Psychophysiology, *34* (1997), 234–240. Cambridge University Press. Printed in the USA.

METHODOLOGY

A comparison of methods for the analysis of event-related potentials in deception detection

JOHN J. B. ALLEN[a] AND WILLIAM G. IACONO[b]

[a]University of Arizona, Tucson, USA
[b]University of Minnesota, Minneapolis, USA

Abstract

We previously reported that a Bayesian-based event-related potential memory assessment procedure (Allen, Iacono, & Danielson, 1992, *Psychophysiology*, *29*, 504–522) was highly accurate at identifying previously learned material, regardless of an individual's motivational incentive to conceal information. When a bootstrapping procedure (Farwell & Donchin, 1991, *Psychophysiology*, *28*, 531–547) is applied to these same data, greater motivational incentives appear to increase the accuracy of the procedure. Receiver operating characteristic (ROC) curves were used to examine these two procedures and a new procedure. ROC curves indicated that all three methods produce extremely high rates of classification accuracy and that the sensitivity of the bootstrapping procedure to motivational incentive is due to the particular cut points selected. One or the other method may be preferred depending upon incentive to deceive, the cost of incorrect decisions, and the availability of extra psychophysiological data.

Descriptors: Deception detection, Memory, ERPs, P3, ROC

Several methods exist for using event-related brain potentials (ERPs) to detect concealed information with high levels of accuracy (Allen, Iacono, & Danielson, 1992; Boaz, Perry, Raney, Fischler, & Shuman, 1991; Farwell & Donchin, 1991; Rosenfeld, Angell, Johnson, & Qian, 1991; Rosenfeld et al., 1988). The aim of the present study was to compare the accuracy of two such procedures in the same individuals by applying the bootstrapping procedure (Wasserman & Bockenholt, 1989) of Farwell and Donchin (1991) to the data of Allen et al. (1992). Additionally, the available data allow for an examination of the influence of motivational incentives on the accuracy of detecting concealed information.

Other than the approach of Boaz et al. (1991), who used an ERP procedure based on the N400 component, the existing ERP procedures are essentially variants of the guilty knowledge technique (Lykken, 1959) in which previously seen but concealed items (*concealed*) appear infrequently among new information (*nontargets*). These procedures make it likely that the previously seen but concealed information, if recognized, will elicit a larger P3 amplitude of the ERP than will the frequent nontargets. In addition, to ensure that each stimulus is processed rather than simply ignored, relevant targets (*targets*) are included in these procedures (see Table 1). These infrequently occurring task-relevant targets are not related to the previously seen but concealed information; participants must identify these targets to ensure that they are attempting to make discriminations between stimuli. Responses to these targets additionally provide a reference for how an individual responds electrophysiologically to previously seen information.

Table 1. *Categories of Stimuli in P3-Based ERP Deception-Detection Procedures*

Category	Frequency of occurrence	Description
Concealed	rare	memory-relevant items that were previously learned, but participant attempts to conceal this knowledge
Targets	rare	task-relevant items that were recently learned, relevant to current task to ensure attention to all items
Nontargets	frequent	new items that are neither memory-relevant nor task-relevant

Note: If concealed items are recognized by participants, then these items should elicit a large P3 response similar to that of targets; if concealed items are not recognized, they should be treated just like nontargets and therefore should elicit a small P3 response like that of the nontargets.

Portions of the present data set were presented at the 1991 annual convention of the Society for Psychophysiological Research.

This research was supported in part by NIMH research training fellowship 5T32-MH17069-07 and by a grant from the McDonnell-Pew Program in Cognitive Neuroscience.

We thank Sheri Boril and Kurt Danielson for assistance with participants, Peter Rosenfeld and Larry Farwell for comments on an earlier draft of this manuscript, several anonymous reviewers for their helpful comments, and Mark Borgstrom for statistical assistance and consultation.

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

The essential obstacle that all deception-detection procedures face is providing a statistically supported decision for a single individual. Each of the ERP deception-detection methods uses the participant as his or her own control, attempting to minimize the influence of individual differences irrelevant to the detection of memories, by focusing on relevant within-person differences in the ERP signatures to concealed versus nonconcealed material. A brief overview of two such methods for ERP deception-detection procedures follows.

Bayesian analysis (for a detailed explanation, see Allen et al., 1992) computes the probability that items from each of several sets of items are familiar to an individual. As indicated in Table 1, two of the sets of items are those that the participants have learned, and five of the sets of items have not been learned. The procedure uses several ERP indicators (e.g., P3 amplitude), each of which imperfectly distinguishes previously seen from new material. In practice, any given indicator may distinguish such items very well but seldom perfectly. These particular indicators included P3 amplitudes and four measures derived from factor analyses of ERPs for each participant. These indicators were based upon ERPs that included only trials for which participants did not respond incorrectly.

