COGNITIVE NEUROPSYCHOLOGY, 2011, 28 (5), 322–337

Psychology Press
Taylor & Francis Group

Bootstrap analysis of the single subject with event related potentials

Ipek Oruç[1,2], Olav Krigolson[3], Kirsten Dalrymple[4], Lindsay S. Nagamatsu[4], Todd C. Handy[4], and Jason J. S. Barton[1,2,4]

[1]Department of Ophthalmology and Visual Science, University of British Columbia, Vancouver, BC, Canada
[2]Department of Medicine (Neurology), University of British Columbia, Vancouver, BC, Canada
[3]Department of Psychology, Dalhousie University, Halifax, NS, Canada
[4]Department of Psychology, University of British Columbia, Vancouver, BC, Canada

Neural correlates of cognitive states in event-related potentials (ERPs) serve as markers for related cerebral processes. Although these are usually evaluated in subject groups, the ability to evaluate such markers statistically in single subjects is essential for case studies in neuropsychology. Here we investigated the use of a simple test based on nonparametric bootstrap confidence intervals for this purpose, by evaluating three different ERP phenomena: the face-selectivity of the N170, error-related negativity, and the P3 component in a Posner cueing paradigm. In each case, we compare single-subject analysis with statistical significance determined using bootstrap to conventional group analysis using analysis of variance (ANOVA). We found that the proportion of subjects who show a significant effect at the individual level based on bootstrap varied, being greatest for the N170 and least for the P3. Furthermore, it correlated with significance at the group level. We conclude that the bootstrap methodology can be a viable option for interpreting single-case ERP amplitude effects in the right setting, probably with well-defined stereotyped peaks that show robust differences at the group level, which may be more characteristic of early sensory components than late cognitive effects.

Keywords: Event-related potentials; Bootstrap analysis; Single subject analysis; N170; Error-related negativity; P3.

In identifying the contribution of different cerebral regions to specific cognitive functions, neuropsychological case studies have long played a key role, which has been further enhanced by the ability of modern neuroimaging to provide detailed structural information about cerebral lesions in living patients. Since the structural impact of pathological processes such as tumours, strokes, infections, and trauma are unique to each patient, it is often inappropriate to consolidate the results of different patients within a group analysis. Rather, each case needs to be

Correspondence should be sent to Ipek Oruç, Department of Ophthalmology and Visual Sciences, VGH Research Pavilion, 828 W 10th Ave, Vancouver, BC V5Z 1L8, Canada (Email: ipor@mail.ubc.ca).

J.B. was supported by a Canada Research Chair. K.D. was supported by NSERC (Natural Sciences and Engineering Research Council of Canada) and MSFHR (Michael Smith Foundation for Health Research) fellowships.

http://www.psypress.com/cogneuropsychology http://dx.doi.org/10.1080/02643294.2011.648176

considered individually. The difficulty in interpreting the data of an individual is the loss of the between-subject estimates of variability in the experimental measure. This creates a challenge in determining whether a behavioural or neurophysiological marker is present or absent, and whether any anomalies are due to the brain damage or simply an experimental or sampling error.

An alternative approach to the use of between-subject variance as a gauge of the consistency of the experimental measure is to assess within-subject variability. This can be difficult, though, as the data of repeated trials within a single subject may be more variable and noisy than the subject mean data and may not fulfil certain assumptions such as normality, which conventional parametric methods require. One method that is relatively free of assumptions is the bootstrap approach. This method is based on forming a *sampling distribution* for the statistic of interest (e.g., mean) by resampling from the raw data with replacement a large number of times. The bootstrap distribution can then be used to determine confidence intervals or for hypothesis testing. Since the bootstrap distribution is obtained empirically by iteratively resampling the original data, it is useful in settings where assumptions for normality and equality of variances are not met in principle or when working with small or unequal sample sizes.

The bootstrap method has been used in a few event-related potential (ERP) studies to date (e.g., Charest et al., 2009; Philiastides, Ratcliff, & Sajda, 2006; Philiastides & Sajda, 2006; Rousselet et al., 2010; Rousselet et al., 2009). It has been used to assess reliability of electrode locations showing maximal P3 activity across subjects (Fabiani, Gratton, Corballis, Cheng, & Friedman, 1998), to examine the latencies in which differences in N170 amplitudes between objects and faces emerge (Rousselet, Husk, Bennett, & Sekuler, 2008), to provide confidence intervals for ERP waveform amplitudes and compare those across two participant subgroups (Caryl, Golding, & Hall, 1995), and to obtain group-level significance for N170 amplitude and behavioural performance (d') differences, as well as the correlation between those two (Vizioli, Foreman, Rousselet, & Caldara, 2010), for example. However, it is unclear from these specialized applications of the bootstrap method how useful this technique is in the evaluation of the single subject in neuropsychology, particularly with respect to the ability to determine simply whether a particular ERP component is present or absent in a given patient.

