend that they stole it and were being examined using a polygraph. They were encouraged to beat the test. Then they watched a display which presented at random one of nine words every two seconds while ERPs were recorded. One of the nine words corresponded to the chosen "stolen" item; the eight others described novel items. The chosen item was a deviant "oddball" stimulus (Duncan-Johnson & Donchin, 1977) because it alone had been among the nine items recently exposed to and chosen by the subject. Thus it was expected to and did evoke the late positivity in the ERP usually referred to as the P3 component. Control subjects, who saw an experimenter-chosen novel item in place of the actually chosen item, did not show P3 responses.

The deception detection test described here is closer to a control question test of the type used in pre-employment screening situations. We did *not* (in the present report) request that subjects imagine having committed a pretended crime, nor did we ask subjects to commit a mock crime as is commonly done in laboratory analogs of control question and guilty knowledge tests. What we did was to ask subjects about nine undesirable acts with reasonable probabilities in our student-subject population, e.g., cheating on tests, using false identifications, etc. We arranged a situation in which we voiced our suspicion that a given subject may have done four of nine possible acts. In the operationally defined guilty group, we further arranged the situation so that the subjects would be actually guilty of just one of the four accused acts, and of no other among the nine possible acts. In the operationally defined innocent group, the situation was arranged so that the subjects would be innocent of all nine acts. The falsely accused (but innocent) items were intended to serve as analog control questions for all subjects: It was thus expected that guilty subjects would not be as responsive to these as to the relevant question, but that to an innocent subject, all falsely accused items would have equivalent P3-evoking potency. In other words, to a guilty but not an innocent subject, the relevant question would have the oddball quality of special meaning and thus evoke the P3 wave. The hypothesis tested was that the difference in P3 amplitude between relevant and control items would be greater in guilty than in innocent subjects. The remaining non-accused acts on which all subjects were innocent were regarded as irrelevant question analogs.

EXPERIMENT 1

Method

Subjects

The subjects were 32 males, aged 18–22 years ($\bar{X}=$ 18.8), obtained from an introductory psychology class, at Northwestern University, who were fulfilling a research participation requirement. All had normal or corrected vision. Sixteen subjects each were randomly assigned to the guilty and innocent groups.

Procedure

Upon entering the lab, the subjects signed a consent form which contained general information about brain wave recording studies, and in particular, the following paragraphs:

> "I understand that I may be asked to respond on a list to personal questions about my behavior and integrity, although I know that no record of my answers is kept, and my responses will remain anonymous."
>
> "I further understand that I may withdraw from the study at any time without prejudice or penalty. I further understand that one of the experimenters will observe me throughout the study, and I am free to inquire any time about any aspect of the study."

The subjects were then led into a room with a recliner and recording equipment and electrodes were applied while the experimenter explained how our laboratory became interested in detection of deception. The aim of this explanation was to impart a serious attitude. Next the experimenter gave the subject a list of 13 acts, with check boxes next to each:

1. "SMOKED POT MONTHLY", 2. "STOLEN A BICYCLE", 3. "CHEATED DURING TEST", 4. "TOOK SCHOOL RECORDS", 5. "USED FALSE MEDICAL", 6. "STOLE AN AUTOMOBILE", 7. "FAILED ONE COURSE", 8. "STOLE SOME CLOTHES", 9. "PLAGIARIZED A PAPER", 10. "WAS COMPUTER CHEAT", 11. "TOOK FRIEND'S MONEY", 12. "USED FALSIFIED I.D.", 13. "BROKEN POP MACHINE."

When the experimenter gave the subject the list, he informed the subject that he would leave the room and shut the door, following which the subject was to turn on a cassette recorder and listen to the loaded tape, which would detail the meanings of the listed acts, as well as instruct the subject about checking "yes" or "no" boxes next to each item. Possibly ambiguous items (e.g., "USED FALSE MEDICAL") were explained (e.g., " 'Used False Medical' means presenting a forged medical note to avoid an exam or term paper deadline") in the tape. All items were defined with respect to a five year period dating back from the date of the subject's present appearance in the lab. Subjects were instructed on the tape to check "yes" or "no" only when they were certain; otherwise they were to write a question mark. Subjects were told that the point of this list filling was to help them clarify in their own minds what acts they were and were not guilty of, and that they could destroy or retain their lists after completion.

.Although some of the listed items (1,3,5,7,9,10,11,12,13) were known from pilot studies to have actual probabilities of 10–50% and others were known to have probabilities <2% (2,4,6,8) in our student population, all items were estimated by subjects to have similar, finite (Mean=22.7%) probabilities, i.e., to represent acts that one might reasonably suspect at least some members of the subject population to have been involved in at one time or another. Our aim in development of this list was to make it likely that most subjects would be guilty of 0–3 items. This would make it possible, in the main control question test analog to come later, to present guilty subjects with a set of items of which only one was a guilty item, and to present innocent subjects with a set of all innocent items. Subjects guilty of more than five acts could not be run because such sets could not be developed with these subjects. Our knowledge of guilt or innocence on the listed items was thus essential not only as *ground truth* to validate our ERP test results, but also to arrange for the appropriate item sets to be given to guilty and innocent subjects.

We obtained this knowledge by secret television surveillance of each subject's list as he checked a "yes" or "no" box next to each listed act. Later debriefing revealed (based on subjects' verbal reports) that all but one of the subjects believed themselves to be unobserved and alone while checking the list boxes. (The exceptional subject's data were not used.) Moreover, in an unpublished pilot study with similar methods, one final, nine-item questionnaire was given at the end of the study (but prior to any debriefing) to this other sample of 30 subjects from the same population. There was only one item of real interest to us on this questionnaire: "I am comfortable that my privacy was respected in this study." (The preceding four and subsequent four items related to subjects' physical comfort, understanding of instructions, experimenter courtesy, etc.) All subjects except two checked "4" or "5" on a 1–5 scale of agreement; the two exceptions checked "3". Thus although it was true that, as we told subjects, the list-filling was intended to make clear in their own minds what their guilty acts were, it was also true (and not clearly told to subjects) that we would be observing their lists so as to (a) arrange our stimulus sets, and (b) ascertain a validating "ground truth" record.

Following the list-fill procedure, subjects watched a video display terminal in privacy while each of eight selected items were flashed on the screen for one second each. No recording was done, but subjects were led to believe that we *were* recording ERPs. Subjects were told to press a button on a counter (unconnected to anything) with their dominant hand if they saw a guilty item. Otherwise, they were to press another unconnected counter-button with their other hand. They were told (truly) that their button responses would be unobserved, and that we wanted them to imprint firmly in their minds which acts they were or were not guilty of prior to the final main run. We also wanted them to believe we were collecting ERP data during the rehearsal, as explained below. In this last rehearsal procedure (hereafter called *second-rehearsal procedure*), guilty subjects saw the eight acts that they would see later on their actual control question test analog. (As described below, there was a ninth item additionally used in this main test, the "yes-target" item.) The set of acts included one guilty item (checked "yes"), three more probable acts with >10% probabilities, but of which the particular subject was innocent, and four improbable acts with <2% probabilities and of which the subject was innocent. Innocent subjects saw the same set during second-rehearsal, but on their actual test to be run later, the guilty item was replaced with a high probability (>10%) innocent item. Thus in the second-rehearsal procedure, a variable mix of guilty and innocent items were removed from the original list, the pattern of removed items varying with the number checked "yes" by a given subject. A *t*-test on the number of guilty items checked "yes" by innocent versus guilty subjects failed to reach significance (p>.3). The mean number of acts checked "yes" was 2.44 across all subjects (see Table 1).

It is acknowledged that the *potentially confounding, second rehearsal and list-fill procedures could not be utilized in the field.* This is considered below in the discussion of Experiment 1, and introduction to Experiment 2, which deals directly with the issue.

Three subjects originally assigned to the guilty group were revealed via surveillance to be guilty of more than five of the nine available high probability acts, and could not be run. One of the 16 originally assigned innocent subjects had 148 artifacts in 256 trials (>50%): Following our *a priori* 50% maximum artifact rate tolerated rule, data from this subject were not considered further. Thus data will be reported for 13 guilty and 15 innocent subjects.

At this point each subject, guilty or innocent, was told that he would be taking a lie detector test of the type used by government agencies to make certain that prospective employees were of sound character and integrity. The subject was told that his goal in the rest of the experiment was to honestly pass or else beat the test (if necessary) so as to obtain a high paying, high responsibility, hypothetical job.

