BOOTSTRAPPING THE P300 IN APPLIED PSYCHOPHYSIOLOGY: EVALUATING PRECISION IN DIAGNOSTIC TESTS

Joseph Olson[1], Gayathri Subramanian[1], Jerzy Wojciechowski[2], Celine Bitegeko[1], and J. Peter Rosenfeld[1]

[1]Northwestern University
[2]University of Warsaw

May 22, 2023

Abstract

Background: In applied psychophysiology, bootstrapping procedures are often used to classify individuals into one of two or more independent states (e.g., high risk vs low risk). Although the number of iterations required for a reliable bootstrap test is not universally agreed upon, some research (Rosenfeld et al., 2017b) suggests that 100 iterations is a sufficient number to obtain reliable results when analyzing P300 from a concealed information test. However, no study to-date has evaluated the diagnostic consistency of the 100 iterations test across repeated examinations. Methods: We evaluated the precision of the 100 iteration test by repeating the test 100 times per participant in a sample of 81 participants. The test was designed to classify participants as either knowledgeable or not knowledgeable of critical information related to a mock crime. Results: We found that the test provided variable classifications in approximately a quarter of our sample (n = 19/81 or 23%), specifically when a participant's score presented near the diagnostic cutpoint. Moreover, the test's diagnostic results varied by as much as +/-15%, in certain cases. Conclusion: Although the test provided reliable results for the majority of our sample, this was not true for a notable number of cases. We recommend that researchers report the variability of their diagnostic metrics and integrate this variability when classifying individuals. We discuss several simple examples of how to take variability into account when making classifications, such as by calculating the probability of one classification state over another given the data.

BOOTSTRAPPING THE P300 IN APPLIED PSYCHOPHYSIOLOGY: EVALUATING PRECISION IN DIAGNOSTIC TESTS

Olson, J. M[1]., Subramanian, G[1]., Wojciechowski, J[2]., Bitegeko, C[1]., Rosenfeld, J. P[1].

[1]Northwestern University, Department of Psychology

[2]University of Warsaw, Faculty of Psychology

Author Note

Joseph M. Olson, ORCID ID:*https://orcid.org/0000-0002-9414-1941*

Gayathri Subramanian, ORCID ID:*https://orcid.org/0009-0009-5651-5402*

Jerzy Wojciechowski, ORCID ID:*https://orcid.org/0000-0001-5188-2590*

The authors have no known conflict of interest to declare.

Correspondence regarding this article should be addressed to:

joeolson@u.northwestern.edu

Abstract

Background: In applied psychophysiology, bootstrapping procedures are often used to classify individuals into one of two or more independent states (e.g., high risk vs low risk). Although the number of iterations required for a reliable bootstrap test is not universally agreed upon, some research (Rosenfeld et al., 2017b) suggests that 100 iterations is a sufficient number to obtain reliable results when analyzing P300 from a concealed information test. However, no study to-date has evaluated the diagnostic consistency of the 100 iterations test across repeated examinations.

Methods: We evaluated the precision of the 100 iteration test by repeating the test 100 times per participant in a sample of 81 participants. The test was designed to classify participants as either knowledgeable or not knowledgeable of critical information related to a mock crime.

Results: We found that the test provided variable classifications in approximately a quarter of our sample (n = 19/81 or 23%), specifically when a participant's score presented near the diagnostic cutpoint. Moreover, the test's diagnostic results varied by as much as +/-15%, in certain cases.

Conclusion: Although the test provided reliable results for the majority of our sample, this was not true for a notable number of cases. We recommend that researchers report the variability of their diagnostic metrics and integrate this variability when classifying individuals. We discuss several simple examples of how to take variability into account when making classifications, such as by calculating the probability of one classification state over another given the data.

Keywords: Bootstrapping, Diagnostics, Precision, Concealed Information Test, P300, Complex Trial Protocol

Introduction

Bootstrapping in Psychophysiology

Bootstrapping is a data-based simulation technique for making statistical inferences (Efron & Tibshirani, 1994). It is especially useful for estimating the repeatability and reliability of a statistical result. Over 30 years ago, Wasserman and Bockenholt (1989) introduced bootstrapping to the field of psychophysiology. It has been shown to be useful in a variety of applications, including research on event-related potentials (ERP) such as the P300 waveform (e.g., Fabbiani et al., 1998). More recently, it has been used to interrogate psychophysiological methods, such as computing standard measurement errors to obtain a universal metric of ERP data quality (Luck et. al. 2021).