To assess the viability of the nonparametric percentile bootstrap technique (Efron & Tibshirani, 1993; Wilcox, 2005; Wilcox & Keselman, 2003) to estimate statistical significance at the single-subject level, we examined three well-documented but very different ERP phenomena, in which amplitude changes are evaluated as markers of underlying cognitive processes. We applied a percentile bootstrap procedure to determine the statistical consistency of these effects at the single-subject level. In addition, to provide an empirical guideline to gauge the potential efficacy of this technique for various other ERP phenomena, we related the results of these single-subject analyses to conventional parametric statistics performed at the group level. Our choices were guided by two criteria: to study well-established ERP phenomena that are consistently obtained in groups of healthy subjects; and second, to include a range of ERP components from early compact perceptual effects to later, broader, cognitive ones.

For our first analysis, we examined the face-selective N170, a negative-going potential observed in occipitotemporal sites occurring between 140 and 200 ms after stimulus onset, which is often larger for faces than for other objects, especially at the right side (Bentin, Allison, Puce, Perez, & McCarthy, 1996; Eimer & McCarthy, 1999; Itier & Taylor, 2004). For our second analysis, we examined the feedback error-related negativity (feedback ERN), a negative deflection in the waveform observed that is maximal over medial frontal cortex (i.e., electrode FCz) typically occurring 200 to 300 ms following feedback delivery, and which is less pronounced or nonexistent following feedback indicating correct performance than after feedback indicating incorrect performance (Miltner, Braun, & Coles,

1997); hence it is considered to be a marker for error-detection activity. In our final analysis, we looked at the P3 waveform, a positive-going potential occurring at midline sites between 300 and 600 ms after stimulus onset, which is larger for unexpected stimuli than for expected stimuli (Eimer, 1996, 1998; Nagamatsu, Liu-Ambrose, Carolan, & Handy, 2009).

EXPERIMENT 1: FACE-SELECTIVE N170 (DALRYMPLE ET AL., 2011)

Method

Subjects

Ten healthy subjects with normal or corrected-to-normal vision participated (7 female, ages 18–59 years). All subjects except one (E.W.) were right-handed. The protocol was approved by the institutional review boards of Vancouver General Hospital and the University of British Columbia, and all subjects gave informed consent in accordance with the Declaration of Helsinki.

Stimuli and procedure

Five faces and five objects (stapler, book, banana, water bottle, and a tea pot) were used. All faces displayed a neutral expression, were cropped to exclude hair, and were unfamiliar to the participants. The stimulus set consisted of four versions of each of the five face and objects (total of 40 images).

Subjects were seated 1 m from the computer screen in an otherwise dark room. Each trial started with a fixation period of 2,700–2,900 ms, followed by the stimulus, either a face or an object, displayed for 100 ms, which was followed by a 300-ms mask. The subjects performed an irrelevant pleasant/unpleasant task by pressing one of two buttons on a joystick. Subjects completed 200 trials for each of the two conditions (face and object).

Electrophysiological recording

Scalp potentials were recorded using a 64-channel Bio-Semi Active 2 system relative to two medial frontal electrodes (CMS and DRL) at 256 Hz. Offline, the electroencephalography (EEG) waveforms were re-referenced to the average of the right and left mastoid electrodes and were low-pass filtered using a Butterworth filter at 25.6-Hz half-amplitude cut-off. Trials where blinks occurred were determined and were rejected based on minimum and maximum thresholds of vertical and horizontal electro-oculograms of individual subjects. Baseline correction was performed by normalizing waveforms relative to a baseline occurring within a 200-ms prestimulus period.