Following this pretest manipulation was our analog accusation/interrogation procedure: The experimenter made one *primary* and three *secondary* accusations: "Based on preliminary data [the bogus ERPs collected during the second rehearsal procedure], we suspect you committed [Act A], but you may also have done [Act B], [Act C], or [Act D]." (Actual acts were given instead of letters in the preceding and subsequent material.) For innocent subjects, Acts A, B, C, and D were high probability acts of which the subject was known (from television surveillance) to be innocent. For guilty subjects, Act B was a guilty act, and the subject was innocent of A, C, and D; again, all (A–D) acts were high probability. We then proceeded to ask subjects if they knew people who committed each of the four acts A–D, if they ever thought about such acts, and what they thought of people who commit such acts. Each question was put once about each act in the order A, B, C, D. Our critical comparison for diagnosis of guilt/innocence in each subject would be the difference between P3 responses to B, the analog relevant question, and C, the secondary control question analog in the middle of the accusation order along with B, the relevant item. A P3 response to A might be expected on the basis of its primary order position and *primary accusational* value. It too could be viewed as a control question analog, but because it was treated uniquely, it was not believed to be the appropriate comparison to make with B. On the other hand, C was in the middle of the accusation order and treated similarly to B in that both were used for *secondary* accusations; i.e., in innocent subjects, we predict B=C but in guilty subjects we predict B>C because B is uniquely a guilty item to which the guilty but not innocent subject must (uniquely) lie in order to escape detection. We utilized A for primary accusation mainly to focus an innocent subject's attention away from the secondarily accused items. We expected some of this effect also for guilty subjects, but reasoned that the relevant (B) guilty item would still retain adequate uniqueness to elicit a sizable P3.

In the final phase following the interrogation/accusation, the experimenter said "We still think you did [A], but could have possibly done [B], [C], or [D]. This last test should provide the answer:

Table 1
Group and stimulus attributes

	Number of Sweeps Per Stimulus Type					Number of Artifacts Per Stimulus Type									
Group	REL	FAL	SEC	IRL	TBY	REL	FAL	SEC	IRL	TBY	Total Artifacts	"Yes", Prob	"?", Prob	"Yes", Impr	"?", Impr
Guilty	10.4	11.5	12.6	11.7	11.2	3.4	3.7	3.2	3.2	3.2	27.6 (20%) Range: 6–89	2.6 Range: 0–5	0.46 Range: 0–2	0 Range: 0	0.3 Range: 0
Innocent	10.8	11.9	12.7	12.8	11.1	4.5	4.2	3.5	3.4	3.1	33.7 (24%) Range: 7–74	2.3 Range: 0–5	0.4 Range: 0–2	0 Range: 0	0 Range: 0

Note.—First five columns: Mean numbers of sweeps per average to Relevant (REL), Primary Accused (FAL), Control (SEC), Irrelevant (IRL), and Target (TBY) stimuli. Second five columns: Mean numbers of eye-movement artifacts (rejected trials) for five stimulus types. Next column: Mean total, rate, and range of artifacts (rejected trials) to all stimuli, "Yes", prob.=Mean number and range of probable acts checked "yes"; "?", Prob., same for probable items to which subjects entered "?". Last two columns: same as preceding two but for improbable acts.

"We will ask you to watch this screen again and the various questions will come up several times each. You say 'yes' or 'no' to the questions by pressing one of these two buttons. [The experimenter here indicated the response buttons next to the dominant ('yes') and non-dominant ('no') hands.] Of course you will want to say 'no' all the time because you are concerned that if you admit guilt on one of these antisocial acts, you won't get the job you want. If you're guilty of an item you'll press 'no' and hope you beat the test by controlling your internal reactions. If you're innocent, you'll certainly want to press 'no' because you are innocent.

"Every now and then, you'll see an item you haven't seen yet; it is 'Took deception test.' You must press 'yes' to this because you truly are taking such a test. If you fail to press 'yes,' we'll know you're not paying attention to the screen and would have to file a negative report on you in a real-life test. And of course your needing to pay attention to this item will force you to notice guilty items."

Hereafter the "Took Deception Test" item will be referred to as the yes-target item; i.e., the target item to be answered with a "yes." There were no subjects whose yes-target counts were off by more than one item, and 20 of 28 subjects' counts were 100% correct. The subject was next given the main test in which he viewed the nine items (eight acts + yes-target). The structures of the nine-item sets seen by both groups of subjects during the main run are shown in Table 2.

It is noted that the accusing/interrogating experimenter did not have prior knowledge of the possibly guilty acts of a given subject. Moreover, cassette tapes were made on a randomly selected set of interrogation sessions. Other individuals in the building housing the lab, upon listening to the taped interrogations, could not correctly identify guilty items. The lack of cues imparted during interrogation/accusation to the subject was planned in that the interrogating experimenter was coached and rehearsed not to emphasize particular acts in any subjects.

It is noted that in this study, subjects were: 1) informed about the specific nature of the questions and procedures to which they were to be exposed prior to their participation; 2) aware that they could withdraw their participation at any time and with no penalty; 3) told they would be observed throughout the study; 4) not identified with the data or their responses to ensure confidentiality; 5) able to request that their data be destroyed (i.e., not analyzed) at the completion of the study; and 6) also debriefed to the extent that all questions asked in response to our solicitation ("Do you have any questions or concerns?") prior to releasing them were answered fully. Subjects were also invited to sign a list requesting full reports of the studies such as the present one. All requests are honored. Therefore, subjects were debriefed such that any potential distress or harm incurred during the study was rectified.

Table 2

Structure of nine stimulus set given to each subject

i/g	1	Secondary Accused: Relevant item (Act B)	probable
i	2	Primary Accused item (Act A)	
i	3	Secondary Accused item (Act C)	
i	4	Secondary Accused item (Act D)	
i	5	Irrelevant item	improbable
i	6	Irrelevant item	
i	7	Irrelevant item	
i	8	Irrelevant item	
T	9	Yes-target item	

Note.—"i"=subject known to be innocent of this item; "g"=subject known guilty; T=yes-target item. Acts A, B, C, D refer to accusation order as explained in text. For guilty subjects, item 1 is "g"; it is "i" for innocent subjects.

Recording Procedures

Silver cup electrodes were placed on F_Z, C_Z, and P_Z locations, and referenced to right mastoid with the left mastoid grounded. Electrodes were also placed supra- and sub-orbitally for EOG recording. Signals were amplified 75,000 times by Grass P511-K preamplifiers with 3dB filters set to pass signals between .3 and 30 Hz.[1] As will be described, for display purposes, off-line digital filtering was done with some waveforms so that the 3dB upper cutoff was reduced to 2.89 Hz (grand averages), 4.23 Hz (individual averages), or 6.11 Hz (single sweeps). Conditioned signals were then led to an 8-bit analog-to-digital converter (Computer Continuum, Inc., Daly City, CA) interfaced to a Commodore C128 computer and sampled every 8 ms (rate=125 Hz). The C128 software systems for stimulus presentation, data acquisition, and analysis were all written by the first author (excepting high resolution displays by Darus, French, & Wallace, 1986). SYSTAT (Wilkinson, 1986) was used for group analyses. Analog-to-digital sampling routines were in 6502/6510 assembly language. Recording began 104 ms prior to stimulus presentation and ended 2.048 s later, although epochs shown in the figures below go out only to 1.92 s.

Stimulus Presentation

A table of 324 randomly selected numbers between one and nine was stored in the computer program and referred to on each trial by the program so as to determine which of the nine stimulus phrases would be presented on that trial. The numbers were previously generated off-line by a random-number generating program and placed into a table as they were generated, subject to the restriction that no two consecutive trials could contain the same number (stimulus phrase).

[1] Although we are aware that longer time constants may be preferable (Duncan-Johnson & Donchin, 1979), in our lab, lower low-pass filter settings would have required a lower gain in order that large-amplitude low-frequency oscillations not cause incoming data to leave the range (0–5 Vdc) of our 8-bit analog-to-digital converter; lower amplification would not have allowed adequate resolution. Our settings may have introduced some distortion, but because we were more concerned with discriminating ERPs between stimulus types than in delineating ERP waveshapes, we were willing to accept the limitations of our available equipment.

Trials containing EOG artifacts (signals > 40 μV) were rejected (i.e., all data erased) and replaced with the next trial number in the quasi-random table. Trials were generated until 108 were collected. Thus there were 12 trials intended for each of nine stimuli. Because of artifact replacements and consequent departures from the stored table's original order, the actual range of numbers of trials averaged for each of the nine stimulus phrases varied across subjects from 9–14. Although Table 1 shows a larger average number of sweeps for control than for relevant stimuli, this held for both guilty and innocent groups. The four waveforms recorded (F_Z, C_Z, P_Z, EOG) on each trial were averaged into nine accumulating sets of four averages, corresponding to the four electrode locations for each of the nine stimulus/phrase types utilized.

Results

Group Data

Grand averaged ERP sets within stimulus type categories and guilty and innocent groups are shown in Figure 1.