Although bootstrapping is used in a variety of fields, it is commonly utilized in the detection of concealed information. For instance, the P300-based Concealed Information Test (P300-CIT), has been studied for over 30 years and reliably uses the bootstrapped P300 to detect concealed knowledge of privileged, typically crime-related, information (Rosenfeld, 2020). In the P300-CIT, a series of *crime relevant*(Probe) and *crime irrelevant* (Irrelevant) stimuli are presented to a participant, and based on their responses to these stimuli one can classify them as being knowledgeable or unknowledgeable of crime relevant information. Another stimulus called a Target is used to maintain the attention of the participants by asking them to produce a unique response (e.g., on a keyboard) everytime they see it (Rosenfeld et al., 2006; Gamer & Berti, 2012). Guilty/knowledgeable individuals will have a differential psychophysiological response to crime relevant information compared to crime irrelevant information. For instance, the probe, which is crime relevant, would be salient only to someone who is "guilty" resulting in a larger P300 for probes in comparison to irrelevants. For this reason, the probe-irrelevant difference (i.e., "CIT effect") is used to determine if the individual is "Guilty" (i.e., knowledgeable) or "Innocent" (i.e., unknowledgeable) (Rosenfeld, 2020).

Bootstrapping the P300 ERP

In the field of deception detection, we are expected to classify an individual as guilty/knowledgeable or innocent/unknowledgeable, and hence analyze the data within-subjects (specifically, within an*individual*). This analysis would use single trial EEG data, which is often noisy and variable, leading to unreliable classification decisions. Therefore, in these situations it is especially useful to use simulation techniques like

bootstrapping to aid intraindividual diagnosis (Wasserman & Bockenholt, 1989). The bootstrapping process involves selecting a set of single sweep ERP data points (with replacement) and averaging them to produce a single ERP amplitude per individual. Since the resampling of data points is done with replacement, the averages produced in each run will be slightly different from other runs. Within a subject, we can bootstrap with replacement a sample of probe trials and calculate the average, and similarly produce an average using irrelevant trials. Then, we compare the difference between resampled probes and irrelevant averages to estimate the CIT effect. In iterating this process n times, one can estimate out of all iterations how many times the probe P300 average exceeded the irrelevant P300 average, enabling investigators to create a diagnostic index for classifying the subject as guilty/knowledgeable or innocent/unknowledgeable. This diagnostic metric is known as the bootstrap iteration score (BSITER) (e.g., Rosenfeld et al., 2015; Rosenfeld & Donchin, 2015), and has been published in psychophysiological experiments for several decades.

Disagreement about iterations

The number of iterations in a bootstrap test refers to the number (n) of times we resample with replacement to calculate the bootstrap average for each individual. The number of iterations can be anything ranging from 10 to 10,000 or more, but generally speaking a larger number of iterations produces more reliable results. Rosenfeld et al. (2017b) provided evidence that a relatively small number of iterations (e.g., 100 iterations) sufficiently yields accurate diagnoses when bootstrapping the P300 ERP. The authors concluded that a smaller number of iterations was effective because the P300 is a robust ERP with a large effect size. However, they also noted that this may not be the case in experiments using other ERPs with smaller effect sizes (e.g., N400). So, Rosenfeld et al., (2017b), suggests that the number of iterations may vary across different applications and may not be "one size fits all". Understandably, concerns have been presented around whether 100 iterations are truly sufficient for accurate diagnosis, even when using P300 (e.g., Zoumpalaki et al., 2015). Although the evidence provided in Rosenfeld et al., (2017b) was suggestive, large correlations between low vs high iteration tests is not highly robust evidence of adequate precision in diagnostic tests. In our view, the concern around a low iteration bootstrap test regards the reliability and repeatability of the classification results of the test, and the most direct way to evaluate this is to simply repeat the test many times, producing many diagnostic results per individual. Fortunately, repeating a bootstrap-based test can be done easily and requires only an investment in time and computation resources. Moreover, we believe that there are several statistical, methodological and diagnostic advantages for repeating a bootstrap test many times, which is the focus of this paper.

Our approach

In this paper, we discuss a method of "repeated bootstrapping" (rBS) where a number of iterations is chosen, and the resampling is repeated a specified number of times. In the traditional bootstrapping method, a diagnosis is made after a specified number of iterations (e.g. 100 iterations). However, if the test were to be repeated, the same diagnosis may not be recommended given the random sampling involved in the bootstrapping procedure11[Note that this limitation is also relevant for any number of diagnostic tests that are performed on the same individual at different points in]
When considering the efficacy of a diagnostic test it is vital to report its precision, or how similar the results of the test are upon re-testing. To our knowledge, no previous study has evaluated the repeatability of the 100 iterations bootstrap test in the P300-based concealed information detection literature, so that is what we aim to do here. Additionally, the rBS technique could be a useful tool to calculate confidence intervals for diagnostic metrics like the BSITER score, improving upon standard methods in our field similar to ERP data quality metrics such as the one recently proposed by Luck et al. (2021).

Methods

Studies Included in the Analysis

For the purpose of the planned analyses, and to minimize potential confound effects, all the data used in the main analyses were recorded in the Rosenfeld Lab, with the same hardware and with the same version of P300-CIT – the Complex Trial Protocol, which is the most modern and countermeasure-resistant version

of the P300-based CIT (Rosenfeld et.al., 2004; Rosenfeld et. al. 2008). Since 2008, many studies have demonstrated high sensitivity and specificity in the CTP (over 90% for autobiographical information and over 80% for incidentally acquired information; with the AUC typically over .9) (see Rosenfeld et.al, 2013 for review).