Data analysis

We focused on the measurements at the right posterior lateral site P8, based on well-replicated data from group studies (Bentin et al., 1996; Eimer & McCarthy, 1999; Jacques, d'Arripe, & Rossion, 2007; Webb et al., 2010) that face selectivity is more prominent in the right N170. For the standard group analysis, we calculated grand averages across all trials of all subjects for the face and object conditions. The peak latencies of N170 for face and object conditions were determined for the group as a whole. Average amplitudes within a 40-ms window of this group peak for each subject were analysed with a repeated measures analysis of variance (ANOVA), with stimulus condition as a within-subjects factor (face, object). Although some group studies focus upon a single peak value to represent amplitude (Bentin et al., 1996; Botzel, Schulze, & Stodieck, 1995), using only a single value is likely to be quite noisy when dealing with analyses based on single trial data. Hence, to improve signal-to-noise ratio in the bootstrap analysis, we represent amplitude by the average values in a small window centred on the peak of the potential. To make our contrast between group ANOVA and single-subject bootstrap comparable, we use the same window for both methods. Similarly, other ERP studies have entered windowed peak data into their group analyses (Eimer & McCarthy, 1999; Itier & Taylor, 2004; Jacques et al., 2007).

For the bootstrap single-subject analyses, trials were first categorized into the two conditions of face and object, and waveforms were averaged for a subject across the 200 trials in each condition. For each subject, peak latency was indexed as the

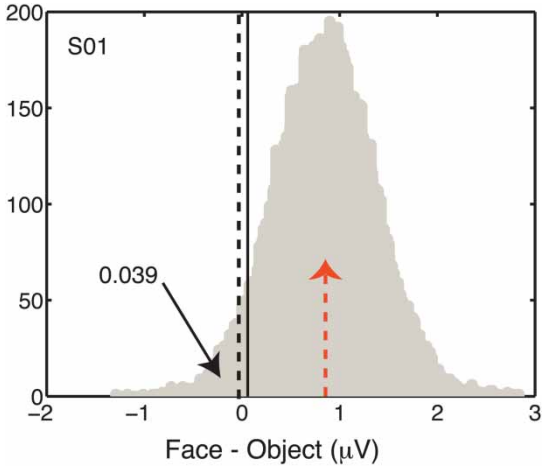

Figure 1. *Illustration of the percentile bootstrap procedure. The bootstrap histogram for the face–object contrast is shown for the N170 data of subject S01. The solid black line marks the 5th percentile value, in this case above zero, indicating a statistically significant difference (one-tailed, face > object). The proportion of resamples that were smaller than zero, in this case .039, yields the exact* p*-value. The face–object difference based on the averaged data for this subject is shown in red (dashed) arrow. To view a colour version of this figure, please see the online issue of the Journal.*

time between stimulus onset and the time when the slope of the mean curve changed sign from negative to positive, within the time interval of 100–200 ms. A temporal window from 20 ms before to 20 ms after the peak of the N170 was determined separately for the face and object conditions, for each individual subject. The difference between the mean potentials in this 40-ms window was taken to represent the face–object contrast. To test whether this contrast was significantly larger than zero for each subject (i.e., larger N170 amplitude for the face condition), we performed a nonparametric bootstrap simulation—a Monte Carlo technique that utilizes the variability across individual trials to determine statistical significance (Efron & Tibshirani, 1993). This simulation involves forming a large number of resampled datasets consisting of 200 trials per condition, with each trial chosen randomly and independently from the original set of 200 trials with replacement. This means that at each random draw, all original trials are available, and thus a given trial can be selected in the resampled dataset more than once. A histogram of face–object contrast values obtained from 50,000 resampled datasets was formed. The lower 5th percentile point of this histogram served as the critical value for (one-tailed) significance at the .05 level. Figure 1 shows the bootstrap histogram of face–object contrast for subject S01. The 5th percentile mark (solid black line) is above zero (i.e., face > object in more than 95% of the resamples), indicating significantly larger amplitudes in the face condition than in the object condition. The proportion of resamples smaller than zero (i.e., object > face) was .039 (dashed black line), yielding the exact *p*-value.

EXPERIMENT 2: FEEDBACK ERROR-RELATED NEGATIVITY (ERN) (KRIGOLSON , HEINEKEY, KENT, & HANDY, 2012)

Method

Subjects

Fourteen healthy subjects with normal or corrected-to-normal vision participated (7 female, ages 19–30 years). Data for two subjects were excluded due to excessive noise. The protocol was approved by the Institutional Review Board of the University of British Columbia, and all subjects gave informed consent in accordance with the Declaration of Helsinki.

Stimuli and procedure

The subjects' task was to estimate a one-second duration following the offset of a short auditory tone (see Holroyd & Krigolson, 2007, for more details). Following the subjects' estimate, visual feedback was provided indicating whether the estimate was correct (a check mark) or incorrect (an X). The criterion for determining correct response changed online throughout the session, getting more or less stringent based on the subject's performance. This was done to produce approximately equal numbers of correct and incorrect feedback trials out of a total of 1,200 trials.

