# Bootstrapping the P300 in diagnostic psychophysiology

PSYCHOPHYSIOLOGY

*Psychophysiology, 54* (2017), 366–373. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/psyp.12789

# Bootstrapping the P300 in diagnostic psychophysiology: How many iterations are needed?

J. PETER ROSENFELD,[a] ANNE WARD,[a] EWOUT H. MEIJER,[b] AND DENIS YUKHNENKO[b]

[a]Department of Psychology, Northwestern University, Evanston, Illinois, USA
[b]Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

**Abstract**

In psychophysiological research, bootstrapping procedures are often used to classify individual participants. How many iterations are required for a reliable bootstrap test is not universally agreed upon. To investigate the number of iterations needed for a stable bootstrap estimate, we reanalyzed P300 data collected in concealed information test paradigms. We also distinguished between the bootstrap and permutations approaches. We compared results in several studies using 100 versus 1,000 versus 10,000 iterations in the bootstrap, and we concluded that 100 iterations were adequate as results from all three iteration numbers correlated highly.

**Descriptors:** Bootstrapping, P300, Permutations

Several applications of psychophysiology require diagnostic classification at the individual level. That is, one often wishes to use physiological responding to distinguish between two or more independent conditions within one individual, and use this distinction to classify an individual as, for example, bipolar versus unipolar, attention deficit hyperactivity disorder (ADHD) versus normally attentive, schizophrenic versus paranoid, or high risk versus low risk. One notable example of a field where decisions at the individual level are crucially important is that of memory detection using the Concealed Information Test (CIT, also referred to as the Guilty Knowledge Test; Lykken, 1959). The CIT assumes that a guilty person would possess information that is known only to the police, the victims, and the person who committed the crime. As such, it aims to classify testees as guilty, inferred from the presence of intimate knowledge of crime details, or as innocent, inferred from the absence of such knowledge. In sum, the CIT detects whether the suspect knows and recognizes the critical crime-related information.

Since the late 1980s (e.g., Rosenfeld, 2011), recognition in the CIT has been indexed in one research area by the fact that the crime-related stimulus item (called the *probe*; e.g., a 356 Magnum revolver) evokes in knowledgeable individuals a large P300 component of the ERP. On the other hand, other items similar to the probe, but not the actual murder weapon used (e.g., 32 Colt, 45 automatic, 9 mm Luger, 9 mm Beretta, etc.) do not evoke a P300 as large as that evoked by the crime-related probe. These other items are typically called *irrelevants*.

Two variants of the CIT exist. In the two-stimulus protocol (Rosenfeld, Shue, & Singer, 2007), the suspect views the probe and irrelevants, one at a time, in a Bernoulli series in one block. He or she presses the same response button for both probes and irrelevants. The classification depends on whether the probe P300 amplitude exceeds the irrelevant P300 amplitude. Alternatively, in the much better known three-stimulus protocol (Farwell & Donchin, 1991; Rosenfeld, 2011), a special third irrelevant stimulus is occasionally presented, requiring a unique button press different from the one pressed in response to probe or irrelevant. This special irrelevant is called a *target* stimulus and is used to hold attention, as well as sometimes for analytic reasons. The statistical question to be answered is then whether the probe stimulus resembles more the target (recognition inferred) or resembles more the irrelevant (no recognition inferred).

Distinguishing between probe and irrelevant P300 waveforms is not restricted to concealed information testing. In a variety of other clinical diagnostic situations, the same form of diagnostic question is posed within one test patient/client: Is the rare target P300 larger than the frequent nontarget P300? The answer typically helps decide how to classify the patient—as having some disorder (e.g., attention deficit disorder [ADD], dementia, Alzheimer's, schizophrenia) or not (Polich, 2004). Note that probes and irrelevants can be thought of as special (forensic) cases of more general target and nontarget stimuli in the standard oddball paradigm (Donchin, 1981). That is, both probes and targets are usually relatively rare, occurring in 10% to 30% of the trials, and meaningful—whereas nontargets and irrelevants are usually both frequent (occurring in 70% to 90% of the trials) and of neutral meaning.

We also note that bootstrapping may be used with psychophysiological variables other than P300 (e.g., heart rate variability, blood pressure, EEG measures [alpha asymmetry, beta-theta ratio, etc.], and so on; Wasserman & Bockenholt, 1989).

We are grateful to Ulf Bockenholt of Kellogg School of Management, Northwestern University, Evanston, IL, for valuable consultation.

Address correspondence to: J. Peter Rosenfeld, Department of Psychology, Northwestern University, Swift Hall 102, 2029 Sheridan Road, Evanston, IL 60208, USA. E-mail: jp-rosenfeld@northwestern.edu

### Groups Versus Individuals

In nondiagnostic P300 studies, in which one compares the effect of an independent variable such as acoustic loudness on P300 amplitude, one might have a high loudness group and a low loudness group, and then compare the averaged ERPs of the low and high loudness across subjects in a between-subjects *t* test. In this case, the basic unit of comparison is the individual, relatively noise-free average of multiple single sweeps, separately averaged for the high and low conditions. Signal averaging of scalp-recorded ERPs is customary since single sweeps are noisy. In the diagnostic situation, however, in which ERPs are compared within one subject, if one ran the usual *t* test on the effect of two conditions on the two condition means, one would have to use single-sweep ERP values; that is, the basic unit of comparison is the (noisy) single sweep, and the *t* test results are typically insensitive (Rosenfeld, Angell, Johnson, & Qian, 1991). One way to obtain actual multiple average ERPs (which are far less noisy than single sweeps) from one subject would be to repeat the study at least 20 times. This would yield several averages (each of $n = 20$) for one subject, but the effects of habituation over repeated studies would be fatal. Moreover, in the case of the P300 CIT, the multiple repetitions would likely make the irrelevant stimuli familiar and increasingly relevant through repetition, so that they would become difficult to distinguish from the probe.