In the Complex Trial Protocol (CTP), contrary to the traditional three-stimulus protocol, the probe, irrelevant and target are separated into two parts (Rosenfeld, et.al., 2008). In the first phase, probes and irrelevants are presented and the participants are asked to make a button response to indicate that they saw the stimulus (same button for both stimuli). In the second phase, targets and non-targets are presented during which the participants are instructed to respond differently to targets compared to non-targets. These targets and non-targets are usually a series of digits that are not relevant to the first phase (e.g., "11111" is the target, and "22222", "33333", "44444","55555", "66666" are non-targets), such that more cognitive resources can be dedicated to phase 1, where the diagnostic analysis takes place. This protocol is to be used as an example throughout the paper. So, data collected from the CTP can further be analyzed through statistical tools like Bootstrapping to make classification results more accurate and effective.

Critically, we reanalyzed the data from Rosenfeld et al. (2017b) using our rBS technique, specifically regarding the following two studies:

1) Rosenfeld, et.al (2017a) (Experiment 2 - n=52 guilty participants with semantic stimuli) – this study explored the possibility of a memory suppression effect on P300 amplitude while using the Complex Trail Protocol with semantic stimuli. No differences in amplitude or latency were found between suppression versus non-suppression groups, so all participants were pooled in Rosenfeld et al. (2017b). Consequently, we also merged suppression and non-suppression groups in the current analysis. 2) Ward, Rosenfeld (2017) (n=29 guilty participants with episodic stimuli) – this experiment also verified the memory suppression effect on P300 amplitude in the CTP, however with episodic memory. Since there were no differences in P300 amplitude or latency between experimental groups, data were pooled in Rosenfeld et al. (2017b) and also in the current analysis.

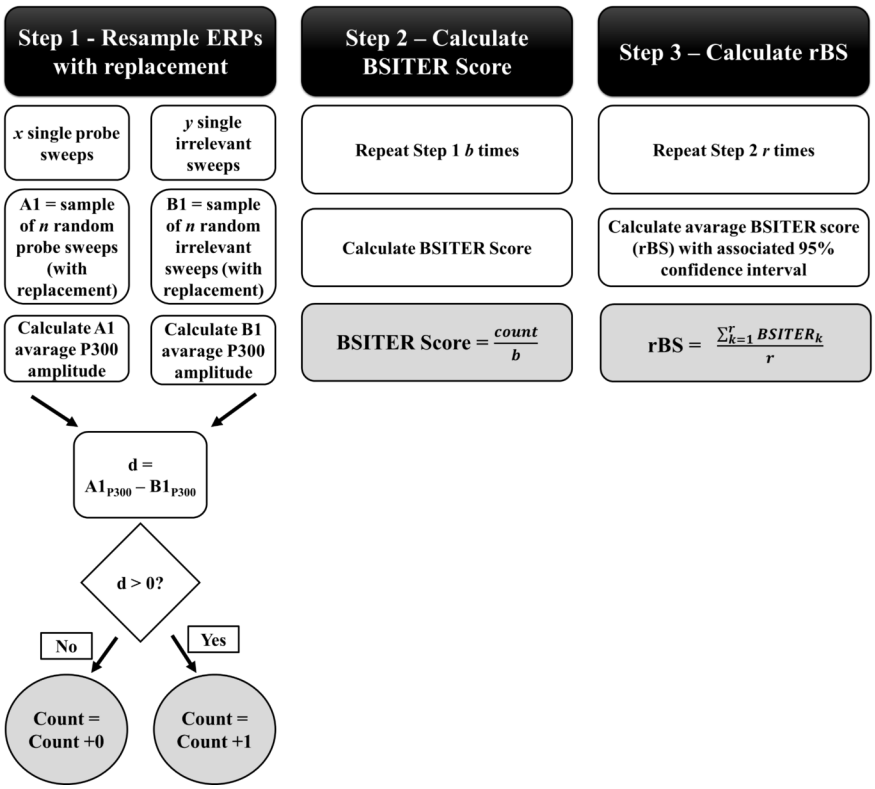
Please refer to Rosenfeld et al., (2017a) and Ward & Rosenfeld (2017) for details on participant recruitment and EEG/ERP data processing methods.

The bootstrapping and repeated bootstrapping procedure

As was explained in the Introduction, bootstrapping P300 is a technique for randomly sampling (with replacement) ERPs from a single subject's dataset. In the studies concerning detection of concealed information with the CTP, a single dataset for one participant usually consists of ˜30 probe sweeps (i.e., single trials) and ˜150 irrelevant sweeps (e.g., 5 irrelevants presented 30 times each). Each resampling (bootstrap iteration) is based on the same set of sweeps (e.g., 180 sweeps total). Because the sampling with replacement method is used, the probability that subsequent samples of sweeps will be identical to previous samples is reasonably low.