By ascertaining the proportion of ERPs with the presence or absence of these indicators (e.g., Indicator 1 suggests items were learned, Indicator 2 does not) that were evoked by previously seen items, the procedure calculates the probability for each participant that any given ERP is in response to previously seen items. For example, if 95% of ERPs with a given combination of indicators were evoked by previously seen items (and 5% of ERPs with such a combination were evoked by new items), then there exists a 95% probability that such an ERP is indicative of prior knowledge of the items that evoked it. Across three groups of 20 participants each (Allen et al., 1992), a Bayesian combination of five ERP indicators identified learned material as familiar (regardless of a person's behavioral response) with 94% accuracy and classified unlearned material as unfamiliar with 96% accuracy. Moreover, the rates of accuracy of the procedure did not differ as a function of motivational incentive across three groups of participants (learned material correctly identified 93%, 92.5%, and 95% of the time; unlearned material correctly identified 96%, 94%, and 98% of the time).

The rationale for Farwell and Donchin's (1991) bootstrapping procedure follows from the expectation that both sets of familiar items (concealed and targets), if recognized, should elicit a large P3 response because these items appear infrequently (see Table 1) against a background of new material (nontargets). If, by contrast, the concealed items are not recognized as distinct from the background of new nontarget items, then the concealed items should produce a P3 response that is small and similar to the P3 in response to nontarget items. Thus, if a person recognized the covered items as familiar, then the cross-correlation between the ERP to concealed items and that to the target items (concealed–target correlation) should be higher than the cross-correlation between this ERP to concealed items and the ERP to the nontargets (concealed–nontarget correlation). If by contrast, the concealed items were not recognized as familiar, the concealed–nontarget correlation ought to be greater than the concealed–target correlation.

Simply comparing two cross-correlations, however, does not allow one to make statistically supported inferences. Bootstrapping (for a review of methodology, see Wasserman & Bockenholt, 1989; for specifics on application, see Farwell & Donchin, 1991) provides an answer to this problem. Bootstrapping as applied to the deception-detection paradigm involves repeatedly calculating average ERP waveforms based on different random subsamples of trials and then calculating the correlation values described above so that a distribution of correlations is obtained. One hundred iterations are performed, and on any given iteration, an equal number of epochs are selected at random (with replacement) from each of the three categories of epochs: concealed items, target items, and new nontargets. These randomly selected epochs are then used to compute an ERP for each category. The three resultant ERPs are then used to compute two cross-correlations: concealed–target correlation and concealed–nontarget correlation. The bootstrap statistic summarizes the number of iterations in which the latter correlation is larger than the former; a lower bootstrap statistic therefore is indicative of higher confidence that the concealed information is recognized.

Farwell and Donchin (1991) established cut points of a bootstrap statistic of <10 to index familiarity of >70 to index nonfamiliarity, and values between 10 and 70 were deemed indeterminate. In Farwell and Donchin's sample of 20 persons who were familiar with a critical set of information that they attempted to conceal, the procedure produced 90% correct determinations and 10% indeterminate verdicts. In their sample of 20 persons who did not have such familiarity, the procedure correctly deemed 85% of them as being unfamiliar with the material, and 15% were classified as indeterminate.

The purpose of the present investigation was to verify the utility and accuracy of the bootstrapping procedure in another data set and to examine whether the accuracy of the bootstrapping procedure was affected by motivational manipulations. Whereas several studies have found that increased motivational incentives increase the ability to detect known information and deception using skin conductance methods (Elaad & Ben-Shakhar, 1989; Gustafson & Orne, 1963), other studies have failed to find such a relationship (Davidson, 1968; Furedy & Ben-Shakhar, 1991; Horvath, 1979; Lieblich, Naftali, Shmueli, & Kugelmass, 1974). The data set to which the bootstrapping procedure was applied here is that used in the Bayesian classification of Allen et al. (1992).

Method

Participants

Sixty undergraduate students (36 women, 24 men) participated for extra credit in their psychology courses. All participants were native English speakers, reported normal or corrected-to-normal vision, and gave informed consent to participate.

Apparatus and Procedure

Full details of the procedure are presented elsewhere (Allen et al., 1992). Participants learned a list of six words from a semantic category (e.g., animals) to a criterion of perfect serial recitation, following which they participated in a simple recognition task to ensure that they could discriminate these words from new nontargets. Following a 30-min break, participants learned another list of six words from a different semantic category (e.g., clothing).

Participants then performed an ERP task in which they were to conceal the fact that they had learned the first list of words (concealed) and they were to acknowledge that they had learned the second list of words (targets). Items from each of these learned lists of words appeared on 1/7 trials, and unlearned items from five other semantic categories appeared on 5/7 trials (nontargets). All words, previously seen and new nontargets, were repeated in five

blocks of 42 words each. Stimulus duration was 306 ms, and onset-to-onset interval was 1,999 ms.