Multiple solutions to this dilemma of comparison of probe and irrelevant averages within an individual exist. One is the bootstrap method (Efron, 1979), introduced into psychophysiology by Karis, Fabiani, and Donchin (1984). It was later adapted by Farwell and Donchin (1991), Rosenfeld, Sweet, Chuang, Ellwanger, and Song (1996), and Ellwanger, Rosenfeld, Sweet, and Bhatt (1996) for the problem of discriminating knowledgeable and not-knowledgeable individuals in a P300-based CIT. Rather than unrealistically repeating a study within an individual, these researchers repeated random selection (with replacement) of single-sweep data subsets, each of which was averaged into a bootstrapped average ERP for both the probe and irrelevant categories (resampling without replacement substituted for replication). Because of the bootstrapping without replacement process, it is unlikely that any bootstrapped average will be exactly the same as any other, or be exactly the same as the actual sample average based on all the original single sweeps. Rosenfeld's lab (Rosenfeld, 2011; Rosenfeld, Hu, Labkovsky, Meixner, & Winograd, 2013) simply generated a set of $n1$[1] resampled (with replacement) from the original sample set of $n1$ probe sweeps, averaged to yield one bootstrapped average probe ERP, and likewise from the original sample set of $n2$ irrelevant single sweeps, to yield one bootstrapped average irrelevant ERP. The computed P300 difference[2] between each pair of probe and irrelevant bootstrapped averages was placed in a distribution of probe-minus-irrelevant differences (P − I), the process repeated multiple times, and a decision of knowledgeable was based on the finding that 90% or more of these bootstrapped P − I P300 differences was > 0. Of course, one could choose another criterion (e.g., 85% or 95%). These methods have been verified empirically: They were used to determine knowledgeability based on experimentally manipulated knowledgeable and not-knowledgeable groups. Twenty-two knowledgeable groups (including 10 using countermeasures) were reviewed by Rosenfeld et al., (2013), and the mean area under the ROC curve (AUC) using also 22 not-knowledgeable groups was .931.

A second method that can compare probe and irrelevant averages within an individual is Fisher's permutation test (Efron & Tibshirani, 1994, p. 205), which is related to but not identical to the bootstrap. These two resampling methods have been a source of confusion. Here is a typical situation in which either bootstrapping or permutations methods may be applied: One has two sample sets of single-sweep ERPs, which in our research are (1) probe (say $n = 30$), and (2) irrelevant (say $n = 150$) ERPs, from one subject who may be from either of two groups of subjects: K (knowledgeable of crime details) and N (not knowledgeable). The K group members have seen and know the probe item. The average recognized (rare and meaningful) probe P300 from the K group should exceed the averaged irrelevant P300. To members of the N group, the probe item is meaningless and is just another irrelevant, so that there should be no difference between the average probe and irrelevant P300s. The task is to determine if the given single subject is K or N. That determination depends on deciding within the individual if there is an improbably large difference (e.g., $p < .05$ or .1) between his probe and irrelevant P300s. Standard *t* tests cannot be used here as they are insensitive (see Rosenfeld, 2011; Rosenfeld & Donchin, 2015).

As outlined above, the bootstrapping method resamples—but with replacement of the selection after each selection—the original set of 30 probe sweeps 30 times. Again, each resampling will rarely yield the original set of probe sweeps due to sampling with replacement. Then, for each set of these 30 resampled single sweeps, an average bootstrapped ERP is computed, and the average P300 value over 30 bootstrapped, randomly selected single sweeps is determined; where P300 is defined, for example, as the largest 100-ms segment average in a window running from 400–800 ms poststimulus. The same is done with the set of 150 irrelevant single sweeps to yield an average irrelevant P300, also based on 30 resamplings (from the original 150 single sweeps). Now, a P − I value is determined for this pair of first bootstrapped probe and irrelevant values. The process is repeated—in our hands—100 times, so that we now have a distribution of 100 P − I differences. Typically, we then determine if 90% (or 85% or 95% or whatever criterion percentage is chosen) of these differences are > zero. If so, the subject is classified K; if not, then N. As noted by Efron and Tibshirani (1994, p. 5), the term *bootstrap* seems most apt for this algorithm, since one is repeatedly generating average P300s, and then averaging these into a grand average by resampling the bootstrapped single sweep set; one in effect pulls up the grand average by its boot straps—the individual resampling averages. (See Figure 1 for a bootstrapping flow chart.)