In each iteration, two sets of sweeps are sampled with replacement from whole datasets, separately for probes and irrelevants (see Figure 1). Then, an average probe ERP and an average irrelevants ERP is computed. Next P300 amplitude values are calculated. We used the peak-to-peak (p-p) method, which has been shown to yield the highest classification performance in P300-based CITs (Soskins, Rosenfeld & Niendam, 2001). The methods for the p-p analysis are described in detail in several recent reports (Olson, Rosenfeld, Perrault, 2019; Lukács et al., 2016). Importantly, the p-p method derives P300 amplitude using averages across time windows, *not* peak amplitude values that are easily susceptible to variation and error. When the Probe (P) and Irrelevants (I) P300 amplitudes are determined, the P-I difference (d) is calculated and is marked as 1 if P-I $>$ 0 or 0 if P-I $<=$ 0. The whole process is repeated b times (usually 100) and provides the researcher with the proportion of times the P-I differences were larger than 0 (see Figure 1, steps 1-2). For example, such an analysis might result in BSITER score = 92, which indicates that in 92/100 of comparisons the Probe amplitude was larger than the Irrelevants amplitude. Such a result in concealed information studies is usually interpreted as a sign of "guilt", or recognition of the probe item.

In the repeated bootstrap procedure (rBS) the calculation of a BSITER score is performed r times (see Figure 1, step 3) which provides new information about the consistency of results and their distribution, among other things. Figure 1 presents the whole procedure graphically.



Note: n is equal to the number of probe sweeps in the analyzed data set (usually around 30-40 sweeps).

Results

Summary and Aim of the Paper

Bootstrap iteration (BSITER) scores are often used in the P300 CIT literature to make classification decisions about individual participants. The BSITER score is typically derived using 100 or 1,000 iterations. Although random sampling is inherent to the calculation of these scores, it is generally assumed that this variability does not meaningfully impact classification performance. However, this assumption has never been tested, to our knowledge, particularly regarding the 100 iterations version of the bootstrap test. Although Rosenfeld et al., (2017b) observed robust correlations between the 100, 1,000 and 10,000 iterations tests, which suggests high precision across the majority of subjects, a correlation approach is not a thorough evaluation of the internal validity of the 100 iterations test. Therefore, we opted for a more rigorous alternative to evaluate the precision of the 100 iterations bootstrap test in the current report.

To evaluate the precision of the 100 iterations test, we repeated it 100 times per participant. Because each test yields a BSITER score, our procedure (henceforth the "repetition bootstrap" or rBS) produced 100 scores per participant, allowing us to observe and report their variability within each individual subject.

The Precision of the 100 Iterations Bootstrap Test on Different Memory Types

The rBS test was conducted independently for each subject and the resulting 100 BSITER scores were then used to calculate each of the following measures per participant: a mean BSITER score and associated

standard error, 95% confidence interval (CI), and range (see Table 3). Note that these measures of variability represent how participants' BSITER scores changed due to random sampling alone. The scores in parentheses are the between-subjects standard deviations for each measure. Importantly, because differences in BSITER scores tend to be observed between memory types (e.g., Olson et al., 2022), we conducted this procedure separately for Semantic and Episodic memory information (see Table 3).

Table 1

The mean repetition bootstrap scores (rBS) and their associated standard error, 95% confidence interval (CI), and range for Episodic and Semantic groups.

Group	*Mean rBS Score*	*Standard Error*	*95% CI*	*Range*
Semantic (n=52)	95% (7.77%)	1.32% (1.46%)	2.63% (2.91%)	8% (8.73%)
Episodic (n=29)	90% (12.55%)	2.44% (1.29%)	4.88% (2.58%)	15% (7.74%)

As is typical of guilty/knowledgeable populations in the P300 CIT literature, the mean BSITER score for both memory groups was at least 90%. Importantly, we observed that the standard error and respective confidence intervals for these scores varied to a small degree (within-subjects), which could result in inconsistent diagnostic or classification performance. We also noted that the within-subjects variability of the Episodic group was nearly twice that of the Semantic group (trending towards significance at $p<0.1$), suggesting that a participant's BSITER score may be more variable (generally speaking) when they are derived from less familiar Episodic information. The following sections detail the results for each memory type separately.

Semantic Memory

The results in Table 3 suggest that a knowledgeable/guilty individual tested on highly familiar, Semantic information with an estimated BSITER score of 95% may in fact present anywhere between 92.37% – 97.63%, due to random sampling. In very rare cases, a subject's BSITER score may vary as much as +/-8% (87% – 100%). Although this level of variability is likely not concerning for participants with large BSITER scores (e.g., 95-100%), it could result in diagnostic complications for participants who scored near the 90% threshold, which we examine next.