From visual inspection, it appears that prominent positive waves appear in the yes-target records of both guilty and innocent subjects as expected, but these waves were found only for the guilty subjects in response to the relevant question. Because this component is parietally maximal (confirmed statistically; see below), positive, and appears at a latency (550–650 ms) where P3 components in response to complex visual stimuli have been previously reported (e.g., Fabiani, Karis, & Donchin, 1986), we will assume the component to be P3. An apparently negative-going component appears immediately following P3 (from 1000–1420 ms). Given our low-frequency cutoff of .3 Hz (corresponding to an approximately .7-s time constant), it is possible that this component represents some distortion in P3 recovery to baseline (Duncan-Johnson & Donchin, 1979). Nevertheless, because we have found in unpublished pilot work (with the .3 Hz cutoff) that P3 amplitude estimates based on differences between this late negative component and P3 have consistently higher reliability than the standard P3 estimate referenced to prestimulus baseline, and we routinely find the negative component and the P3 peak to negatively covary in averaged ERPs, one quantitative estimate of P3 in this report will be the *peak-to-peak* difference between the negative component and P3 (see footnote 1). This procedure is furthermore consistent with our cross-correlation algorithms, utilized in data analysis within individuals, in which cross-correlation coefficients are calculated on the waveforms from 468–1420 ms, i.e., a window that includes both components. However, for evaluation of group data we will additionally utilize the standard baseline-to-peak P3 measure in which the value of the positive component noted above is subtracted from the prestimulus baseline average. Specifically, for all ERPs, a first search window from 468–1052 ms is utilized. The maximum positive 104-ms segment in this time window is taken as the positive component amplitude. The midpoint of this maximum segment is defined as P3 latency. From this latency to 1420 ms, a second time window for finding the peak negative component is used. The maximally negative 104-ms segment in it is taken as the value of the negative component, and the difference between these negative and positive components is taken as the peak-to-peak measure of P3 amplitude. For the baseline-to-peak measure, the positive component value is subtracted from the average of the first 104 ms of the epoch that precedes the stimulus onset.

Because the yes-target channel was the one expected, based on our standard target operations, to

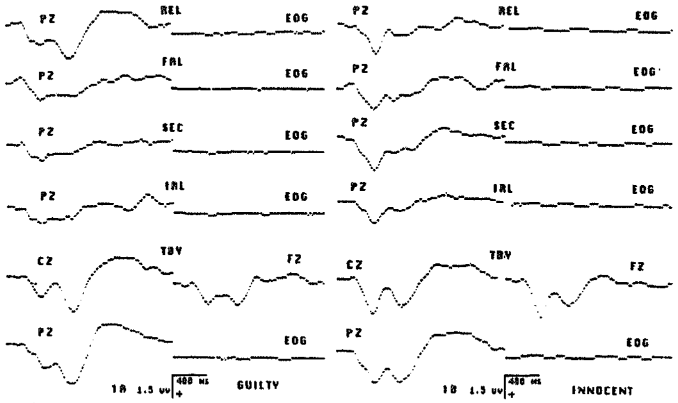


Figure 1. Grand averaged ERPs and EOG traces in guilty (1A) and innocent (1B) groups. For TBY (Yes-target), C_Z, P_Z, and F_Z responses are shown. Only P_Z is shown for others. Stimulus is presented 104 ms after start of epoch whose duration is 1920 ms. REL = relevant response, FAL = primary falsely accused response, SEC = secondary accused control response, IRL = irrelevant response. The averages were filtered to be 3dB down above 2.89 Hz.

contain a P3 response in both guilty and innocent groups, a two-way MANOVA[2] was done on the yes-target P3 responses with Site (F_Z vs. C_Z vs. P_Z) as the repeated-measures variable and Group (guilty vs. innocent) as the between-groups variable. The results for the peak-to-peak P3 index showed significance only for site ($F(2/52)=18.976$, $p<.001$) and confirmed visual impressions (Figure 1) of parietally maximum (13.1 μV) and frontally minimum (10.5 μV) waves (with C_Z at 12.8 μV); i.e., neither the group effect nor the interaction between group and site were significant ($p>.8$). For the baseline-peak P3 index, the results were similar with significance only for site ($F(2/52)=7.33$, $p<.003$), and with the mean values $F_Z=6.7$ μV, $C_Z=7.14$ μV, and $P_Z=7.56$ μV. In all results to follow, only the P_Z data will be analyzed.

Figures 2 and 3 illustrate P3 means across subjects and within groups for peak-peak amplitude and latency as defined above. Again, as predicted, the most important feature of Figure 2 for present purposes is the clearly enhanced P3 to the relevant item in guilty but not in innocent subjects. Separate repeated-measures analyses were performed on the amplitude and latency data to confirm the specific prediction given above. The between-group variable was guilt versus innocence and the repeated measures variable was relevant versus control ("REL" and "SEC" respectively, in Figures 2 and 3) response levels. The results on the peak-to-peak index indicated no significant latency effects, but for amplitude, there was a highly significant Group × Stimulus Type interaction ($F(1/27)=15.9$, $p<.001$), indicating a greater difference between relevant and control responses in guilty subjects than in innocent subjects as expected. There was also a significant main effect of stimulus type, $F(1/27)=5.727$, $p<.03$, which appears to be entirely carried by the guilty group because the difference between relevant and control responses is not only greater in the guilty group, but is slightly negative numerically in the innocent group. The group main effect (guilty vs. innocent) was not significant, $p>.3$. Also, no large systematic difference is visibly evident in Figure 2 between the amplitudes of responses to the primary ("FAL") and secondary ("SEC") control items. The baseline-to-peak P3 analysis was consistent in showing a highly significant Group × Stimulus Type interaction ($F(1/27)=11.6$, $p<.002$), but the effect of stimulus type just failed to reach significance ($F(1/27)=3.13$, $p<.08$), and the between-group effect was not significant.

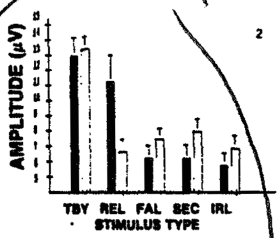


Figure 2. Computer determined P3 amplitude means and standard deviations within guilty (filled bars) and innocent (open bars) groups. TBY, REL, FAL, SEC, IRL as in Figure 1.

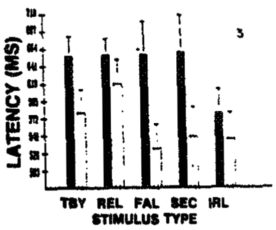


Figure 3. Computer determined P3 latency means and standard deviations within guilty (filled bars) and innocent (open bars) groups. TBY, REL, FAL, SEC, IRL as in Figure 1.

Although latency did not distinguish guilty and innocent subjects on the predicted (relevant-control) comparison, Figure 3 suggests a difference between groups on other items. Thus a 3×2 MANOVA was performed on latencies to primary accused, yes-target, and irrelevant items with item type as the repeated measures variable, and guilty vs. innocent as the between-group variable. The results showed a significant group effect ($F(1/27)=4.503$, $p<.05$), but no other significant effects.

Individual Data

Although the predicted group amplitude effects were obtained, the practical use of deception detection is more concerned with accuracy of prediction within an individual. At present, there are two documented possible approaches to diagnosing individual guilt or innocence:

1) Utilizing the ERPs within a subject, one could use the cross-correlation approach suggested by Farwell and Donchin (1988) and amplified by Wasserman and Bockenholt (1989). In their study a guilty knowledge procedure was utilized in which three kinds of ERPs were arranged: (a) The response to the guilty item, expected to contain a P3 and comparable to our present relevant item response; (b) a response to a target item, also expected to contain P3, and similar to our present yes-target item response, and (c) an irrelevant item response *not* expected to contain a P3. This latter item's response is comparable to our irrelevant items' responses, *but not necessarily to our control item responses* with which we wanted to compare re-

What is the relevant item for innocents?

[2] Multivariate MANOVAs were performed in order to reduce the likelihood of false positive errors (Vasey & Thayer, 1987). As long as multivariate results agreed with univariate results, the latter values are reported here. Also, in the SYSTAT "MGLH" module we used, if there are only two levels of a repeated measures variable (as will sometimes be the case in the present studies), only the univariate test is done because there is no concern about non-sphericity (with only two levels of the repeated measure).

sponses to our relevant item; i.e., our control items were subjected to false accusation and on this basis might be expected to contain small to moderate-sized P3 responses. Farwell and Donchin.compared the cross-correlation coefficient of the responses to guilty and target items with that of the responses to irrelevant and guilty items, utilizing the "bootstrap" approach to establish a confidence interval (Wasserman & Bockenholt, 1989) for cross-correlation differences. The reasoning was that because a P3 was expected to target and to guilty items (in guilty subjects) but not at all to irrelevant items, the guilty-target correlation should exceed the irrelevant-guilty correlation. We utilize a cross-correlational approach also (however, our bootstrapping procedure is different from theirs as described below), and we substitute our *control* item for their *irrelevant* items. Nevertheless, we anticipated at least three limitations in its utility: (a) The cross correlation of two similar (e.g., sinusoidal) waveforms differing only but distinctly in amplitude will be very high; thus a small but clear P3 in the control channel could correlate as highly with the yes-target response as would a large P3 in the relevant channel, thus making a guilty subject appear innocent on a cross-correlation comparison criterion; (b) as illustrated in the individual averages in Figure 4, some subjects can be more responsive to the relevant than to even the yes-target item. In this case the cross-correlation of the large, phasic relevant response with the yes-target is significantly *less* than that of the smaller, broader control and yes-target responses; (c) related to the previous point, to the extent that P3 *latencies* are more similar in control and target responses and less similar in relevant and target responses, the relevant-target correlations are reduced and control-target correlations increased. Simple cross-correlations on waveforms not adjusted for latency then become misleading. Because of these three issues, the addition of other decision criteria (described in the next section) in a multi-step diagnostic decision algorithm seems in order.