In comparison, here is what the typical permutations method does with the original sets of probe and irrelevant single sweeps (there are anomalous other permutations methods as described below). First, it creates a combined data set distribution: It takes the 30 probe and 150 irrelevant single sweeps and shuffles them together (as one might shuffle 30 royal playing cards—Kings, Queens, Jacks—together with 150 number cards numbered 1–10) into a single pooled set of 180 sweeps ("cards"). Now one "cuts the cards" by randomly selecting a set of 30 sweeps (cards) for one new subsample from the pooled distribution. The 150 single sweeps (cards) remaining in the pooled distribution become a

1. $n1$ was the number of sweeps in the originally collected sample of probe single sweeps, and $n2$ was the same for original irrelevant sweeps, multiplied by a fraction that would bring $n2$ to be equal to $n1 \pm 1$.

2. We note that, although P300 is usually thought of as the peak—one data point—of a positive wave between 300 and 800 ms poststimulus, in the Rosenfeld lab, we find that it improves the S/N ratio to compute the mean amplitude of the most positive 100-ms segment in the 300–800 ms "look window."

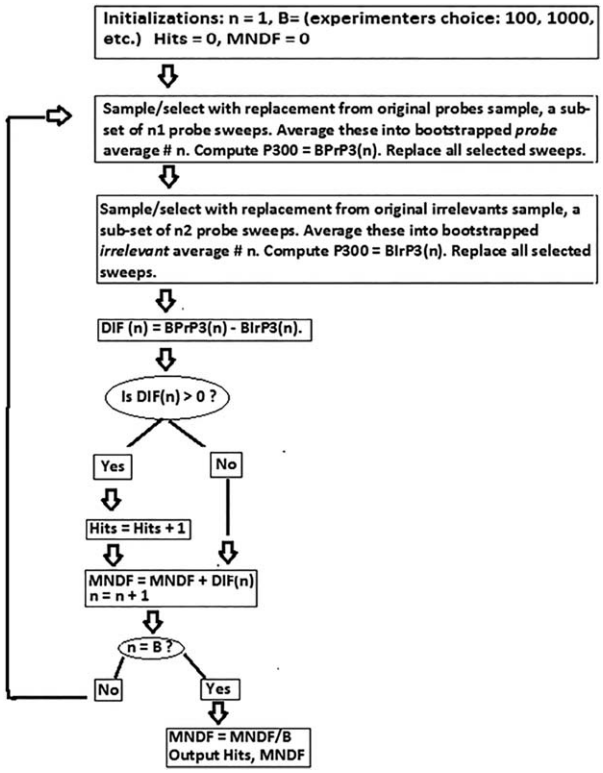


**Figure 1.** Flow chart showing the basic structure of the bootstrap. Main outputs: MNDF = bootstrap-estimated probe-irrelevant difference; Hits = number of times a bootstrapped probe average exceeds a bootstrapped average irrelevant average; B = experimenter-defined number of bootstrap iterations. Input is a set of originally collected probe sweeps and an originally collected set of irrelevant sweeps. BPrP3(n) is the bootstrapped probe P300 for the *n*th iteration. BIrP3(n) is the bootstrapped irrelevant P300 for the *n*th iteration.

second subsample from the pooled distribution. One now has two data sets whose expected mean difference value is zero, since they are samples from the same population distribution. The means of these two data sets are computed, and the average difference (probably small but not exactly equal to zero) found. This process is also repeated some large number of times, and one obtains a distribution of iterated null differences with mean = zero. Next, the actually obtained individual average difference from the original sets of probe and irrelevant P300s is tested to see if it is within the top (5% or 10%) tail of the null distribution. If it is, the subject is deemed to be in the K group, if not, then the N group. (See Figure 2 for a permutations method flow chart.)

As Zoumpoulaki, Alsufyani, and Bowman (2015) demonstrated, if one uses this permutation method to estimate the maximum (peak) value of a single data point, one gets something systematically different from what one gets by using a bootstrap algorithm. This fact clearly indicates that the permutation method is quite distinct from the bootstrap method. Efron and Tibshirani (1994, p. 207) nevertheless noted that "the permutation algorithm is quite similar to the bootstrap algorithm ... The main difference is that sampling is carried out without replacement, rather than with replacement." However, this could be unintentionally misleading, because in Fisher's permutation test, only one sample pair (of equal number, unlike what we described above) can be randomly selected from the combined distribution for each iteration, which determines the remaining second sample. But with each new iteration, the combined distribution is reconstituted (i.e., all data are replaced prior to the next sample drawing). In any case, our previous simulation (Rosenfeld & Donchin, 2015) demonstrated rather clearly that if one bootstraps or permutes mean values of 100-ms long ERP segments (as opposed to single peak maxima), both resampling techniques (bootstrapping and permutations) agreed about 98% of the time in thousands of repeated simulations.

An alternative permutations method, directed toward hypothesis testing on a group of subjects, was presented by Blair and Karniski (1993). It is critical to point out here that Blair and Karniski were dealing with an entirely different research question than the one relevant in CIT research described above. Blair and Karniski aimed to determine if there are differences in ERP amplitude—and where they are temporally located—to two kinds of stimuli between two grand-averaged waveforms (each consisting of the average of 15 individual averages). Obviously, this is not an intraindividual diagnostic problem. It is analogous to a repeated measures *t* test.