Participants were given one of three sets of instructions, which differed in terms of salience of deception and in terms of the incentive to deceive.

Conceal. The first 20 participants were told to press a button labeled *yes* if the word on the screen was one of the words they just learned and to press the *no* button for all other words, including those they had learned earlier. They were informed that "most people's brain waves give an indication of the words they learned before." They were told to conceal the fact that they learned those earlier words, that is, to keep their brain waves from letting the experimenter know which words they learned earlier.

Lie. The next 20 participants were told to press a button labeled *yes* for all words they had learned that day. They were then told to lie about having learned the first list of words. They were informed that "most people's brain waves give an indication of the words they learned before." Moreover, they were told to conceal the fact that they learned those earlier words and to conceal that they were lying about those words, that is, to keep their brain waves from letting the experimenter know that they were lying about those words they learned earlier. The essential differences between the conceal and the lie conditions involved (a) use of the word *lie* in the lie condition and (b) potentially greater conflict in response to the previously learned items. In the lie condition, the instructions provided conflicting information about the required response to previously learned items, that is, respond *yes* if learned but lie about a subset of learned words. The instructions in the conceal condition, by contrast, made explicit the response requirements, that is, respond *yes* if most recently learned, otherwise respond *no*.

Lie + money. The final 20 participants were given instructions identical to those in the lie condition, but they were also told that they would be given a $5 reward if they were successful in concealing the fact that they were lying about the set of words they had learned earlier.[1]

Recording Procedure

Scalp potentials (electroencephalogram [EEG]) were recorded from Cz, Pz, and two off-midline sites at the junction of the parietal and temporal lobes using Ag-AgCl electrodes, all referenced to linked mastoids. Only the data from Pz were included in the analysis to be consistent with Farwell and Donchin (1991) and Allen et al. (1992). Moreover, the data from the additional sites did not improve the accuracy of the procedure in the study of Allen et al. (1992).

A ground clip was affixed to the right ear. Eye movements (electrooculogram [EOG]) were recorded with Ag/AgCl electrodes in a bipolar arrangement, with superior orbit referenced to the outer canthus of the right eye. Scalp and mastoid electrode impedances were less 5 kΩ, and EOG impedances were less than 10 kΩ. Signals were amplified with Beckman AC differential amplifiers using a 30 Hz low-pass filter and a 1-s time constant and digitized on line at 200 Hz. Off-line analysis included correction for blink artifact using the method of Gratton, Coles, and Donchin (1983) and digital filtering using 51-point 5.75-Hz half-amplitude finite impulse response zero-phase-shift low-pass filter with a stopband cutoff of 12 Hz.[2] The filtered blink-corrected epochs extended from −185 ms to +1,065 ms.

Bootstrapping Analysis

The blink-corrected, digitally filtered epochs were divided according to the types of items that elicited them: concealed, targets, and nontargets. For each participant, this procedure resulted in a pool of 30 epochs for the concealed items, 30 epochs for targets, and 150 epochs for nontargets. On each iteration of the bootstrap procedure, 30 epochs were sampled with replacement from each of the three categories, and the sampled epochs were averaged to form three ERPs, one in response to each stimulus type. Two unlagged cross-correlations were then computed: ERP in response to concealed items with ERP in response to targets, and ERP in response to concealed items with ERP in response to new nontargets. Although lagged cross-correlations could be used for this purpose if substantial latency variations are present among target, concealed, and nontarget conditions, we used unlagged cross-correlations to be consistent with Farwell and Donchin (1991) and because no large latency variations among word classes were observed.

Per Farwell and Donchin (1991, p. 535), *double-centered* correlations were used, subtracting the grand average of all trials from each sampled ERP prior to computing the cross-correlations. The distributions of correlations resulting from 100 iterations were used to compute the bootstrap statistic. The bootstrap statistic ranges from 0 to 100 and reflects the number of iterations for which the concealed–nontarget cross-correlation is larger than the correlation between ERPs to learned material (concealed–target). A low bootstrap statistic thus reflects great similarity between the ERPs to the two sets of learned items and is an indication of familiarity. The cut points established by Farwell and Donchin were used, with a bootstrap statistic of <10 indicating familiarity and of >70 indicating nonfamiliarity, and values between 10 and 70 resulted in an indeterminate verdict.

Results

Bootstrapping of Cross-Correlations

The upper half of Table 2 presents the results of the bootstrapping analyses. The overall rate of accuracy of the procedure across the three studies is 87%, a rate very close to Farwell and Donchin's (1991) rate of 90% for their guilty participants. To test the effect of instructional set on the accuracy of the procedure, a chi-square test was conducted on the 2×3 table, excluding the Not-Recognized row because no observations occurred in this row and because the distribution of the chi-square statistic is sensitive to small cell sizes. The accuracy of the bootstrapping procedure at detecting concealed information was influenced by instructional set ($\chi^2 = 8.08$, $p = .02$), with higher rates of accuracy occurring as the motivational incentive to deceive increased.