It is noted that the bootstrap approach to establishing a significant difference between correlation coefficients has been questioned (e.g., Rasmussen, 1987) and other approaches exist: Hotelling (1940) established a parametric *t*-test on this difference whose assumptions are difficult to satisfy with the present data; there is not much information about how robust the test is to such violations, however. Olkin (1967) developed still another confidence interval approach to the problem with less stringent assumptions. Unfortunately, the Olkin approach is such that high intercorrelations among the three appropriate waveforms (relevant, control, yes-target) may lead to terms in Olkin's formula whose square root cannot be calculated as called for. We here report results of all three methods with some interest in comparison. We utilized 1000 iterations to develop the bootstrapped distributions.

2) The alternative to the cross-correlation approach is direct comparison within an individual of control versus relevant P3 response sizes. Storage of single sweeps would allow a familiar, repeated measures *t*-test on mean P3 amplitude differences between response types. We did this in Experiment 2 (below) and observed insensitivity of this parametric *t*-test with only 22 degrees of freedom (12 control + 12 relevant sweeps − 2) on noisy data, and even after digitally filtering the single sweeps with a 3dB high cutoff of 6.11 Hz.

Alternatively, we have here, for the first time, utilized a bootstrap approach with amplitude differences (between relevant and control responses within a subject), utilizing only the averaged, within-subject ERPs. In this procedure, instead of using our regular maximum segment determinations on the actual ERPs, we repeatedly randomly sampled the ERPs between 468 and 1420 ms poststimulus and re-ordered with respect to time the randomly selected 120 data points (for 1420−468=952 ms at 8-ms resolution) so as to generate bootstrapped ERP segments for relevant and control ERPs, P_z derivations. Now our regular maximum segment determination procedure was applied, exactly as described above, to the bootstrapped ERP segments so as to determine a P3 difference estimate between bootstrapped relevant and control P3 values. This procedure was repeated 1000 times so as to obtain the mean ($\overline{X}$) and standard deviation (SD) of the bootstrapped distribution of relevant-minus-control P3 differences. A confidence interval was now set up extending from $\overline{X}$−2SD to $\overline{X}$+2SD. If it contained zero (0), then a diagnosis of innocent was appropriate because no difference was concluded to exist between relevant and control responses. If it ~~contained only~~ values <0, a diagnosis of innocent was also made because this implied a greater control than relevant response. Only if both ends of the interval were >0 would it be appropriate to conclude with 95% confidence that the P3 response to the relevant question exceeded that to the control.

Not sure I understand these statistics

This finding alone does not necessarily lead, however, to an automatic guilty diagnosis: it could be the case that the relevant response is flat in the P3 time domain but will test as greater than a control response that happens to be *negative* in the P3 region. (This situation could also lead to a false positive outcome in cross-correlation analysis.) It must therefore be additionally established that there *is* a normative P3 waveform in the relevant channel. There are various ways (e.g., template matching) one might do this, but in the present paper, we shall utilize the relevant-target correlation coefficient (R_{RT}) as a standard. It is assumed that the yes-target will contain a P3; thus if the relevant response also has a P3, even if smaller or of a somewhat different morphology, it should still correlate with the response to the target some minimal amount (provided relevant and target P3 latencies are not grossly out of phase; this is discussed below). The value of +.5 was chosen as a minimum standard on the following basis: The senior author visually inspected all relevant responses in a related, unpublished study, not knowing whether the subject sources were guilty or innocent. All responses believed, on visual inspection, to contain P3 were found to have >.5 cross correlations with the respective yes-target waves. (Data in the present study given below indicate that the minimum value could be at least as high as .52.)

The foregoing background justifies a (minimally) four-step diagnostic algorithm in determining individual guilt and innocence:

1. *A P3 response must be present in the relevant channel for a guilty diagnosis.* In this report, an $R_{RT} \geq +.5$ must be obtained, although this criterion is necessary but not sufficient for the guilty diagnosis. (Moreover, if there are gross latency/phase differences among the key waveforms, latency-adjusted data should be used. This was not necessary in the present report.) If this criterion is met, one proceeds to step 2.

2. *Parametric* t*-test.* If this conservative test finds a significant positive relevant-control difference and the R_{RT} criterion is satisfied, a diagnosis of guilty is made. No further test is necessary. If not or if single sweeps are not available, one proceeds to the next step.

3. *Bootstrap P3 amplitude difference.* Diagnostic sub-criteria were described above. If guilt is established, no further test is necessary. If guilt is not established, one proceeds to the next step.

4. *Cross-correlation tests, as described above.* Here, for a guilty diagnosis, it is *not* simply sufficient that the cross-correlation of relevant and target responses exceed that of the control-target response correlation; e.g., the former could be ≤0, indicating a lack of similarity or negative relation between relevant and target, and the latter some large negative value (−.8). Thus, just as for the relevant-control amplitude difference tests, there should be an *additional* requirement that the R_{RT} be positive and greater than some value specified *a priori*, which in this report will be +.5, as discussed above.

It should be pointed out that the bootstrap approach utilized by Farwell and Donchin (1988) bootstrapped a distribution by repeatedly taking subsamples of single sweeps from which to recalculate average waveforms and cross-correlations. Given that single sweeps were not available to us in Experiment I, we repeatedly took subsamples of the average ERPs for each condition on which to recalculate both cross-correlations and amplitude differences. The Farwell and Donchin (1988) method has the distinct advantage of directly preserving trial-to-trial fluctuations as a variance source in averaged data. The source of variability in our bootstrapped distributions is determined by randomly varying selections of data points within one time segment of the average. We assume that trial-to-trial peak *latency* variability, especially given our relatively small number of trials per average, will generate a significant component of variance in average P3 amplitude during a critical time window. By randomly selecting points in this window, our method would then indirectly reflect this source of trialwise variance. Our method is, at least, objective and, as will be noted, it appears to work.

Table 3 illustrates for guilty and innocent subjects, respectively, the outcomes of all the procedures noted above (excepting the parametric *t*-test approach because single sweeps were not available in Experiment 1). In addition, there is a final column giving the decision of the first author (blind to the group membership of each subject) as to guilt or innocence based solely upon his visual inspection of the waveforms. Examples of various outcome types in this study are given in Figures 4–9.

It is striking in Table 3 that in the guilty group, with two easily explainable exceptions, the indices were consistent; i.e., the amplitude difference results (B_A values in Table 3) were consistent with the cross-correlation results (denoted X_H, X_O, X_B in Table 3). It is also noted that although in two cases, the Olkin procedure could not be used ("< 0" outcome), in all other cases, the cross-correlation comparison procedures produced consistent results. One case (GA15) in which the B_A test yielded a "+" while two of the cross-correlation tests yielded "−" outcomes can be easily explained with reference to Figure 4B. This is a visually obvious guilty case anticipated above in which the relatively phasic, large relevant response exceeded the more rounded target and control responses, which therefore showed a higher mutual correlation (R_{CT}) than R_{RT}. P3 amplitude ratios (relevant to target, $P3_R/P3_T$) were >1 in this Case GA15, as well as in three other guilty cases (GA13, GB1, GB13) and indeed, the case of GB13 is similar to that of GA15. In the cases of GA13 and GB1, the R_{CT} values were much

Table 3

Outcomes of statistical procedures, and the decision of the experimenters as to guilt or innocence