To be confident that a participant's BSITER score is reliably above or below a diagnostic threshold it is intuitive to expect that the 95% confidence interval for their score must not overlap with the threshold. If the 95% CI overlaps with the diagnostic threshold it follows that we cannot have at least 95% confidence in that participant's diagnosis. Therefore, we tallied how many participants in the Semantic group had mean BSITER scores with 95% CIs that overlapped with the 90% diagnostic cutpoint.

Following this analysis, we noted that at least 95% confidence was obtained in the majority of Semantic participants' diagnoses (41/52 or 78.85%). However, we did not have 95% statistical confidence in nearly a quarter of the sample's BSITER scores (11/52 or 21.15%). Notably, all 11 participants had scores $<$ 95%. Importantly, 7/11 of these participants also had scores >90%, meaning they were correctly classified as "guilty/knowledgeable" (true positives, since these subjects were knowledgeable of crime-relevant information), but the guilty/knowledgeable classification was not made with 95% statistical confidence.

To summarize, 100 iterations appears to yield consistent and reliable results for the majority of our Semantic sample; however, 100 iterations was insufficient to produce precise results in nearly a quarter of participants (21.15%). For these participants, conducting the 100 iterations test once (as is traditionally done) could result in a score either above or below the 90% threshold depending only upon chance resampling in the derivation of their BSITER scores. In other words, although these classifications were technically correct, the classification decisions were made absent the level of statistical rigor we and others advocate for in diagnostic psychophysiology. Consequently, we recommend A) increasing the number of iterations (e.g,. to 10,000) when deriving the BSITER score in order to reduce the variability of the score, and/or B) using the

rBS test to calculate, report, and integrate the 95% CI for participant's BSITER score during individual classification (we will explore two detailed examples of this in the Discussion).

Episodic Memory

Regarding the 100 iterations bootstrap test's performance on less familiar, Episodic information, we found that participants with an estimated BSITER score of 90% may present anywhere between 85.12% – 94.88%. In rare cases, the score may vary as much as +/-15% (75% – 100%). Once again, although participants with very large BSITER scores may be unlikely to encounter issues with diagnostic precision, participants who scored near the 85% threshold may be at risk for inconsistent classification.

Therefore, we tallied how many participants in the Episodic group had mean BSITER scores with 95% CIs that overlapped with the 85% cutpoint. Again, we found that 100 iterations produced reliable diagnoses for the majority of participants (21/29 or 72.41%); however, we had <95% confidence in over a quarter of the sample's diagnoses (8/29 or 27.59%). All 8 of these participants had BSITER scores <= 90%. Importantly, 7/8 of these participants had BSITER scores equal to or exceeding the 85% threshold, meaning they were correctly classified as "guilty/knowledgeable", but not with 95% confidence.

In conclusion, we found that the 100 iterations bootstrap test provided consistent and reliable results for the majority of our Episodic and Semantic groups (n = 62/81 or 77%); however, participants whose BSITER scores occurred near-threshold (approximately a quarter of our sample), were not classified with at least 95% statistical certainty. We recommend increasing the number of iterations used in the bootstrap test and/or calculating, reporting, and integrating the 95% CI for a participant's BSITER score during individual classification analyses. We believe these steps are both important and necessary for improving statistical rigor in our field.

Discussion

Summary of study concept and major results

This paper aimed to carefully evaluate the 100 iterations bootstrap test, which is a statistical resampling technique commonly used in the P300 concealed information detection literature. Although the accuracy of the 100 iterations test is debated, Rosenfeld et al., (2017b) argued that the test is equally reliable to its more rigorous 1,000 and 10,000 iterations peers. Although these results are compelling, they do not rigorously describe the precision of the 100 iterations test, which we aimed to do here. Thus, the current report is the first and only to describe and evaluate the precision of the classification results of the 100 iterations bootstrap test.

We felt the most intuitive way to interrogate the precision of the 100 iteration bootstrap test was to simply repeat the test many times. Therefore, we repeated the 100 iterations test 100 times using a technique we call the repeated bootstrap (i.e., rBS), which is the mathematical equivalent to using 10,000 iterations. In this paper, we used the rBS technique on amplitude values of the P300 ERP from a concealed information test (i.e., the CTP) - critically, the same knowledgeable/guilty participants analyzed in Rosenfeld et al., (2017b). In diagnosing a participant as knowledgeable/guilty or unknowledgeable/innocent, the bootstrap iteration (i.e., BSITER) score is compared to a threshold to determine the diagnosis. In our analyses, we used data from two CTP studies: one using semantic or highly salient information like birth dates, and another using less salient, episodic information like knowledge acquired during committing a mock crime. In the semantic group, we found that, in rare cases, an individual's BSITER score may vary by +/-8 %, and the episodic group was nearly twice as variable at +/-15%. This result was not surprising given that highly rehearsed semantic information is more memorable than episodic information, so P300s are generally smaller and more variable from trial to trial in the episodic protocol (Olson et al., 2020). Fortunately, we find it unlikely that the variability observed would affect the diagnosis of individuals so long as their BSITER score was quite large (i.e., >= 95%). However, for participants with lower scores (e.g., near the diagnostic threshold) this variability could produce unreliable or inconsistent diagnoses. The intraindividual variability of BSITER scores has never previously been analyzed or even reported, to our knowledge, and based on these results we

suggest researchers use rBS (or a similar technique) to report the variability of the BSITER score within a subject, as we have done here.