In their first example, a set of data values (say, mean P300 values) is tabulated for three subjects in two conditions: probe versus irrelevant presentations:

| Subject | Probe | Irrelevant |
|---|---|---|
| 1 | 8 | 5 |
| 2 | 4 | 3 |
| 3 | 6 | 4 |

A repeated measures *t* test is now computed, yielding $t = 3.46$. Assuming the null hypothesis, the values of probes and irrelevant P300s for each subject are interchangeable, so that there are $2^3 = 8$ variations of the above table. One can now calculate the probability (one-tailed) of obtaining $t = 3.46$ under a true null hypothesis. It is 1/8 = .125; (two-tailed: 2/8 = .25). The value .125 becomes the one-tailed null rejection criterion. Note the results are in terms of exact probabilities since all permutations are used. Note again that this test is not a within-individual test, but an alternative within-groups *t* test based on all existing actual data. To apply this method to the intraindividual case, the three rows above would have to be conceptualized as obtained from one subject, and this would involve resampling from a single individual. Indeed, even in their second example, a larger group test case, Blair and Karniski (1993) consider a data set involving 15 subjects, noting that 32,768 permutations are required if one requires an exact probability test based on all permutations of all existing data. But they then show that, by using a randomly reselected set of, say, 10,000 permutations, the test results are quite similar to the exact case. Thus, in the commonly utilized Fisher permutations test applied to the individual case, as outlined above and in Figure 2, one does indeed combine all actual probe and irrelevant P300 values, then divides this null distribution into two fractions (pseudoprobe and pseudoirrelevant subdistributions), computes the mean difference—and, by repeating this process, generates a difference distribution of null mean differences. One then examines where the actually obtained true probe-irrelevant difference sits. If it is beyond the criterion value (e.g., .01, .05, .1, etc.), the no-difference hypothesis is rejected. Thus, with both bootstrapping and, as it is commonly used, permutations, all the data are repeatedly resampled, and the statistical test outcome becomes more stable as the number of iterations increases.

## Number of Iterations

The number of iterations used to calculate the statistical outcome has varied between studies. In this article, we test how many

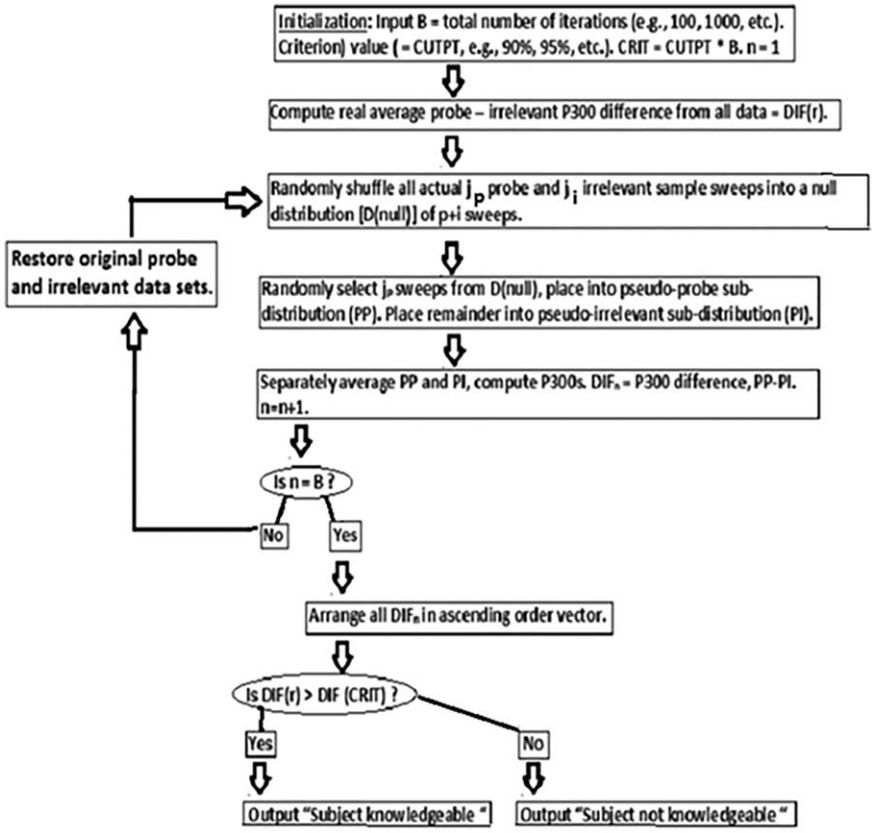


**Figure 2.** Flow chart showing the basic structure of the Fisher's permutation test (Efron & Tibshirani, 1994) method. One major output is whether or not the single test subject is knowledgeable versus not knowledgeable. Input is a set of originally collected probe sweeps and an originally collected set of irrelevant sweeps.

iterations are needed for reliably classifying individuals in the P300-based CIT situation. Because bootstrapping has been mostly used in diagnostic psychophysiology based on P300, this is the method we study here so as to determine adequate numbers of iterations. Critically, in determining how many iterations/resamplings one must use in either permutations or bootstrap methods, one must primarily consider the signal/noise (S/N) ratio for the component of interest, since the smaller its S/N ratio (as with effect size) the larger must be the sample examined. Our probe to irrelevant ratios are usually about 14% to 86% (1 to 7); that is, our oddballs are rare, meaning there will be large probe P300s. But this is hardly true for many other (smaller) ERPs (such as N200, the old-new late positivity, N400, and so forth), whose stimulus ratios are usually 50% to 50%.