Subject	$P3_R/P3_T$	$P3_C/P3_T$	B_A	R_{RT}	R_{CT}	X_H	X_O	X_B	Eye	Diagnosis
GB11	.78	.24	+	+.54	+.26	+	+	+	G	G
GB10	.79	.56	+	+.81	+.57	+	+	+	G	G
GB9	.96	.30	+	+.67	+.42	+	+	+	G	G
GPET	.86	.42	+	+.84	+.30	+	+	+	G	G
GA15	1.33	.84	+	+.78	+.86	[−]	<0	[−]	G	G
GB1	1.08	.27	+	+.78	−.02	+	+	+	G	G
GA13	1.89	.83	+	+.57	−.32	+	+	+	G	G
GB7	.83	.35	+	+.81	+.64	+	+	+	G	G
GB5	.39	.22	+	+.70	+.23	+	+	+	G	G
GB3	.57	.35	+	+.81	+.21	+	+	+	G	G
GB8	.61	.49	+	+.64	+.12	+	+	+	G	G
GB6	.30	.57	[−]	[+.23]	+.67	[−]	[−]	[−]	[I]	[I]
GB13	1.11	.86	+	+.73	+.69	[N]	<0	[N]	G	G
IA7	.52	.43	N	+.11	+.14	N	N	N	I	I
IA6	.41	.68	−	[+.68]	+.72	N	<0	N	I	I
IA9	.35	.41	N	[+.64]	+.81	−	−	−	I	I
IA5	.39	.90	−	+.40	+.62	−	−	−	I	I
IIV	.25	.11	[+]	+.05	+.05	N	N	N	I	I
IA2	.34	.47	N	+.39	+.21	N	N	N	I	I
IKS	.17	.26	N	−.07	−.31	[+]	N	[+]	I	I
ILA	.63	.86	−	+.40	+.40	N	N	N	I	I
IYS	.34	.61	−	[+.55]	+.88	−	−	−	I	I
IJC	.94	1.03	N	[+.59]	+.45	N	N	N	I	I
INS	.70	.68	N	+.26	+.03	N	[+]	[+]	I	I
IA1	1.10	.85	[+]	[+.67]	+.79	−	−	−	[G]	[G]
IA3	.20	.81	−	−.53	+.60	−	−	−	I	I
IA11	.81	.70	[+]	[+.68]	+.36	[+]	[+]	[+]	[G]	[G]
IB4	.70	.79	N	[+.51]	+.59	N	N	N	I	I

Note.—B_A = Bootstrapped relevant-minus-control amplitude difference result: "+" = guilty diagnosis, "−" or "N" (significant negative difference and no significant difference, respectively) = not guilty diagnosis. R_{RT} = cross correlation of P3 responses to relevant and target items. R_{CT} = cross correlation of control and target responses. R_{RT} must $\geq +.5$ for guilty diagnosis, no matter what other results are found on cross-correlation tests. X_H, X_O, and X_B are results of Hotelling, Olkin, and Bootstrap (respectively) tests (see text) on significance of differences between R_{RT} and R_{CT}: "+" = guilty diagnosis, "N" or "−" (no significant difference or significant negative difference, respectively) indicate not guilty diagnosis. The result "<0" can occur only with the Olkin test and means that the test could not be performed (see text). $P3_R/P3_T$ = the ratio of the P3 amplitudes of the relevant and target responses. $P3_C/P3_T$ is the ratio of control and target responses. EYE = diagnosis based on visual inspection of waveforms; "G" = guilty diagnosis; "I" = Innocent diagnosis. DIAG = final diagnostic conclusion based on 4-step algorithm described in text. "Guilty" and "Innocent" as for "EYE." Data not consistent with assigned group (guilty or innocent) are bracketed; e.g., if $R_{RT} < +.5$ in the guilty group or $\geq +.5$ in the innocent group, or if X_O, X_H, X_T, or B_A is "N" or "−" in the guilty group, or "+" in the innocent group, etc. Bracketed letters in the DIAG column indicate erroneous diagnoses. The top set of 13 subjects with "G" prefixes is the guilty group. The other subjects ("I" prefix) comprise the innocent group.

less than the corresponding R_{RT} values, so that B_A and X_H, X_O, and X_T values were consistent. Subject GB6 clearly "beat the test" on all measures with perfect consistency. Thus the algorithm correctly diagnosed 12/13 = 92% of the guilty subjects. Figure 5 shows a completely consistent and representative guilty subject (GB7).

Results for the innocent group were not as consistent as those for the guilty group; however, following the four-step algorithm allowed unambiguous diagnoses in all 15 subjects. The R_{RT} values varied from −.53 to +.68 (compare guilty data). Not only were there innocent cases for whom the X_H, X_O, and X_B outcomes did not agree (IKS, INS), but in four cases the B_A and X_B, B_H, X_O values were inconsistent: IIV, IKS, INS, and IA1[3] (see Table 4). It is easy to explain Case IA1 (Figure 4A); it is very similar to GA15 (Figure 4B), as indicated by the $P3_R/P3_T > 1$. Our algorithm made an erroneous guilty diagnosis in this innocent subject. In the case of IIV (Figure 6), the B_A outcome was positive probably because the relevant response was *less negative* (i.e., vs. *more positive*) in the P3 time domain than was the control. Thus, the R_{RT} value was near 0 (+.05) and our algorithm requires an innocent diagnosis which is consistent with the visual impression of minimal positivity in the relevant response. (Compare with cases GA15 in Figure 4B, GB13, or IA1 in Figure 4A, which also have "+" outcomes on B_A but "−" or "N" outcomes for cross-correlation. In these cases, however, the guilty diagnosis

[3] We consider the "−" or "N" outcomes both to indicate innocence. Thus "inconsistent" results include a "+" (i.e., guilty) on some measures but "−" and/or "N" (both = innocent) on others.

Table 4

Outcome analysis by group

Category	N/N	N/+	+/N	N/−	−/N	−/−	+/−	−/+	+/+
Number of Guilty Subjects	0	0	1	0	0	[1]	1	0	10
Number of Innocent Subjects	4	2	1	1	2	3	[1]	0	[1]

Note.—In the category row, the symbol to the left of the slash (N, +, or −) refers to the B_A outcome; the symbol to the right refers to the outcome in at least two of the three cross-correlation tests. The table was derived from Table 3. The bracketed numbers identify the 1 guilty and 2 innocent subjects who were misdiagnosed.

Figures 4–9 (right). Figure 4. Two individual sets (4A and 4B) from two individual cases of three averaged P_Z traces of relevant (top), control (middle), and yes-target (bottom) responses filtered to be 3dB down above 4.23 Hz. For these traces and for Figures 5–9 and 11–15, there are from 10–14 sweeps averaged per trace (see Table 1). Stimulus is presented 104 ms after start of epoch whose duration is 1920 ms. The two cases illustrate a situation in which the relevant response is greater than both control and yes-target responses, leading to greater R_{CT} than R_{RT} values. A guilty subject (GA15) is in Figure 4B; the (misdiagnosed) innocent subject IA1 is in Figure 4A.

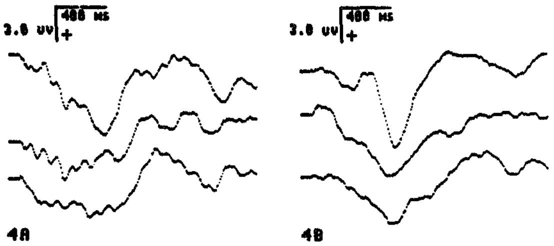
Figure 5. Three traces as in Figure 4A or 4B, but for a more typical guilty subject (GB7) with the yes-target response at least as large as the relevant response, and with both > control response.

Figure 6. Three traces as in Figure 5 but for innocent subject IIV, discussed in text.

Figure 7. Three traces as in Figure 5 but for innocent subject IKS, discussed in text.

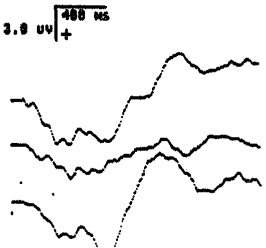
Figure 8. Three traces as in Figure 5 for erroneously diagnosed subject IA11.

Figure 9. Three traces as in Figure 5 for a representative correctly diagnosed innocent subject.

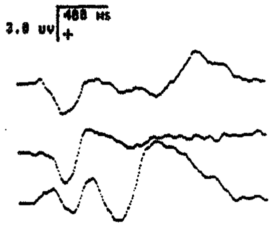


4A

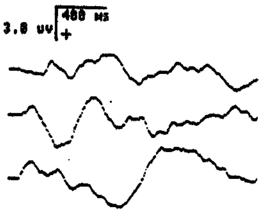
4B



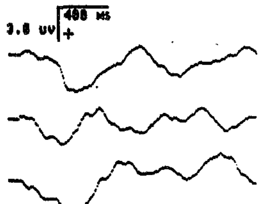
5



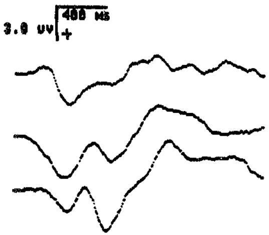
6



7



8



9

was warranted by the R_{RT} values of +.78, +.73, and +.67, respectively.) In the very similar cases of IKS (Figure 7) and INS, as anticipated above, the R_{RT} correlations were indeed larger than the R_{CT} correlations, but only in the sense of being less negative (IKS −.07 vs. −.31) or weakly more positive (INS, +.26 vs. +.03) than the corresponding R_{CT} values. The algorithm requires the innocent diagnosis; moreover, the three cross-correlation tests were not consistent in these cases, which suggests that the correlation differences were marginal anyway. The remaining innocent subjects had in all except one case (IA11; Figure 8) consistent B_A and X_H, X_O, and X_B outcomes consistent with an innocent diagnosis (see representative Figure 9). IA11 was another erroneous (but consistent) guilty diagnosis. The four-step algorithm, then, correctly diagnosed 13 of 15 innocent subjects = 86.6%.