Electrophysiological recording

Scalp potentials were recorded using a 32-channel Bio-Semi Active 2 system relative to two medial frontal electrodes (CMS and DRL) at 256 Hz. Offline, the EEG waveforms were re-referenced to the average of the right and left mastoid electrodes and were low-pass filtered using a Butterworth filter at 25.6-Hz half-amplitude cut-off. Trials where blinks occurred were determined and were rejected based on minimum and maximum thresholds of vertical and horizontal electro-oculograms of individual subjects. Baseline correction was performed by normalizing waveforms relative to a baseline occurring within a 200-ms prestimulus period.

Data analysis

The analysis steps were the same as those in Experiment 1 except where otherwise indicated below. Our analysis focused on the measurements at the medial frontal site FCz (Gehring & Fencsik, 2001; Holroyd & Krigolson, 2007; Krigolson & Holroyd, 2006). For the group analysis, grand averages across all trials of all subjects were calculated for the error and correct conditions. The peak latency for error and correct conditions were determined once at the group level. Average amplitudes within a 40-ms window of this group peak for each subject were analysed with a repeated measures ANOVA with stimulus condition as a within-subjects factor (correct, error). This method deviates slightly from the usual way of analysing the ERN, which usually calculates a mean difference waveform between error and correct trials and finds the peak difference in that difference waveform. However, if our individual subjects show slight intrasubject variations in the latency rather than the amplitude between error and correct potentials, this too could lead to a significant difference waveform, despite the lack of any real change in amplitude. By allowing the window of measurement to vary separately for activity peaks in error and correct waveforms, we minimize the impact of latency differences and provide a more conservative measure of the effect in both the group and the single subject.

For the bootstrap analyses of single subjects, trials were first categorized into the two conditions: error and correct feedback, and waveforms were averaged across all trials (roughly 600 trials per condition) in each condition. For each subject, peak latency was indexed as the time the slope of the mean curve changed sign from negative to positive, within the time interval of 200–400 ms following feedback onset. A 40-ms window around the peak latency was determined separately for the error and correct conditions, for each individual subject. An error–correct contrast was obtained as the averaged difference score within the 40-ms window around the peak. To test whether this contrast was significantly larger than zero for each individual subject (i.e., larger negativity for the error condition), we performed a nonparametric bootstrap simulation as described above. A histogram of error–correct contrast values obtained from 50,000 resampled datasets was formed. The lower 5th percentile point of this histogram served as the critical value for (one-tailed) significance at the .05 level.

EXPERIMENT 3: P3 (NAGAMATSU ET AL., 2009)

Method

Subjects

Ten healthy subjects with normal or corrected-to-normal vision participated (all female, ages 66–74 years). The protocol was approved by the institutional review board of the University of British Columbia, and all subjects gave informed consent in accordance with the Declaration of Helsinki.

Stimuli and procedure

Subjects were seated 1 m away from the computer screen in an otherwise dark room. Each trial started with the presentation of a central cross for 1 s, which the subjects were instructed to fixate for the duration of the trial. The fixation cross was followed by an arrow pointing either to the left or to the right, which remained on the screen until the end of the trial; 900–1,100 ms

after the onset of the arrow, the target stimulus (an "X") appeared 80% of the time on the side indicated by the arrow, and 20% of the time on the opposite side—that is, the arrow served as a cue that was predictive 80% of the time. The subject's task was to indicate which side the target stimulus appeared as quickly and accurately as possible by pressing a button with their left hand if the target was on the left and vice versa.

Electrophysiological recording

Scalp potentials were recorded using a 32-channel Bio-Semi Active 2 system relative to two medial frontal electrodes (CMS and DRL) at 256 Hz. Offline, the EEG waveforms were re-referenced to the average of the right and left mastoid electrodes and were low-pass filtered using a Butterworth filter at 25.6-Hz half-amplitude cut-off. Trials where blinks occurred were determined and were rejected based on minimum and maximum thresholds of vertical and horizontal electro-oculograms of individual subjects. Baseline correction was performed by normalizing waveforms relative to a baseline occurring within a 200-ms prestimulus period.