In recent years, there has also been some confusion about the proper number of iterations, which we denote as *B*. For example, DiNocera and Ferlazzo (2000) used the bootstrapping and permutations terms, in effect, interchangeably, and dealt with a question that somewhat resembles the CIT bootstrapping question posed here. These authors provided their participants with a list of 160 words, and then presented those 160 words, alternated with 160 new words. The question was whether or not the authors, based on ERPs, could accurately diagnose—within each single subject—whether the subject had previously seen a word on a list or not. One thousand (1,000) resamplings were used so as to offer a diagnosis with the permutations (not bootstrap) method. Importantly, these authors started out with 160 trials in each condition, considerably more than typically used in CIT research.

For another example, authors such as Blair and Karniski (1993) recommended 10,000 iterations for the permutation test they described (which again, importantly, was directed to groups, not individuals). This value of 10,000 was actually recommended (by an action editor for *Psychophysiology*) to one of us (JPR) as proper for our P300-based CIT work. In contrast, we and others (e.g., Farwell & Donchin, 1991; Meijer, Smulders, Merckelbach, & Wolf, 2007) use only 100 iterations. Given the modern laptop's processing speed, there is little cost to running 10,000 and more iterations. We are more concerned with reestablishing the validity of the dozens of previous papers that have used 100 iterations. One size (number of iterations) does not fit all, and we would argue that 10,000 are not needed in our type of P300 bootstrap analyses (though it probably doesn't hurt to use such a large number). This will be illustrated in formal comparisons with real data sets below. We felt it important here to make two points that many people do not appear to appreciate. They are that (1) bootstrapping is not the same thing as permutations (in either its typical or anomalous forms), and (2) the number of resamplings/iterations required for a particular research question depends on the nature of that question, and this number is best found out at the onset of the research by comparing various numbers of iterations in their pilot subjects—as we do now in previously collected data sets. One of us (AW) had four data sets in immediately accessible files.

We note that a highly rigorous treatment of these issues is available (e.g., Andrews & Buchinsky, 2000). It is probably too technical for the typical reader of psychophysiology journals, which is why we provide here a more intuitive approach. Additionally,

**Table 1.** *Comparison of Bootstrapping Results for Different Numbers of Iterations*

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | Study 1, $n = 52$ | | | |
| P > I% 10^2 | P > I% 10^3 | P > I% 10^4 | PIDF10^2 | PIDF10^3 | PIDF10^4 |
| 95.7 | 95.7 | 95.6 | 10.2 uV | 9.8 uV | 10.2 uV |
| | P > I correl® matrix | | | PIDF correl matrix | |
| | R10^2 | R10^3 | | R10^2 | R10^3 |
| R10^2 | | | R10^2 | | |
| R10^3 | .987 | | R10^3 | .927 | |
| R10^4 | .988 | .998 | R10^4 | .998 | .930 |
| | | | Study 2, $n = 29$ | | |
| P > I% 10^2 | P > I% 10^3 | P > I% 10^4 | PIDF10^2 | PIDF10^3 | PIDF10^4 |
| 93.8 | 91.8 | 91.9 | 6.32 uV | 6.24 uV | 6.24 uV |
| | P > I correl® matrix | | | PIDF correl matrix | |
| | R10^2 | R10^3 | | R10^2 | R10^3 |
| R10^2 | | | R10^2 | | |
| R10^3 | .935 | | R10^3 | .995 | |
| R10^4 | .951 | .994 | R10^4 | .996 | .999 |
| | | | Study 3a, $n = 8$ | | |
| P > I% 10^2 | P > I% 10^3 | P > I% 10^4 | PIDF10^2 | PIDF10^3 | PIDF10^4 |
| 88.0 | 88.4 | 88.1 | 4.59 uV | 4.71 uV | 4.68 uV |
| | P > I correl® matrix | | | PIDF correl matrix | |
| | R10^2 | R10^3 | | R10^2 | R10^3 |
| R10^2 | | | R10^2 | | |
| R10^3 | .995 | | R10^3 | .997 | |
| R10^4 | .992 | .998 | R10^4 | .998 | .999 |
| | | | Study 3b, $n = 8$ | | |
| P > I% 10^2 | P > I% 10^3 | P > I% 10^4 | PIDF10^2 | PIDF10^3 | PIDF10^4 |
| 99.0 | 98.9 | 98.9 | 9.38 uV | 9.34 uV | 9.29 uV |
| | P > I correl® matrix | | | PIDF correl matrix | |
| | R10^2 | R10^3 | | R10^2 | R10^3 |
| R10^2 | | | R10^2 | | |
| R10^3 | .919 | | R10^3 | .998 | |
| R10^4 | .944 | .996 | R10^4 | .999 | .999 |

*Note*. Comparisons of iteration numbers (100 = 10^2 vs. 1,000 = 10^3 vs. 10,000) used in four studies with numbers of subjects shown. Also shown for each study are the cross correlation matrices (e.g., at intersections of R10^2 with R10^3 is the Pearson correlation of numbers based on 100 vs. 1,000 iterations). P > I% = percent of iterations in which the probe was greater than the irrelevant at a given iteration number test. PIDF = average P − I P300 difference estimated by the bootstrap.

although we will be speaking about distinguishing between probe and irrelevant P300 waveforms in guilty knowledge suspects, in a variety of other clinical diagnostic situations (Polich, 2004), the same form of diagnostic question is posed within one test patient/client: Is the rare target P300 larger than the frequent nontarget P300? The answer typically helps decide how to classify the patient—as having some disorder (e.g., ADHD, dementia, Alzheimer's, schizophrenia, and so on) or not. Note that probes and irrelevants can be thought of as special (forensic) cases of more general target and nontarget stimuli in the standard oddball paradigm (Donchin, 1981).