Discussion

1. Subject to the limitations discussed below (under 2.) the present P3-based procedure provided a relatively accurate analog of a control question screening test. The overall hit rate was (12+13)/(13+15)=89.3%. The two errors made in the innocent group could be explained by the chance occurrence of large P3-like deflections in the relevant channel, but there is another possibility suggested (especially in subject IA1, Figure 4A) by the rather distinct-looking P3 waves in these subjects' relevant responses: It is possible that the relevant items presented to these innocent subjects were acts of which the subject was indeed personally innocent, but which had special meaning for the subjects anyway; e.g., perhaps the subject's roommate or sibling (etc.) was involved in the act, and upon seeing this relevant item, the subject associated uniquely to it, leading to a P3. We have no direct evidence for this possibility here; post-test interrogation of such innocent subjects could be done in future work. We did ask our one guilty subject (GB6) who "beat the test" how he did it. He admitted no strategy and inasmuch as his yes-target count accuracy was 100%, we have no present explanation.

We noted that in most guilty subjects, the various numerical indices were mutually consistent, as well as being consistent with visual inspection. The exceptions, GA15 and GB13, were easily identified as guilty utilizing our four-step algorithm that gives a greater weight to the bootstrapped amplitude difference test (B_A) than to the cross-correlation comparison tests (R_{RT} vs. R_{CT}) when the relevant-to-target P3 amplitude ratio ($P3_R/P3_T$) is >1 (as explained above). It might be additionally noted that in the case of GB13, the control and target P3 waves were more in phase than the relevant and target waves as indicated by the divergent latencies of relevant and control response (736 and 776 ms, respectively) and identical latencies of control and target P3s (776 ms). Such effects suggest use of latency-adjustment procedures prior to R_{RT} and R_{CT} computation in future work.

2. The present control question screening analog is limited by the shortcomings of all lab analogs (Ekman, 1985) and has some further problems specifically inherent in the present methods: In particular, our preliminary list filling and second-rehearsal procedures could clearly not take place in a field setting, and may have added a confounding element. In one way, our list-filling procedure is analogous to the mock crime scenarios used in other lab studies of deception detection in the sense that it provides validation of diagnosis which, though it cannot be as certain as a mock crime, is probably more natural: In a mock crime situation the subject usually knows that at least one experimenter—the one who directs the crime act—knows who is and is not guilty. The subject may thus feel defeated prior to his/her test. In our situation, as noted above, subjects are probably not aware of our surveillance of their list filling and probably go into the test believing that they alone know the truth. Moreover, the crimes involved are not externally directed, mock crimes: they are the real antisocial acts of the subjects.

Our second-rehearsal procedure surely heightens, intentionally, if perhaps artificially, the guilty subject's awareness of his guilty act, and does so on the same day as and prior to his test. It could be replied, however, that in a real field situation, it should be unnecessary to heighten awareness because in a real test setting, a guilty person with full knowledge that one particular guilty disclosure could cost him his job or freedom would probably be highly focused on that potential disclosure, without the need for a second rehearsal-like procedure. However, a carefully orchestrated preliminary interrogation, with no operator knowledge of guilty acts, could be developed to stimulate such focus anyway. What is required is that any such preliminary activation be *non-selectively distributed across control and relevant items.* Experiment 2 was designed, in part, to deal with these issues.

At least two purposes could have been achieved by the second-rehearsal procedure in Experiment 1: 1) The subject is forced to clearly fix in his mind what his guilty act is. Thus, he reinforces the initial mental commitment forced by the list-fill procedure. 2) The subject's attention is forced to one particular item, the guilty item, prior to and on the same day as his test. This provides a potential confound of interpretation of the P3 in the relevant response: is the response due to the rehearsal being temporally close to the real test, or to guilt, or to a facilitative interaction of the two factors? If this confound interpretation is correct, it would predict that if a subject were given a list of neutral, arbitrary stimuli (e.g., numbers) and told to select and remember one in particular as *his* number, and was then later tested in an oddball paradigm for P3 response to members of the list—which included not only the selected/to-be-remembered item, but also a designated response target—P3 responses would be obtained to both selected-remembered and target stimuli. Nasman and Rosenfeld (1990) found, however, that in such a situation, only the target stimulus evokes the P3, and that personally selected but neutral, to-be-remembered stimuli are easily overshadowed by other stimuli experimentally endowed with more P3-evoking potency. Thus it is here hypothesized that in the present study, it was not simply the list-filling or second-rehearsal procedure that isolated an item by forcing a unique response to it: the item must have had inherent special significance such as that of a truly guilty item.

It is noted that the second-rehearsal procedure was introduced in Experiment 1 here because in an earlier unpublished pilot experiment that did not use it, hit rates were at 90% for 20 guilty subjects initially guilty of three or less of the nine listed acts (as revealed by TV surveillance) but 75% for 8 guilty subjects who had checked "yes" to four or five acts.

Without having yet attempted the systematic study that this suggestive pilot finding merits, we employed the second-rehearsal procedure here as an especially potent method of narrowing a guilty subject's focus on one rather than on multiple guilty acts.

EXPERIMENT 2

Our hypothesis about the results of Experiment 1 is that it was the subject's knowledge of his guilt, stimulated by the temporally proximal list-fill and second-rehearsal procedures, that produced the P3s in the relevant responses. Further, although in our first study we wanted to be maximally certain of activating guilty self-knowledge and thus knowingly utilized the potentially confounding second rehearsal procedure, we hypothesize for Experiment 2 that: 1) other non-confounding and non-selective methods of activation temporally proximal to the main test can be effective, and 2) the initial mental commitment inherent in list-filling and second-rehearsal methods will have little or no value (i.e., P3-evoking potency) without being temporally proximal to guilty self-knowledge activation.

In Experiment 2, two groups of guilty (and no innocent) subjects were run. List-filling, second-rehearsal, and interrogation procedures were utilized in both groups as in Experiment 1; however, the main run with recording of ERPs was *not* done on the same day; it was delayed by 7–12 days. What differentiated the two groups was that in the subgroup *delay-only*, the main control question run was given on this second day, whereas in subgroup *delay-rehearsal*, an extra, interrogation-like procedure was utilized on the second day and just prior to the main run, the aim of which was to re-activate the subject's guilty self-knowledge, but in a way that did *not* treat the relevant item in any way different from the treatment of the three other falsely accused (control) items. Our hypothesis is that the initial mental commitment and unique response focusing (inherent in the list-fill and second-rehearsal procedures) would *not* be adequate to activate P3 production seven or more days later, and therefore predicts a low hit rate in the delay-only subgroup. Our hypothesis that a non-selective activation process will be effective only with a personally relevant (i.e., guilty) item predicts good results in the delay-rehearsal subgroup.

Finally, it is reasonable to note that our tightly controlled situation (with only one guilty act in a set) could be difficult to arrange in the field. This is considered in the General Discussion, below.

Method

Except for the one week or more separation between the final main run and the key procedures preceding it (list-filling, second-rehearsal procedure), and except as noted below, the methods of Experiment 2 were the same as those used on the guilty subjects of Experiment 1. The only other difference from Experiment 1 for both delay-only and delay-rehearsal subgroups was the delay of the accusation procedure until the second week. This was done so as to: 1) avoid complete loss of continuity in the delay-only subjects who would otherwise be run on Day 2 directly following electrode application, and 2) help demonstrate that passive involvement (being accused of four possible acts) is not the key element in effective activation of guilty self-knowledge, as elaborated in the Experiment 2 discussion. There was one other key difference between delay-rehearsal and delay-only groups: Whereas for delay-only subjects, following electrode application, accusation procedure, and general reminder instructions (about trying not to blink or move about excessively), the subjects were run in the main test as in Experiment 1, for delay-rehearsal subjects, interposed between the accusation procedure and main run was a two-phase, non-selective activation procedure. In the first phase, each subject had read aloud to him, one at a time, each one of the eight acts he would see on the final main run. This included one guilty act and seven innocent acts including the three falsely accused acts from the accusation/interrogation. (Order of presentation was systematically varied across subjects.) After hearing each act, the subject was told to create and write out a brief story involving the act, including planning, doing, and reflecting on the act. Following this exercise, the delay-rehearsal subject was read aloud each of only the three falsely accused (control) and one relevant (guilty) acts. These acts were read in the first person and in two sentences, affirmative and negative, e.g., "I have cheated on a test" and "I have not cheated on a test." After hearing each sentence, the subject was told to write down each sentence. This was the second phase of the activation procedure.

It was intended to have 12 subjects per group; however, because two subjects were lost due to >50% artifact rates, and six more could not or would not schedule the second visit, or were guilty of more than five acts, the final sample sizes were eight in both delay-rehearsal and delay-only groups.