Implications for researchers and practitioners

Diagnostic analyses are notoriously difficult, and deriving accurate and representative diagnostic metrics is important for both research purposes and an individual's life (e.g., health, penal consequences, social standing/reputation, etc). Bootstrapping is a useful tool for evaluating the precision of a diagnostic test without the cost associated with repeating the physical test multiple times per patient. We have found that although the 100 iterations bootstrap test yielded statistically reliable classifications for the majority of our sample, it provided inconsistent diagnostic results in approximately a quarter of our participants (n = 19/81 or 23%), particularly when a participant's score occurred near the diagnostic cutpoint. For this reason, increasing the number of iterations performed (e.g., to 10,000) is advisable. However, we note that increasing the number of iterations performed does not guarantee reliable classification, nor does it facilitate the reporting of the intraindividual variability of the BSITER score itself - and thus the stability or reliability of a participant's classification. We believe that describing the intraindividual variability of the BSITER score (or any diagnostic metric) is important to understanding and justifying classification decisions. Therefore, we suggest that researchers and practitioners evaluate the variability of a patient's BSITER score and integrate and describe this variability when providing a diagnosis. We discuss various examples of how researchers and practitioners might do this in the following section.

Intraindividual Diagnostics

Criterion-dependent diagnostics

Criterion dependent diagnostics usually rely on some arbitrary decision criterion (specified a priori) to decide whether a given patient belongs to one classification state or another (e.g., knowledgeable vs unknowledgeable). However, as we have seen in the present report, individuals with scores near a diagnostic threshold may not present consistently above or below that threshold upon re-sampling or re-testing. Therefore, we suggest that researchers take at least one of the following easily derived and intuitive methods into account when conducting criterion-dependent analyses: 1) the 95% confidence interval around the mean, or 2) the probability of each classification state. We note that neither of these options are mutually exclusive and can be used in conjunction to make more informed decisions.

The 95% confidence interval around the BSITER score can easily be calculated for a subject using the rBS method and deriving +/- 2 standard deviations11Note, as discussed by Luck et al., (2021) and Efron & Tibshirani (1994), the standard deviation of the bootstrap mean estimates the standard error, which is why +/- 2 standard *deviations* produces the 95% CI in this example, and not +/- 2 standard *error* . from the mean of that subject's BSITER scores. We propose an intuitive rule for determining patient diagnoses using the 95% CI: if the patient's 95% CI overlaps with the diagnostic threshold, then that patient is labeled as "indeterminate". One can only make a diagnosis with acceptable statistical certainty if the participant's 95% CI does not overlap with the diagnostic cutpoint. Table 1 showcases several illustrative examples of participant results using this method.

Table 2

Several examples of reporting a diagnostic metric (e.g., the BSITER Score), its variability (95% Confidence Interval), and the probability the participant is Guilty/Knowledgeable given the data, illustrated for six participants.

Participant ID	*Memory Type*	*rBS BSITER Score*	*95% CI*	*p(G|data)*
03	Episodic	98.56%	96.26 - 100%	1.0
01	Semantic	100%	100 - 100%	1.0
19	Episodic	89.22%	83.38 - 95.06%	0.95
29	Semantic	93.06%	88.12 - 98%	0.95

44	Episodic	73.22%	64.85 - 81.59%	0.01
55	Semantic	81.06%	72.24 - 90.95%	0.05

For instance, participant 19 had an average BSITER score of 89.22%, with a confidence interval of +/- 5.84%. Therefore, we have 95% certainty that their score falls between 83.38% – 95.06%. Because this range overlaps with the 85% threshold for Episodic memory, we cannot conclude with 95% confidence that Participant 19's BSITER score occurs above threshold, and in result we cannot be confident that they are knowledgeable of crime-relevant information. Therefore, we suggest that the most appropriate label for this participant is "indeterminate".

Note that our assessment differs from the traditional perspective where a participant who scored 89.22% would typically be considered "guilty/knowledgeable" when tested on Episodic, crime-relevant information. However, we note that "indeterminate" labels may be more acceptable for research purposes compared to practice. Indeed, there are applications in diagnostic psychophysiology where an "indeterminate" label is unacceptable (e.g., insurance claims, or investigative work) and a decision must be rendered immediately with the (only) available data. If a classification must be made, then it is important to understand the probability of one decision state or another given the subject's data. The rBS method is also well-suited for this purpose.