## Illustrative Data Sets

### Experiment 1

Experiment 1 (presently in review), $n = 52$, is a study of possible suppression of P300 evoked by semantic stimuli in suppression and nonsuppression groups. A detailed description of the suppression manipulation is found in Hu, Bergström, Bodenhausen, & Rosenfeld (2015), where a knowledgeable group whose participants performed a mock crime, but were instructed to suppress their episodic probe memories, was compared to a simply knowledgeable (guilty of a mock crime) group. The same manipulation was used in the study of semantic probe memory suppression that we reanalyze here in this (Experiment 1) study. There were no differences in this study between suppression versus no-suppression groups in amplitude or latency, so they were pooled for the present reanalysis. In the original study, we used 100 iterations in each subject to determine if the probe P300 > irrelevant P300. For the present study, mean numbers of bootstrapped iterations (within each of 52 subjects) in which bootstrapped probe P300 exceeded irrelevant P300 (P > I) in 100 versus 1,000 versus10,000 iterations are shown in Table 1 (Study 1). Also shown are average P − I P300 differences (in uV) estimated by the bootstraps. The two correlation matrices among the three possible pairs (for each of the two variables) of correlation coefficients are also shown in Table 1. Clearly, one sees that 100, 1,000, and 10,000 iterations produce similar results, such that the intercorrelations of results are high.

### Experiment 2

Experiment 2 (submitted 2016), $n = 29$, is a new study of possible suppression of P300 evoked by episodic memory stimuli in suppression and nonsuppression groups. Again, there were no differences between groups in amplitude or latency, so they were pooled. The results in Table 1 are presented the same way as for Experiment 1. Again, one sees that 100, 1,000, and 10,000 iterations produce similar results, such that the intercorrelations of results are high.

**Table 2.** *Comparison of Bootstrapping Results for Different Numbers of Iterations*

| | | | | | | |
|---|---|---|---|---|---|---|
| Study 1, $n = 24$ | | | | | | |
| P > I% 10^2 | P > I% 10^3 | P > I% 10^4 | | PIDF10^2 | PIDF10^3 | PIDF10^4 |
| 98.5 | 98.5 | 98.5 | | 8.2 uV | 8.2 uV | 8.2 uV |
| | P > I correl® matrix | | | | PIDF correl matrix | |
| | R10^2 | R10^3 | | | R10^2 | R10^3 |
| R10^2 | | | | R10^2 | | |
| R10^3 | .999 | | | R10^3 | .999 | |
| R10^4 | .999 | 1.000 | | R10^4 | .999 | 1.000 |
| BSCOR% 10^2 | | | BSCOR% 10^3 | | BSCOR% 10^4 | |
| 97.2 | | | 97.4 | | 97.5 | |
| | BSCOR correl matrix | | | | | |
| | R10^2 | R10^3 | | | | |
| R10^2 | | | | | | |
| R10^3 | .942 | | | | | |
| R10^4 | 0.951 | 0.996 | | | | |
| Study 2, $n = 24$ | | | | | | |
| P > I% 10^2 | P > I% 10^3 | P > I% 10^4 | | PIDF10^2 | PIDF10^3 | PIDF10^4 |
| 70.9 | 70.4 | 70.6 | | 1.6 uV | 1.5 uV | 1.5 uV |
| | P > I correl® matrix | | | | PIDF correl matrix | |
| | R10^2 | R10^3 | | | R10^2 | R10^3 |
| R10^2 | | | | R10^2 | | |
| R10^3 | .993 | | | R10^3 | .997 | |
| R10^4 | .995 | .999 | | R10^4 | .998 | 1.000 |
| BSCOR% 10^2 | | | BSCOR% 10^3 | | BSCOR% 10^4 | |
| 24.9 | | | 26.1 | | 25.9 | |
| | BSCOR correl matrix | | | | | |
| | R10^2 | R10^3 | | | | |
| R10^2 | | | | | | |
| R10^3 | .993 | | | | | |
| R10^4 | .994 | .999 | | | | |
| Chapter 5, $n = 30$ | | | | | | |
| P > I% 10^2 | P > I% 10^3 | P > I% 10^4 | | PIDF10^2 | PIDF10^3 | PIDF10^4 |
| 91.9 | 92.4 | 92.4 | | 4.5 uV | 4.5 uV | 4.6 uV |
| | P > I correl® matrix | | | | PIDF correl matrix | |
| | R10^2 | R10^3 | | | R10^2 | R10^3 |
| R10^2 | | | | R10^2 | | |
| R10^3 | .998 | | | R10^3 | .998 | |
| R10^4 | .997 | 1.000 | | R10^4 | .998 | 1.000 |
| BSCOR% 10^2 | | | BSCOR% 10^3 | | BSCOR% 10^4 | |
| 82.3 | | | 83.4 | | 83.3 | |
| | BSCOR correl matrix | | | | | |
| | R10^2 | R10^3 | | | | |
| R10^2 | | | | | | |
| R10^3 | .991 | | | | | |
| R10^4 | .993 | .999 | | | | |