Results

Grand averages (P_Z only) for relevant, control, and target stimuli in the delay-only and delay-rehearsal groups are shown in Figure 10 (A and B).

There are apparently distinct P3 waves in the yes-target channels of both delay-only and delay-rehearsal groups but in the relevant channel of only the delay-rehearsal group average. It happens that, as shown below, three of the eight delay-only group members did have P3 responses (and were diagnosed by algorithm as guilty). These were apparently not enough in phase (520 ms, 592 ms, and 936 ms) to have much effect on the averages. It also appears that there may be a small P3 response in

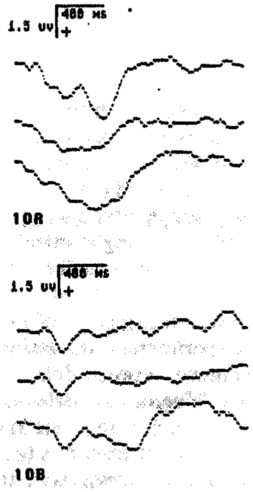


Figure 10. Two sets of grand averaged ERPs, otherwise each set of 3 is as in Figure 5. Figure 10A is for the delay-rehearsal group, Figure 10B is for the delay-only group of Experiment 2.

the control channel of the delay-rehearsal subjects. Table 5 gives the amplitudes and latencies for relevant and control responses in the two groups, and reflects visual impressions.

Because the delay-only and delay-rehearsal groups were both guilty groups, we planned, *a priori*, not to combine them in a MANOVA, utilizing the interaction term (of guilt × relevant/control) to assess group differences, as was appropriate in Experiment 1. Instead we examined the differences between relevant and control P3 amplitudes and latencies separately within each group in correlated *t*-tests. (The control response selected for comparison was C, as in Experiment 1.) For peak-to-peak P3 amplitude in the delay-rehearsal group, $t(7)=4.01$, $p<.01$. For the delay-only group, amplitude effects failed to reach significance. Baseline-to-peak P3 results were consistent; $t(7)=3.66$, $p<.01$ in the delay-rehearsal group ($p>.05$ in the other group). There were no significant latency effects for either delay-only or delay-rehearsal groups.

Results within subjects are given in Table 6 and Figures 11–15. (These figures, as with Figures 4–9, show filtered waveforms; the calculations were done on unfiltered data.) The algorithm correctly diagnosed 7/8 (87.5%) of the delay-rehearsal subjects but only 3/8 (37.5%) of the delay-only subjects. A typical correctly diagnosed guilty subject in the delay-only group, Case S3, is shown in Figure 11. An incorrectly diagnosed counterpart (S7) is shown in Figure 12.

As expected, the correlated *t*-test (A_T) index was extremely insensitive, for both negative and positive differences, relative to B_A and visual inspec-

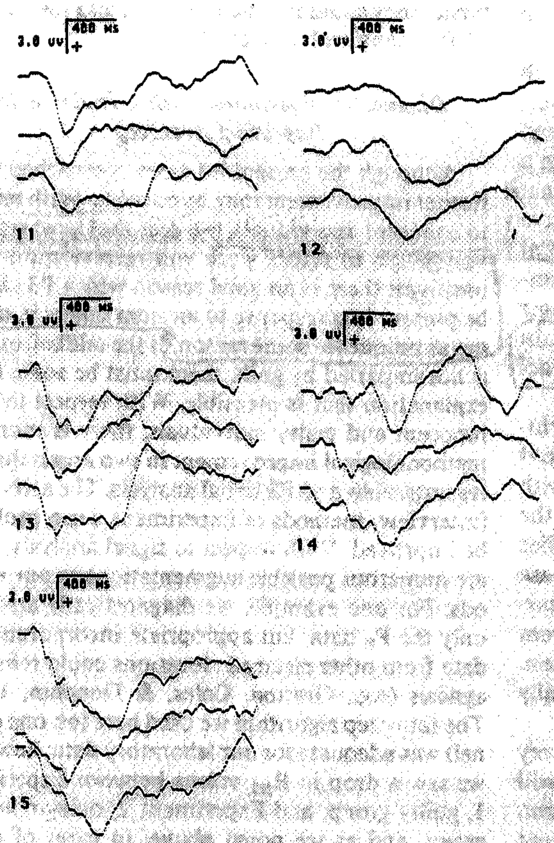
Table 5
Mean relevant and control response amplitudes and latencies at P_Z for delay-rehearsal and delay-only groups

	Mean P_Z Amplitude (μV)		Mean P_Z Latency (ms)	
Groups	Relevant	Control	Relevant	Control
Delay-Rehearsal	10.03 ± 3.5	6.04 ± .26	607 ± 56.5	600 ± 83.9
Delay-only	5.59 ± 2.35	4.7 ± 2.96	658 ± 186.1	617 ± 213.8

Table 6
Individual results for Experiment 2

Subject	$P3_R/P3_T$	$P3_C/P3_T$	B_A	A_T	R_{RT}	R_{CT}	X_H	X_O	X_B	Eye	Diagnosis
S1	.63	.86	−	−	−.03	+.08	−	N	−	[i]	[i]
S3	.67	.11	+	+	+.50	−.13	+	+	+	g	[g]
S7	.18	.42	−	−	+.31	+.34	N	N	N	[i]	[i]
S8	.49	.10	+	+	+.51	−.59	+	+	+	g	g
S9	.59	.15	+	N	+.50	+.05	+	+	+	g	g
S10	.9	1.15	−	N	+.73	+.71	N	N	N	[i]	[i]
S11	.6	.39	N	N	+.35	+.46	N	N	N	[i]	[i]
S12	.50	.47	N	N	+.42	+.24	+	+	+	[i]	[i]
D1	.73	.56	+	N	+.77	+.76	N	<0	N	g	g
D2	.68	.68	N	N	+.12	+.26	N	N	N	[i]	[i]
D3	.39	.14	+	N	+.53	+.46	N	N	N	g	g
D4	1.77	.73	+	+	+.51	+.29	N	N	+	g	g
D5	.77	.36	+	N	+.52	+.14	+	+	+	g	g
D6	1.16	.64	+	+	+.72	+.78	N	<0	N	g	g
D7	1.5	.95	+	N	+.65	+.69	N	<0	N	g	g
D8	.96	.70	+	N	+.75	+.33	+	+	+	g	g

Note.—Legend of Table 3 applies with the extra entry A_T = results of parametric, correlated *t*-test on single sweeps, Relevant versus Control: "+", "−", and "N" have same meaning as with B_A.



Figures 11–15. Figure 11. Three traces as in Figure 5, but for a correctly diagnosed delay-rehearsal subject (S3).

Figure 12. Three traces as in Figure 5, but for an incorrectly diagnosed delay-rehearsal subject (S7).

Figure 13. Three traces as in Figure 5, but for an exceptional case (S12) discussed in text.

Figure 14. Three traces as in Figure 5, but for a correctly diagnosed delay-rehearsal subject (D6) with $R_{CT} > R_{RT}$.

Figure 15. Three traces as in Figure 5, but for a correctly diagnosed delay-rehearsal subject (D7) with $R_{CT} < R_{RT}$.

tion. The other indices were consistent in the delay-only group except in Case S12 (Figure 13). The reasons for the significant cross-correlation differences here are probably spurious. The failure of R_{RT} to reach .5 requires the innocent diagnosis.

In the delay-rehearsal group, there were five of seven cases for whom the guilty diagnosis resulted from the algorithm even though the cross-correlation outcomes were negative (all measures agreed in the other two correctly diagnosed delay-rehearsal subjects). Cases D6 and D7 (Figures 14 and 15, respectively) illustrate why. In both cases, visual inspection suggests even larger responses in the relevant than in both control and yes-target channels (confirmed by $P3_R/P3_T > 1.0$). The R_{CT} correlations thus are misleadingly large (as was seen in cases GA15, GB13, and IA1 in Experiment 1). In addition, the latencies are not well aligned between relevant and target responses (especially obvious in Figure 15 and unshown cases D1 and D4), and such misalignment tends to depress R_{RT} misleadingly.

GENERAL DISCUSSION

P3-Based Deception Detection Paradigm

Experiment 1 demonstrated that if a subject can be confronted with one known guilty act in a set of acts in which each is perceived to have finite probability in his subject population, the oddball effect (Duncan-Johnson & Donchin, 1977) will force a P3 response that is usually detectable on an individual basis. False accusation on control[4] (innocent) questions does not have a significant effect in innocent or in guilty persons. The method of Experiment 1 did, however, use a method of activating guilty self-knowledge that could not be used in the field: a rehearsal procedure that required a guilty subject to privately acknowledge (to himself) a guilty act, prior to and on the same day as the final control question test analog. (The reasons for use of this procedure were discussed above.) We hypothesized that the purpose served by the rehearsal/self-commitment procedure was to powerfully activate focused self-knowledge, and that the initial mental commitment *per se* involved in the self-acknowledgement would not be sufficient to focus activated self-knowledge if not applied in close temporal proximity to the control question test. We also hypothesized that other, more natural and unconfounded procedures to activate focused self-knowledge would be effective if applied in close temporal proximity to the test. Experiment 2 confirmed these predictions; requiring subjects to construct stories and write sentences about various acts, innocent and guilty, was adequate to activate guilty self-knowledge. Moreover, these *non-selective* (unconfounded) procedures were the first ones contemplated and tried. Further research could lead to even more potent methods of non-selective activation.