Figure 1 displays the ERP waveforms grand averaged across participants under each of the instructional sets. Because the accuracy of the bootstrapping procedure was affected by the instructional set, analyses were conducted to determine whether instructional set influenced P3 amplitude differentially by item type (unlearned, concealed, target), which would be indicated by a significant interaction in a 3 (instructional set: conceal, lie, lie + money) × 3 (item type) repeated measures analysis of variance (ANOVA) on P3 am-

[1] Actually, all participants were paid the $5.

[2] This filter also produced no ripple greater 0.1% of the original signal amplitude in frequencies beyond 12 Hz.

Table 2. *Bootstrapping Outcomes as a Function of Instructional Set, When Concealed Items Are Familiar or Not Familiar*

	Instructional set		
Verdict	Conceal	Lie	Lie + money
Familiar to participants			
Recognized	14	18	20
Indeterminate	6	2	0
Not recognized	0	0	0
Not familiar to participants			
Recognized	0	0	0
Indeterminate	7	6	4
Not recognized	13	14	16

plitude (defined as the maximum positive value in a search window of 350–850 ms). P3 amplitude was significantly influenced by the type of item, $F(2,114) = 125.17$, $p < .001$, $\epsilon = 0.91$, and although P3 amplitude differed as a function of instructional set, $F(2,57) = 3.23$, $p = .05$, instructional set did not interact with type of item, $F(2,114) = 1.08$, *ns*, $\epsilon = 0.91$). Post hoc contrasts for instructional set revealed larger P3 amplitudes in the lie + money condition than in the lie condition ($p < .02$), with no other pairwise comparisons significant. Post hoc tests for item type revealed that P3 amplitude was significantly larger in response to target items than in response to concealed items ($p < .001$) and significantly larger in response to concealed items than in response to unlearned items ($p < .001$).

If P3 amplitude for each item type is not differentially affected by instructional set, then the other features of the ERP waveforms may have been. Examination of P1, N1, P2, and N2 amplitudes, however, revealed no evidence for a significant interaction between instructional set and item type, all $Fs(4,114) \leq 1.06$, *ns*, all $\epsilon s \geq 0.78$. Similarly, examination of latency revealed no significant interactions between instructional set and item type for P1, P2, N2, or P3 components, although the interaction was significant for N1 latency, $F(4,114) = 3.23$, $p < .05$, $\epsilon = 0.96$. Decomposition of this interaction revealed that whereas N1 latency did not differ as a function of list in the conceal or the lie + money conditions, in the lie condition N1 latency was shorter ($p < .01$) to concealed items ($M = 193.8$ ms, $SD = 15.6$ ms) than to unlearned nontargets ($M = 204.5$ ms, $SD = 11.5$ ms), with no other significant pairwise comparisons.

The preceding analyses revealed very few interactions of instructional set with item type. The morphology of the waveforms, therefore, must have been affected in a manner other than simple amplitude and latency differences. To characterize the similarity of the waveforms, the cross-correlations derived from the bootstrapping analysis were analyzed. For each participant, the bootstrapping procedure tallied the median concealed–target correlation and the median concealed–unlearned correlation. These median correlations for each participant, after Fisher *z*-transformation to render the correlations suitable for parametric analysis, were subjected to a one-way ANOVA with instructional set as the between-participants variable. Although the concealed–unlearned correlation was unaffected by instructional set, $F(2,57) = 1.55$, *ns*, the concealed–target correlation was significantly larger in the lie + money condition than in the conceal condition, $F(2,57) = 3.39$, $p < .05$, post hoc test by the Tukey studentized range method, $p < .05$, with no other significant pairwise tests (mean concealed–target correlations after converting from mean Fisher *z* scores: conceal = .46, lie = .51, lie + money = .67).

Simulation of Innocent Participants

All participants in the current study were "guilty" of knowing the concealed information and were therefore comparable to Farwell and Donchin's (1991) guilty participants. To examine the accuracy of the bootstrapping for innocent participants, the epochs in response to concealed items were replaced by a set of an equal number of epochs in response to nontarget items. These substituted nontarget items were then treated as concealed items rather than as nontarget items. This procedure resulted in a pool of 30 epochs for these substituted concealed items, 30 epochs for targets, and only 120 epochs for nontargets. As shown in the lower half of Table 2, the bootstrap procedure correctly assessed 72% of these participants to be unfamiliar with the concealed material, and indeterminate verdicts resulted 28% of the time. This finding compares with 85% accuracy for Farwell and Donchin's innocent participants. The accuracy of the bootstrapping procedure, however, was not affected by instructional set for these simulated innocent participants ($\chi^2 = 1.12$, *ns*, for the 2×3 table excluding the Recognized row).