*Note*. Comparisons of iteration numbers (100 = 10^2 vs. 1,000 = 10^3 vs. 10,000 = 10^4) used in three studies that also included target stimuli. P > I% and PIDF are the same as in Table 1. BSCOR = number of iterations in which the probe-target correlation exceeded probe-irrelevant correlation.

### Experiment 3a and 3b

These were from a pilot study (from a laboratory course with AW as instructor) for Rosenfeld, Ward, Frigo, Drapekin, and Labkovsky (2015), with $n = 8$ in each of two conditions, with auditory (3a) versus visual (3b) probe stimuli, respectively. Results are reported in Table 1 the same way as for Experiment 1 and 2. Again, one sees that 100, 1,000, and 10,000 iterations produce similar results, such that the intercorrelations of results are high.

### Experiment 4

Data are from Experiment 1 of Meijer et al. (2007). Twenty-four participants were presented with six pictures of faces. Two of these were familiar (brother or sister and best friend), two were not. One of the familiar pictures served as a target stimulus (i.e., participants were instructed to press one of two buttons upon presentation of this face). The other button was pressed upon presentation of all other faces including the other recognized face (probe). Results are shown in Table 2 (Study 1). In addition to the probe-irrelevant difference, we also compared the correlation of probe and target with that of probe and irrelevant (Farwell & Donchin, 1991). BSCOR refers to the number of iterations in which the probe-target correlation exceeded the probe-irrelevant correlation. That is, in the previous discussion, for each iteration, each bootstrapped average probe ERP and each bootstrapped average irrelevant ERP had P300 amplitudes computed, and then the P − I P300 amplitude difference was placed into a distribution of differences. It was required

that 90% or more of these differences were greater than zero for a knowledgeable decision. In the BSCOR computation, the P300 elicited by a third stimulus, the target, as described above, becomes important. As noted before, targets are rare irrelevant stimuli to which a unique button press is required. They are rare ($p < .2$, usually) and meaningful due to their unique response requirement. Therefore, they too should evoke P300 that can be used as a template or benchmark P300 (but see Rosenfeld, 2011). Thus, knowledgeable individuals should show P300s to probes and targets, but not-knowledgeable people should show P300s only to targets. Thus, one expects the probe-target cross-correlation (Rpt) over most of the entire sweep to exceed that of the probe-irrelevant cross-correlation (Rpi) for knowledgeable persons, but the reverse relation for not-knowledgeable persons. The bootstrap test based on this approach computes these two cross-correlations on each iteration of bootstrapped ERPs, and the knowledgeability criterion is that Rpt must exceed Rpi on 90% of the iterations.

### Experiment 5

Data are from Experiment 2 of Meijer et al. (2007). Design was similar but familiar pictures represented teachers of courses the participants had completed. Results are given in Table 2 (Study 2).

### Experiment 6

Data from Chapter 5 of Meijer (2008). Thirty participants committed a mock crime, and then completed a P300-based CIT with questions relating to the mock crime details. Results are presented in Table 2 (Chapter 5).

## Results and Discussion

It is evident from these tables that values obtained in all data sets do not appreciably differ among bootstrap tests with 100 versus 1,000 versus 10,000 iterations, and that, indeed, the tables of correlations show high values, all exceeding 0.9. To validate these impressions, we did statistical group hypothesis tests on data from Experiment 1 only, which had the largest number of subjects. The independent variable was iteration number (100 vs. 1,000 vs. 10,000) regarding either numbers of P > I% outcomes or estimated probe-irrelevant difference (DF) based on those iteration numbers. A $1 \times 3$ repeated measures analysis of variance (ANOVA) with iteration number as independent variable yielded $F(2,102) = .05$, $p > .94$ for P > I%. The scaled JZS Bayes factor favoring the null hypothesis was a strong 6.45. A follow-up paired $t$ test testing P > I% effect of iteration number in 100 versus 10,000 iterations (suggested by the above introductory material) yielded $t(51) = .233$, $p > .81$. The scaled JZS Bayes factor favoring the null hypothesis was also 6.45.

The same two tests were applied to the estimated difference (in uV) between P and I bootstrapped P300s. The ANOVA yielded $F(2,102) = 1.89$, $p > .15$. The scaled JZS Bayes factor favoring the null hypothesis was 2.71. The paired $t$ test comparing 100 versus 10,000 iterations yielded $t(51) = .704$, $p > .48$. The scaled JZS Bayes factor favoring the null hypothesis was also 2.71.

In sum, across all seven of our datasets, 100 iterations seem sufficiently stable, and adding more iterations added relatively little information. This leads us to conclude that in typical P300-based CIT research—where the goal is to classify knowledgeable and not-knowledgeable participants, based on probe-minus-irrelevant P300 differences—one does not require more than 100 bootstrap iterations.