EX 1
MAIN
FINDING

Ex 2
MAIN
FINDING

Regarding real field use, there remains the important question of how important and possible it

[4] In standard polygraphic practice, a control question is typically defined (Reid & Inbau, 1977) as one pertaining to a general area in which the subject could at some time have been involved, but which the investigator is actually not primarily concerned about. Our control items are similar in the sense of our asking and probing about them while being actually uninterested in them, but these items do involve known innocent acts.

would be to arrange a stimulus or question set like ours, involving one guilty/relevant act in a set of seven innocent acts. Our method of arranging such a situation (i.e., the use of hidden surveillance, which also provided us with an estimate of ground truth for validation purposes) is possible only in a laboratory. There are really two issues here: 1) What is the largest number of guilty acts one could have (in a set of seven acts) with each still capable of evoking the P3/oddball response? (i.e., as the number approaches 50%, the term "oddball" no longer applies); and 2) how can an investigator maintain the ratio of guilty to innocent acts within the acceptable range?

Future Study?

Regarding (1), there have been numerous demonstrations that P3 amplitude is inversely correlated to oddball or oddball/target probability (with the added task-relevance requirement shifting the function considerably), but most of these studies have used the auditory modality and in any case do not directly apply to present concerns if the target requirement was present, because a guilty item cannot be an explicit target in a real field situation. However, the specific question (1) is empirically answerable.

Regarding (2), there probably is no infallible way to be certain that a given set of plausible acts will contain all acts of which a given subject is *innocent*. The method we used was to utilize as innocent items those acts of controlled (low) probability in the subject population, but which were plausibly probable to the subject. Table 1 (Experiment 1) revealed that no subjects (in 28) were guilty of acts we had predetermined to be improbable (<2%). This was also true for the 24 subjects initially scheduled in Experiment 2. Yet as noted above, an earlier study on the same subject population showed that the actually improbable acts we selected are *perceived* to have the same probabilities as the actually probable acts. There are, of course, more serious acts perceived (probably accurately) by our student-subject population to have a low probability in the population, e.g., selling heroin; producing pornography, etc. We do not use such items on our tests. It seems plausible that lists of acts with differential perceived and actual probabilities could be developed for any population of interest. Finally, although there is surely a finite probability that an occasional subject's control (or irrelevant) items will include one of which he/she is guilty, such an occurrence can not lead to a false positive (misdiagnosed, actually innocent person), but only to a false negative diagnosis. It follows from the preceding considerations that although our present methods are in no way to be regarded as highly tuned, they could in time be refined and then adapted for many real applications.

Diagnostic Algorithms Within Individuals: Improving Accuracy

Although the present hit rates approached 90%, further improvement may be possible. With respect to innocent individuals, we feel that post-test investigation in future work will resolve most false positives: there is no good reason why a P3 should be present in a response to an item unless it stands out as unique for some reason. If the oddball quality is not imparted by guilt, there must be some other explanation that is plausible. With respect to both innocent and guilty individuals, there is room for methodological improvement in two areas; the pretest interview and P3 signal analysis. The activation (interview) methods of Experiment 2 can probably be improved. With respect to signal analysis, there are numerous possible augmentations to our methods: For one example, we diagnostically analyzed only the P_Z data, but appropriate incorporation of data from other electrode locations could refine diagnosis (e.g., Gratton, Coles, & Donchin, 1989). The four-step algorithm we used here (on one channel) was adequate for our laboratory data; however, we saw a drop in R_{RT} values between Experiment 1, guilty group, and Experiment 2, delay-rehearsal group, and as we noted above, in cases of phase shifting, simple cross-correlation data could produce misleading results. Moreover, our use of latency-unadjusted R_{RT} values to determine presence or absence of a P3 waveform in the relevant response is not an optimal method.

COULD THIS BE DON IN THE FIELD THO

The questions to be answered by an appropriate algorithm are really twofold and simple: 1) Is there a P3 present in the relevant response? 2) If so, is it larger in amplitude than the one possibly also seen in the control response? Because there are various reasons why latencies in the various to-be-compared responses could fail to align, it is reasonable that all data be first adjusted with respect to latency via the use of a standard template. After that, a minimal criterion for latency-adjusted correlation of relevant and any reasonable P3 standard waveform is applied. Then, the only remaining question pertains to amplitude ratio, relevant-to-control. Bootstrapped amplitude difference approaches can be utilized here.

"Oddball" Basis of Present Paradigm

It has been noted here that in procedures like ours, the relevant item is an oddball for guilty but not innocent subjects. The term "oddball" is not meant to imply only the attribute of low probability. It is apparent that relevant items are also especially meaningful and therefore attention-absorbing for guilty subjects. An attentional response to a stimulus, even if it is involuntary (e.g., in a guilty subject attempting to avoid detection), is what is ordinarily required in order to perform a task-relevant response in a more familiar P3/oddball task. Thus it is explicitly recognized here that attention, task relevance, meaningfulness, and other attributes in addition to low subjective probability, may explain the P3-evoking property of relevant items for guilty subjects.

REFERENCES

Darus, D., French, K., & Wallace, L. (1986). C-128 ultra hi-res graphics, part 2. *Run, 3*(5), 34–39.

Duncan-Johnson, C.C., & Donchin, E. (1977). On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology, 14*, 456–467.

Duncan-Johnson, C.C., & Donchin, E. (1979). The time constant in P300 recording. *Psychophysiology, 16*, 53–55.

Ekman, P. (1985). *Telling lies*. New York: Norton & Co.

Fabiani, M., Karis, D., & Donchin, E. (1986). P300 and recall in an incidental memory paradigm. *Psychophysiology, 23*, 298–308.

Farwell, L.A., & Donchin, E. (1986) The "brain detector": P300 in the detection of deception [Abstract]. *Psychophysiology, 24*, 434.

Farwell, L.A., & Donchin, E. (1988). Event-related potentials in interrogative polygraphy: Analysis using bootstrapping [Abstract]. *Psychophysiology, 25*, 445.

Forth, A.E., Hart, S.D., Hare, R.D., & Harpur, T.J. (1988). Event-related potentials and detection of deception [Abstract]. *Psychophysiology, 25*, 446.

Furedy, J.J. (1986). Lie detection as psychophysiological differentiation. In M.G.H. Coles, E. Donchin, & S.W. Porges (Eds.), *Psychophysiology: Systems, processes, and applications* (pp. 683–701). New York: Guilford Press.

Gratton, G., Coles, M.G.H., & Donchin, E. (1989). A procedure for using multi-electrode information in the analysis of components of the event-related potential: Vector filter. *Psychophysiology, 26*, 222–232.

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics, 11*, 271–283.

Kleinmuntz, B., & Szucko, J.J. (1984). Lie detection in ancient and modern times. *American Psychologist, 39*, 776–786.

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

Nasman, V.T., & Rosenfeld, J.P. (1990). Parietal P3 response as an indicator of stimulus categorization: Increased P3 amplitude to categorically deviant target and non-target stimuli. *Psychophysiology, 27*, 338–350.

Olkin, I. (1967). Correlations revisited. In J. Stanley (Ed.), *Improving experimental design and statistical analysis* (pp. 102–129). Chicago: Rand-McNally.

Raskin, D.C. (1986). The polygraph in 1986. *Utah Law Review*, No. 1, 29–73.

Rasmussen, J. (1987). Estimating correlation coefficients: Bootstrap and parametric approaches. *Psychological Bulletin, 101*, 136–139.

Reid, J.E., & Inbau, F.E. (1977). *Truth and deception, the polygraph technique* (2nd ed.). Baltimore: Williams and Wilkins Co.

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

Rosenfeld, J.P., Nasman, V.T., Whalen, R., Cantwell, B., & Mazzeri, L. (1987). Late vertex positivity in event-related potentials as a guilty knowledge indicator: A new method of lie detection. *International Journal of Neuroscience, 34*, 125–129.

Saxe, L., Dougherty, D., & Cross, T. (1985). The validity of polygraph testing. *American Psychologist, 40*, 355–366.

Vasey, M.W., & Thayer, J.F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology, 24*, 479–486.

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

Wilkinson, L. (1986). *SYSTAT: The system for statistics*. Evanston, IL: SYSTAT, Inc.

(Manuscript received October 3, 1989; accepted for publication July 17, 1990)