Exploratory Analyses to Increase Applicability of Bootstrapping Methods

The bootstrapping of cross-correlations is applicable only to those rare situations where there are three classes of stimuli. In the interest of making the technique more widely applicable to situations with two classes of stimuli, we explored a simpler method that relies not on the comparison of two cross-correlations but on a simple comparison of amplitudes. The logic of this set of analyses was that if concealed items were recognized as a distinct, significant, and rare class of stimuli relative to the nontargets, then the amplitude of ERP in the region of the P3 should be larger for concealed items than for nontargets. To operationalize this procedure, on each iteration the 100-ms bin with the maximum amplitude (averaged across the 100 ms bin) in the time window from 350 ms to 850 ms was identified for the nontarget items and for the concealed items. The bootstrap statistic ranged from 0 to 100 and reflected the number of iterations for which the amplitude to concealed items exceeded that to nontarget items. A high bootstrap statistic thus reflected familiarity with the concealed items. Using the same criterion as with the bootstrapping of cross-correlations, that ≥90% of the trials must be indicative of familiarity to conclude that the participant recognized the concealed items, this simplified amplitude bootstrapping procedure correctly identified the concealed material as familiar to the participant 80%, 75%, and 85% of the time across the three groups (conceal, lie, lie + money), a rate not significantly influenced by instructional set ($\chi^2 = 0.63$, *ns*). For comparison purposes, this amplitude-based bootstrapping procedure correctly identified the target items as familiar to the participant 85%, 80%, and 90% of the time across the three studies, a rate not significantly influenced by instructional set ($\chi^2 = 0.78$, *ns*).

To ascertain how well the bootstrapping of amplitudes worked when participants had not learned the concealed information, once again the epochs in response to concealed items were replaced by a set of an equal number of epochs to nontarget items, with these substituted nontarget items treated as concealed items rather than nontarget items in the bootstrapping analysis. This analysis incorrectly identified these substituted items as familiar to par-

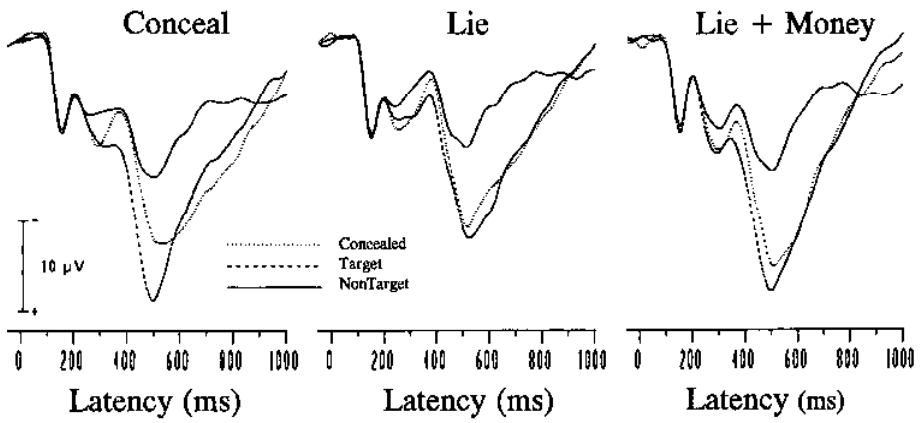

Figure 1. Grand average ERP waveforms under three different instructional sets.

ticipants 0%, 10%, and 5% of the time across the three studies ($\chi^2 = 2.11$, *ns*).

A Comparison of the Methods Using Receiver Operating Characteristic Curves

The above comparison of the Bayesian and the two bootstrapping methods involved the use of cut points established in previous research to differentiate learned from unlearned items or, in the case of the bootstrapping of cross-correlations, learned from unlearned items with the possibility for indeterminate outcomes. To compare the methods directly, receiver operating characteristic (ROC) curves were used to compare the efficiency of the three methods (Bayesian, bootstrapping of cross-correlations, and bootstrapping of amplitudes) without respect to any fixed cut point or points. The ROC curves for the three methods under three different instructional conditions are displayed in Figure 2. Each ROC curve plots the true positive rate (TPR) as a function of the false positive rate (FPR), providing a succinct synopsis of a given method's performance across the entire range of possible cutoff scores (Mossman & Somoza, 1991). Moreover, the analysis of ROC curves allows a comparison of tests that have different metrics, comparing their ability to discriminate learned from unlearned items across the entire range of cutoff scores rather than at an arbitrary fixed cut point or points. The area under the ROC curve (AUC) provides a simple metric that summarizes the method's performance, with an AUC of 1.0 indicating perfect discrimination between learned and unlearned items and an AUC of 0.5 indicating no discrimination whatsoever (Mossman & Somoza, 1991). By testing whether the AUCs are significantly different as a function of instructional set, the performance of each method under different motivational conditions can be compared across the entire range of cut points. The AUC did not differ as a function of instructional set for the Bayesian method (AUC $\pm$ *SE*: conceal = 0.98 $\pm$ 0.02, lie = 0.93 $\pm$ 0.05, lie + money = 0.95 $\pm$ 0.03), for the bootstrapping of cross-correlations (conceal = 0.98 $\pm$ 0.02, lie = 0.97 $\pm$ 0.03, lie + money = 1.0 $\pm$ 0.0), or for the bootstrapping of amplitudes (conceal = 0.94 $\pm$ 0.05, lie = 0.90 $\pm$ 0.07, lie + money = 0.97 $\pm$ 0.02).