## References


Andrews, D. W. K., & Buchinsky, M. (2000). A three-step method for choosing the number of bootstrap repetitions. *Econometrica*, *68*, 23–51. doi: 10.1111/1468-0262.00092

Blair, R. C., & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, *30*, 518–524. doi: 10.1111/j.1469-8986.1993.tb02075.x

Di Nocera, F., & Ferlazzo, F. (2000). Resampling approach to statistical inference: Bootstrapping from event-related potentials data. *Behavior Research Methods, Instruments, & Computers*, *32*(1), 111–119. doi: 10.3758/BF03200793

Donchin, E. (1981). Surprise! ... surprise? *Psychophysiology*, *18*, 493–513. doi: 10.1111/j.1469-8986.1981.tb01815.x

Efron, B. (1979). Bootstrap methods: Another look at the jackknife *Annals of Statistics*, *7*, 1–26. doi: 10.1007/978-1-4612-4380-9_41

Efron, B., & Tibshirani, R. (1994) *An introduction to the bootstrap*. Boca Raton, FL: Taylor & Francis. doi: 10.2307/2532810

Ellwanger, J., Rosenfeld, J. P., Sweet, J. J., & Bhatt, M. (1996). Detecting simulated amnesia for autobiographical and recently learned information using the P300 event-related potential. *International Journal of Psychophysiology*, *23*(1), 9–23. doi: 10.1016/0167-8760(96)00035-9

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology*, *28*(5), 531–547. doi: 10.1111/j.1469-8986.1991.tb01990.x

Hu, X., Bergström, Z. M., Bodenhausen, G. V., & Rosenfeld, J. P. (2015). Suppressing unwanted autobiographical memories reduces their automatic influences: Evidence from electrophysiology and an implicit autobiographical memory test. *Psychological Science*. doi: 10.1177/0956797615575734

Karis, D., Fabiani, M., & Donchin, E. (1984). "P300" and memory: Individual differences in the von Restorff effect. *Cognitive Psychology*, *16*(2), 177–216. doi: 10.1016/0010-0285(84)90007-0

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385–388. doi: 10.1037/h0046060

Meijer, E. H. (2008). *Psychophysiology and the detection of deception: Promises and perils* (Doctoral dissertation, Maastricht University). doi: 10.1016/j.ijpsycho.2012.06.047

Meijer, E. H., Smulders, F. T., Merckelbach, H. L., & Wolf, A. G. (2007). The P300 is sensitive to concealed face recognition. *International Journal of Psychophysiology*, *66*(3), 231–237. doi: 10.1016/j.ijpsycho.2007.08.001

Polich J. (2004). Clinical application of the P300 event-related potential. *Physical Medicine and Rehabilitation Clinics*, *15*, 133–161. doi: 10.1016/s1047-9651(03)00109-8

Rosenfeld, J. P. (2011). P300 in detecting concealed information. In B. Verschuere, G. Ben Shakhar, & E. Meijer, (Eds.) *Memory detection: Theory and application of the concealed information test* (pp. 63–89). Cambridge, UK: Cambridge University Press. doi: 10.1017/cbo9780511975196.005

Rosenfeld, J. P., Angell, A., Johnson, M., & Qian, J. H. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average waveforms. *Psychophysiology*, *28*(3), 319–335. doi: 10.1111/j.1469-8986.1991.tb02202.x

Rosenfeld, J. P., & Donchin, E. (2015). Resampling (bootstrapping) the mean: A definite do. *Psychophysiology*, *52*(7), 969–972. doi: 10.1111/psyp.12421

Rosenfeld, J. P., Hu, X., Labkovsky, E., Meixner, J., & Winograd, M. R. (2013). Review of recent studies and issues regarding the P300-based complex trial protocol for detection of concealed information. *International Journal of Psychophysiology*, *90*, 118–134. doi: 10.1016/j.ijpsycho.2013.08.012

Rosenfeld, J. P., Shue, E., & Singer, E. (2007). Single versus multiple probe blocks of P300-based concealed information tests for self-referring versus incidentally obtained information. *Biological Psychology*, *74*(3), 396–404. doi: 10.1016/j.biopsycho.2006.10.002

Rosenfeld, J. P., Sweet, J. J., Chuang, J., Ellwanger, J., & Song, L. (1996). Detection of simulated malingering using forced choice recognition enhanced with event-related potential recording. *Clinical Neuropsychologist*, *10*(2), 163–179. doi: 10.1080/13854049608406678

Rosenfeld, J. P., Ward, A., Frigo, V., Drapekin, J., & Labkovsky, E. (2015). Evidence suggesting superiority of visual (verbal) vs. auditory test presentation modality in the P300-based, complex trial

protocol for concealed autobiographical memory detection. *International Journal of Psychophysiology*, *96*(1), 16–22. doi: 10.1016/j.ijpsycho.2015.02.026

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology*, *26*(2), 208–221. doi: 10.1111/j.1469-8986.1989.tb03159.x

Zoumpoulaki, A., Alsufyani, A., & Bowman, H. (2015). Resampling the peak, some dos and don'ts. *Psychophysiology*, *52*, 444–448. doi: 10.1111/psyp.12363

(Received June 24, 2016; Accepted October 11, 2016)