Probability

Using rBS, one can calculate the probability that a participant belongs to state A (e.g., guilty/knowledgeable) or state B (e.g., innocent/unknowledgeable), given the data. This is quite simple and intuitive to do, for instance, by counting the number of repetitions that fall above or below a given threshold and dividing by the total number of repetitions conducted. For example, 95/100 (or 95%) of Participant 19's rBS repetitions exceeded the 85% threshold, meaning that 95 times out of 100 Participant 19 would have been diagnosed as "guilty" using the 100 iterations bootstrap test. This is an informative description of the consensus in the results of a diagnostic assessment. Although it is expected that the rate of consensus or agreement across repetitions be quite high, we note that some rate of "disagreement" may be acceptable under specific circumstances. For instance, a lower rate of agreement may be acceptable if the consequences associated with a false positive decision are low. This kind of risk analysis is often overlooked in the CIT literature, and we hope to instigate discussion by drawing more attention to diagnostic precision and consensus across re-sampling or re-testing.

Ultimately, when describing Participant 19 in the context of all available results, we can interpret the data in the following way: "We cannot conclude with 95% statistical confidence that participant 19 recognized information relevant to the crime at-hand. We therefore advise they be labeled "indeterminate" and re-tested using additional crime-relevant information, if available. However, if re-testing is not feasible and a decision must be rendered immediately, it should be noted that the threshold for a "guilty/knowledgeable" decision was satisfied in 95% of the conducted bootstrap repetitions. This may be sufficient for a "guilty/knowledgeable" assessment, assuming the risk associated with a false positive classification is reasonably low. However, if a classification must be made and a judgment rendered "beyond the shadow of a doubt", then an "innocent/unknowledgeable" decision is preferred."

In this modified approach, we acknowledge that a diagnosis may change as a function of both the available data and the context in which the individual is being evaluated. We find this nuanced approach to be more informative and potentially useful in application compared to a single summary statistic supplied by the typical 100 (or even 10,000) iterations bootstrap test. Applying this logic to the remainder of the participants in Table 2, results in the following classification table (Table 3):

Table 3

Example classification decisions for six illustrative participants. "Recommended Classifications" are general-purpose, but decisions may vary depending on the risk associated with the classification. Additional

rows/columns may be added to the table for clarity depending on the sample (e.g., Participant demographics, summary statistics such as true positive or false positive rate, etc).

Participant ID	*Recommended Classification*	*High Risk Classification*	*Low Risk Classification*
03	Guilty/ Knowledgeable	Innocent/ Unknowledgeable	Guilty/ Knowledgeable
01	Guilty/ Knowledgeable	Guilty/ Knowledgeable	Guilty/ Knowledgeable
19	Indeterminate/ Re-test	Innocent/ Unknowledgeable	Guilty/ Knowledgeable
29	Indeterminate/ Re-test	Innocent/ Unknowledgeable	Guilty/ Knowledgeable
44	Innocent/ Unknowledgeable	Innocent/ Unknowledgeable	Innocent/ Unknowledgeable
55	Innocent/ Unknowledgeable	Innocent/ Unknowledgeable	Innocent/ Unknowledgeable
Guilty:	*2/6 (33%)*	*1/6 (17%)*	*4/6 (66%)*
Innocent:	*2/6 (33%)*	*5/6 (83%)*	*2/6 (33%)*
Indeterminate:	*2/6 (33%)*	–	–

Note: Participant 03 is listed as "Innocent/Unknowledgeable" in a high-risk scenario because their 95% CI was not always 100%.

Criterion-free diagnostics

Researchers are often (correctly) discouraged from reporting measures of diagnostic performance derived from a single, arbitrary criterion, and should rather describe the test's performance across all possible criteria. For these purposes, researchers plot the receiver operating characteristic (ROC) curve, from which the area under the curve (AUC) is derived and provides a non-parametric, criterion-free estimate of the test's sensitivity. When deriving the AUC for BSITER scores, we suggest using a large number of iterations (e.g., 10,000). Or, if one has already conducted the rBS test, the average BSITER score from this test can be used.

Limitations & Future directions

We do acknowledge that the analyses presented in this article have some limitations. First, we used data from the Rosenfeld Lab only, so these results may not apply to other labs using other equipment. Further analyses performed on data from different laboratories could not only verify the utility of the presented approach but additionally may or may not also show a difference in BSITER score variability between laboratories, and ultimately could allow for better quality control and understanding of diagnostic metrics in the field of concealed information detection.

Second, we analyzed only "simple guilty" participants - knowledgeable subjects who did not use countermeasure techniques. Therefore, determining the BSITER score variability among other groups, such as innocent subjects or countermeasure users is not possible at this time. Performing the rBS procedure with data from innocent or countermeasure participants is required to provide information about the reliability and repeatability of the classification results in these groups.

Third, the sample size used in these analyses was limited to 81 participants. Although we acknowledge that our results may not generalize from the current sample, we believe that our sample size is not small in the case of an ERP study. Nevertheless, larger and more diverse samples are needed to further support arguments on the utility of the rBS approach in diagnostic psychophysiology.

As we mentioned earlier, in traditional concealed information detection studies, an individual's diagnosis was based on a single-point result (i.e., one BSITER score) without noting its variability. The rBS method

could be a new standard in each future P300-CIT study because it provides information about the stability of the obtained result and the reliability of the classification.