These analyses highlight that all methods produce impressive classification accuracy (all AUCs > .90) and indicate that the apparent sensitivity of the bootstrapping method to motivational instructions is a result of the particular cut point selected. These cut points were selected in previous studies, however, to maximize classification accuracy while minimizing the FPR. ROC curves can also be used to evaluate the performance of two tests at a given FPR, which may involve different cut points for different curves. At the statistically customary FPR of 5%, the TPRs (tested using their associated nonsymmetric confidence intervals; Metz & Kronman, 1980) differed significantly as a function of instructional set for the bootstrapping of cross-correlations but not for the other two methods (middle panel of Figure 2; TPR = 1.0 at FPR = 0.05 for the lie + money condition). Using the bootstrapping of cross-correlations, the true positive rate of the lie + money condition was significantly greater than that of the conceal or the lie conditions, which did not differ from one another.

The ROC curves provided a means of directly comparing the three methods, collapsed across instructional set (see Figure 3). In comparing the methods to one another, the procedure of Metz, Wang, and Kronman (1984) was used to account for the fact that the ROC data matrices are correlated when the methods compared involve the same participants. The AUC was significantly lower for the bootstrapping of amplitudes (0.95 $\pm$ 0.02) than for either the Bayesian method (0.97 $\pm$ 0.02) or the bootstrapping of cross-correlations (0.99 $\pm$ 0.01). The latter two methods did not differ significantly from one another.

Discussion

Bootstrapping is a highly flexible procedure that can be used to derive a distribution for virtually any statistic (Wasserman & Bockenholt, 1989). When cross-correlations between ERP waveforms were used in the present study and when previously established cut points (Farwell & Donchin, 1991) were applied, the overall accuracy of the bootstrapping procedure at detecting concealed material that participants recently learned was highly similar (87% vs. 90%) to that of the study of Farwell and Donchin (1991). When this procedure was applied to the simulated innocent participants in the current studies, the procedure was somewhat less accurate (72% vs. 85%) than previously reported (Farwell & Donchin, 1991). However, both in this study and in that of Farwell and Donchin, the procedure has yet to produce a false positive verdict when applied to innocent participants.

When a simpler procedure with potentially wider applicability was applied to these data, the method of bootstrapping of amplitudes correctly identified concealed material as familiar to partici-