Future studies and analyses should evaluate whether BSITER score variability could differ between protocols (e.g., standard 3SP vs. CTP or single probe vs. multiple probe protocols) or between experimental conditions (e.g., motivated vs. unmotivated participants). The rBS method could also provide new indicators for individual diagnosis. Apart from calculating the mean BSITER score and its confidence interval, rBS allows the analysis of the distribution of eg. 100 BSITER scores for every individual, from which many statistics can be derived. For instance, we can hypothesize that the distribution of BSITER scores should be normally distributed in the case of innocent individuals - since they are unknowledgeable, their mean BSITER score should be around 50% (for innocents, the probe is just another irrelevant stimulus, see eg. Meixner and Rosenfeld, 2011). However, for guilty individuals, since their expected mean BSITER score is close to the boundary (close to or above 85-90%), we can also expect that the distribution of their single BSITER scores will be leptokurtic and left skewed. Therefore, the normality test could be another metric employed in diagnosis. This is one example of an empirical question that could be addressed in future studies and analyses.

References

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.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* . CRC press.

Fabiani, M., Friedman, D., & Cheng, J. C. (1998). Individual differences in P3 Scalp Distribution in older adults, and their relationship to frontal lobe function. *Psychophysiology* ,*35* (6), 698–708. https://doi.org/10.1111/1469-8986.3560698

Gamer, M., & Berti, S. (2012). P300 amplitudes in the concealed information test are less

affected by depth of processing than electrodermal responses.*Frontiers in Human Neuroscience, 6*, 308

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology* ,*58* (6), e13793.

Meixner, J. B., & Rosenfeld, J. P. (2011). A mock terrorism application of the P300-based concealed information test. *Psychophysiology, 48* (2), 149-154.

Olson, J., Rosenfeld, J. P., & Perrault, E. (2019). Deleterious effects of probe-related versus irrelevant targets on the "CIT effect" in the P300-and RT-based three-stimulus protocol for detection of concealed information. *Psychophysiology, 56* (12), e13459.

Olson, J. M., Rosenfeld, J. P., Ward, A. C., Sitar, E. J., Gandhi, A., Hernandez, J., & Fanesi, B. (2022). The effects of practicing a novel countermeasure on both the semantic and episodic memory-based complex trial protocols. *International Journal of Psychophysiology, 173* , 82-92.

Polich, J. (2004). Clinical application of the P300 event-related brain potential. *Physical*

Medicine and Rehabilitation Clinics, *15*(1), 133-161.

Rosenfeld, J. P. (2020). P300 in detecting concealed information and deception: A review.

Psychophysiology, *57* (7), e13362.

Rosenfeld, J. P., Biroschak, J. R., & Furedy, J. J. (2006). P300-based detection of concealed autobiographical versus incidentally acquired information in target and non-target paradigms. *International Journal of Psychophysiology, 60* (3), 251-259.

Rosenfeld, J. P., & Donchin, E. (2015). Resampling (bootstrapping) the mean: A definite do. *Psychophysiology* , *52* (7), 969-972.

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* (2), 118-134.

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., & Chedid, E.

(2008). The Complex Trial Protocol (CTP): A new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology, 45*(6), 906-919.

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to

P300-based tests of detection of concealed information.*Psychophysiology, 41*(2), 205-219.

Rosenfeld, J. P., Ward, A., Drapekin, J., Labkovsky, E., & Tullman, S. (2017a). Instructions to suppress semantic memory enhances or has no effect on P300 in a concealed information test (CIT).*International Journal of Psychophysiology, 113* , 29-39.

Rosenfeld, J. P., Ward, A., Meijer, E. H., & Yukhnenko, D. (2017b). Bootstrapping the P300 in diagnostic psychophysiology: How many iterations are needed?. *Psychophysiology* , *54* (3), 366-373.

Rosenfeld, J. P., Ward, A., Thai, M., & Labkovsky, E. (2015). Superiority of pictorial versus verbal presentation and initial exposure in the P300-based, Complex Trial Protocol for concealed memory detection. *Applied psychophysiology and biofeedback, 40* , 61-73.

Soskins, M., Rosenfeld, J. P., & Niendam, T. (2001). Peak-to-peak measurement of P300

recorded at 0.3 Hz high pass filter settings in intraindividual diagnosis: complex vs. simple paradigms. *International Journal of Psychophysiology, 40*(2), 173-180.

Ward, A. C., & Rosenfeld, J. P. (2017). Attempts to suppress episodic memories fail but do produce demand: evidence from the p300-based complex trial protocol and an implicit memory test. *Applied psychophysiology and biofeedback, 42* (1), 13-26.

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology.

Psychophysiology , *26* (2), 208-221.

Zoumpoulaki, A., Alsufyani, A., & Bowman, H. (2015). Resampling the peak, some dos and

don'ts. *Psychophysiology* , *52* (3), 444-448.