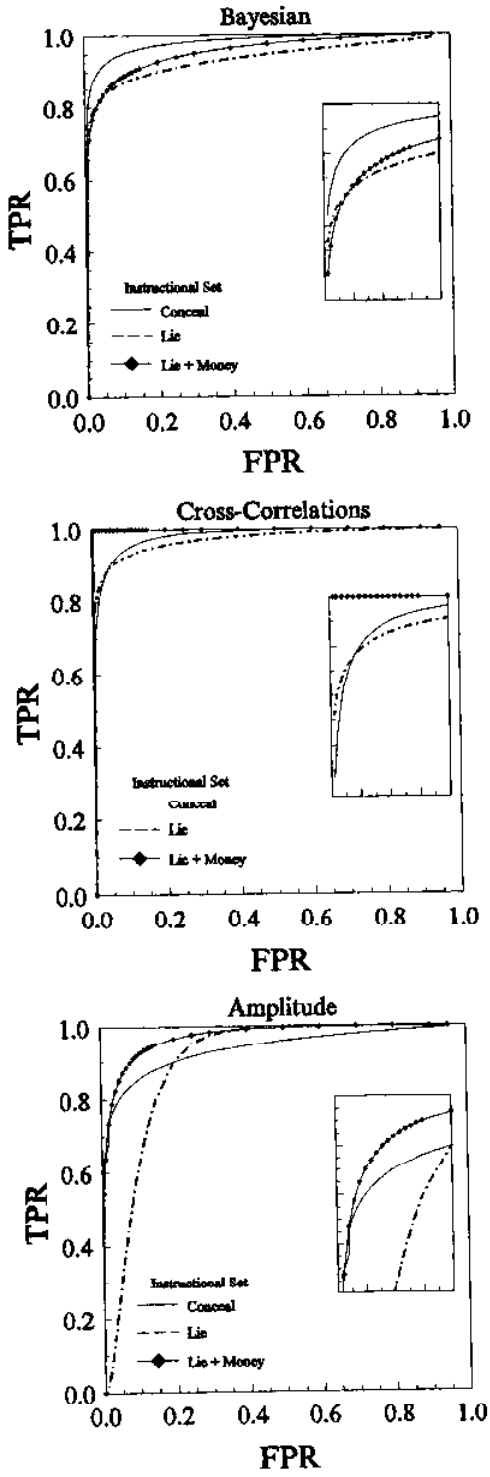

Figure 2. Receiver operating characteristic (ROC) curves for the three analytic procedures under three different instructional sets. Bayesian = Bayesian combination of ERP indicators; cross-correlations = bootstrapping of cross-correlations; amplitude = bootstrapping of amplitudes. The true positive rate (TPR) for the identifying learned material is compared with the false-positive rate (FPR) for identifying unlearned material in the simulated innocent participants. The inset in each ROC curve enlarges the region from TPR = 0.6–1.0 and from FPR = 0.0–0.20.

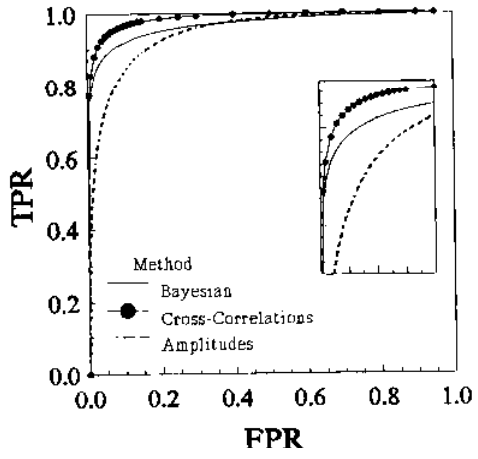

Figure 3. Receiver operating characteristic (ROC) curves for the three analytic procedures collapsed across the three instructional sets. Bayesian = Bayesian combination of ERP indicators; cross-correlations = bootstrapping of cross-correlations; amplitudes = bootstrapping of amplitudes. The true positive rate (TPR) for the identifying learned material is compared with the false-positive rate (FPR) for identifying unlearned material in the simulated innocent participants. The inset ROC curve enlarges the region from TPR = 0.6–1.0, and from FPR = 0.0–0.20.

pants 80% of the time (instead of 87% with the cross-correlation procedure) and incorrectly identified unlearned material as familiar in simulated innocent participants 5% (instead of 0%) of the time. Whether a simple bootstrapping of amplitudes might prove useful in other applications remains an empirical question. In paradigms where an investigator is interested in comparing ERP waveforms for two classes (rather than three classes) of stimuli, such a procedure offers an alternative and as yet not fully tested means of making determinations for individuals.

As the motivational incentive to deceive increases, the accuracy of the cross-correlational bootstrapping procedure at detecting concealed knowledge in guilty participants also increases, based on the preexisting cut points established by Farwell and Donchin (1991). The results of the ROC curve analyses essentially corroborated these findings. Although there was not a significant effect of incentive to deceive on the AUC, at a statistically customary FPR of 5%, the highest motivational incentive produced a TPR that is significantly higher than that under the two lower motivational incentives. This finding suggests that should such an ERP-based assessment procedure be used in field applications, in situations where the motivational incentive to deceive would likely be high, the procedure may actually be more accurate than the analog studies suggest. Empirical work is clearly needed to assess this possibility and to provide cross-validation of the present findings. The motivational incentives in some field applications (e.g., avoiding an aversive consequence such as imprisonment) are quite different than those in the lab (e.g., obtaining a small monetary reward). By contrast, in other field applications, such as malingering a memory problem to obtain financial compensation, the motivational incentives may be similar to but larger than those incentives provided in the lab.

Why the procedure is more accurate as the incentive to deceive increases is open to multiple interpretations. Although a case could be made invoking motivational, emotional, or deception-related

mechanisms, it is parsimonious to explain these results in terms of a the triarchic model of the P3 class of components (Johnson, 1986). The instructional sets differ in terms of how much salience or significance is placed on the items to be concealed, ranging from "conceal them" to "lie about them" to "lie successfully about them and receive cash rewards." P3 amplitude is, in part, a function of stimulus significance[3] (Johnson, 1986). The instructional set with the greatest incentive (monetary reward) produced larger P3 amplitudes than did the instructional set with the least incentive, although the effect of the incentive was not different for unlearned, concealed, or target items; all produced larger P3 amplitudes in the condition with the greater incentive. These results suggest that increased incentive to deceive increases stimulus significance and thereby increases P3 amplitude for all item types, not solely for the concealed items.

The data from the present study using the bootstrapping of cross correlations can be compared with the results of the Bayesian analysis of the data for these same participants (Allen et al., 1992), where the accuracy of the procedure was not affected by instructional set. Using the cut points established and cross-validated in those studies, the concealed list was detected as familiar for 92% of the participants using the Bayesian approach (87% using bootstrapping), whereas nontarget material was correctly classified as not learned for 96% of the participants (72% using bootstrapping).

A direct comparison of all three methods using ROC curves suggested that whereas all three methods produced high rates of classification accuracy, the bootstrapping of amplitudes produced significantly less accurate classifications than did either the Bayesian method or the bootstrapping of cross-correlations. All methods compare favorably to skin conductance procedures, where AUC estimates in the range of 0.77–0.85 have been observed (Elaad & Ben-Shakhar, 1989). In fact, there is no overlap between the AUC confidence intervals of the skin conductance measures (Elaad & Ben-Shakhar, 1989) and those of any of the three ERP methods in the present study, indicating that the ERP methods produce statistically superior classification accuracy. Of course, a direct comparison within participants of these methods would be important to verify that the methods (ERP vs. skin conductance) differ per se. It remains possible that other variables (e.g., the population under study) account for the superiority of the ERP methods.

Three essential differences exist between the bootstrapping procedure and the Bayesian procedure. First, at a customary FPR of 5%, the accuracy of the bootstrapping of cross correlations but not that of the Bayesian procedure is influenced by motivational instructions. Second, the distribution of bootstrapping statistics is such that the use of indeterminate verdicts improves classification relative to a dichotomous classification rule, whereas the distribution of probabilities from the Bayesian method does not produce a benefit from an indeterminate classification. This difference in distributions allows for the possibility of indeterminate verdicts for the bootstrapping of cross-correlations, whereas errors of classification using the Bayesian method resulted in false-positive and false-negative verdicts. Third, the Bayesian procedure can be applied to other paradigms and can include other measures in addition to ERP data. These considerations suggest that each method may be preferred under different circumstances, depending upon the nature of the application, the incentive to deceive, the cost of incorrect decisions, and the availability of extra-psychophysiological data. These considerations might be fruitfully investigated in future laboratory and field studies.

[3]Stimulus significance may reflect emotional as well as cognitive determinants. The present study does not differentiate between these factors nor does it provide a basis for speculation concerning the extent to which each may be involved.

REFERENCES

Allen, J. J., Iacono, W. G., & Danielson, K. D. (1992). The development and validation of an event-related-potential memory assessment procedure: A methodology for prediction in the face of individual differences. *Psychophysiology, 29*, 504–522.

Boaz, T. L., Perry, N. W., Raney, G., Fischler, I. S., & Shuman, D. (1991). Detection of guilty knowledge with event-related potentials. *Journal of Applied Psychology, 76*, 788–795.

Davidson, P. O. (1968). Validity of the guilty-knowledge technique: The effects of motivation. *Journal of Applied Psychology, 52*, 62–65.

Elaad, E., & Ben-Shakhar, G. (1989). Effects of motivation and verbal response type on psychophysiological detection of information. *Psychophysiology, 26*, 442–451.

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

Furedy, J. J., & Ben-Shakhar, G. (1991). The roles of deception, intention to deceive, and motivation to avoid detection in the psychophysiological detection of guilty knowledge. *Psychophysiology, 28*, 163–171.

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55*, 468–484.

Gustafson, L. A., & Orne, M. T. (1963). Effects of heightened motivation on the detection of deception. *Journal of Applied Psychology, 47*, 408–411.

Horvath, F. (1979). Effect of different motivational instructions on detection of deception with the psychological stress evaluator and the galvanic skin response. *Journal of Applied Psychology, 64*, 323–330.

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

Lieblich, I., Naftali, G., Shmueli, J., & Kugelmass, S. (1974). Efficiency of GSR detection of information with repeated presentation of series of stimuli in two motivational states. *Journal of Applied Psychology, 59*, 113–115.

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

Metz, C. E., & Kronman, H. B. (1980). Statistical significance tests for binormal ROC curves *Journal of Mathematical Psychology, 22*, 218–243.

Metz, C. E., Wang, P. L., & Kronman, H. B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In F. Deconinck (Ed.), *Information processing in medical imaging: Proceedings of the eighth conference* (pp. 432–445). The Hague: Martinus Nijhoff Publishers.

Mossman, D., & Somoza, E. (1991). ROC curves, test accuracy, and the description of diagnostic tests. *Journal of Neuropsychiatry, 3*, 330–333.

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

Rosenfeld, J. P., 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, 42*, 157–161.

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

(Received August 10, 1995; Accepted Accepted June 28, 1996)