Data analysis

The analysis steps were same as those in Experiment 1 except where indicated otherwise. Our analysis focused on the measurements at the medial electrode sites Fz, Cz, and Pz. For the group analysis, grand averages across all trials of all subjects were calculated for the cued and uncued conditions. Based on the group data (Figures 2c–2e), we created the following broad temporal windows for analysis: a 350-ms time window starting at 355 ms after stimulus onset for the Fz data, a 300-ms time window starting at 380 ms after stimulus onset for the Cz data, and a 300-ms time window starting at 450 ms after stimulus onset for the Pz data. Average amplitudes within these temporal windows for each subject were analysed with three separate repeated measures ANOVAs, one for each of Fz,

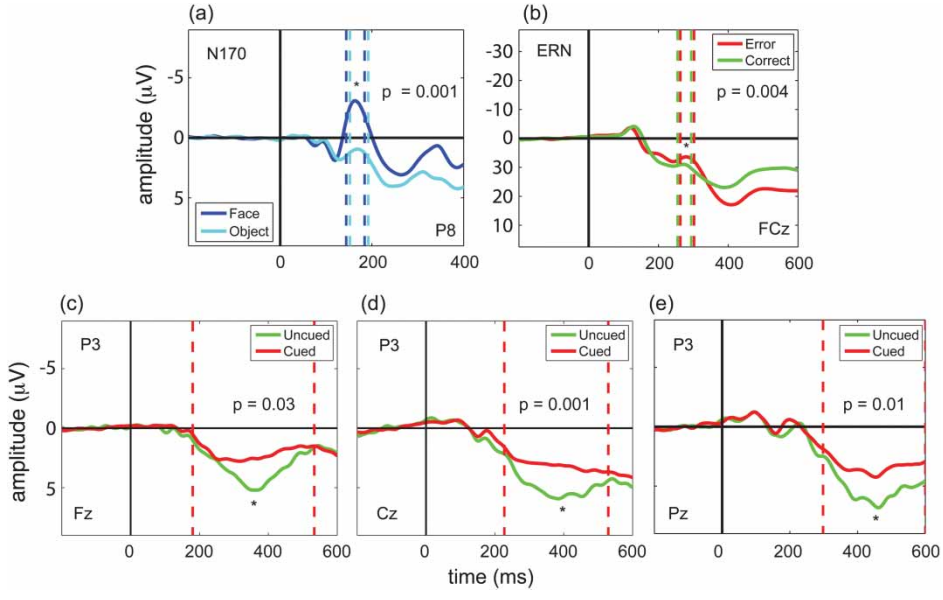

Figure 2. *Group data for all studies. Dashed lines indicate the time window that was included in the analysis determined at the group level. An asterisk indicates a significant contrast, based on a group-level analysis of variance (ANOVA), with the corresponding* p*-value indicated on the panel. (a) Experiment 1, N170, P8 electrode. (b) Experiment 2, ERN (error-related negativity), FCz electrode. (c) Experiment 3, P3, Fz electrode. (d) Experiment 3, P3, Cz electrode. (e) Experiment 3, P3, Pz electrode. To view a colour version of this figure, please see the online issue of the Journal.*

Cz, Pz, with stimulus condition as a within-subjects factor (cued, uncued).

The bootstrap analysis proceeded slightly differently in this analysis. Given the nature of the P3 as a broad component with a relatively loosely defined structure, and the broad temporal windows thus derived, we let the group analysis guide the choice of a fixed and relatively wide time window applied to all participants, rather than attempting to tailor a narrow time window on a subject-to-subject basis. Trials were first categorized into the two conditions—cued and uncued—and waveforms were averaged across all trials in each condition. We used the same fixed temporal windows as those in the group analysis of the Fz, Cz, and Pz data. An uncued−cued contrast was obtained as the averaged difference score within the specified time window. To test whether this contrast was significantly larger than zero for each individual subject (i.e., larger amplitude for the uncued condition), we performed a nonparametric bootstrap simulation as described above, separately for the Fz, Cz, and Pz data. In each case, a histogram of uncued−cued contrast values obtained from 50,000 resampled datasets was formed. The lower 5th percentile point of this histogram served as the critical value for (one-tailed) significance at the .05 level.

RESULTS

For the right face-selective N170, the group analysis (Figure 2a) replicated the finding of larger amplitude for the face than for the object conditions, $F(1, 9) = 24.60$, $p = .001$. The bootstrap analysis showed that all 10 subjects show a significant face-selective N170 (Figure 3).

For the error-related negativity, the group analysis (Figure 2b) confirmed a difference between the error and correct conditions, $F(1, 11) = 13.09$, $p = .004$, with larger amplitudes for the error condition. The bootstrap analysis showed that 10 out of 12 subjects show a significant error > correct effect, and 1 additional subject showed a trend in that direction ($p < .1$), with only 1 subject failing to show any difference (Figure 4).

For the P3, the group analysis showed a difference between uncued and cued conditions at each site—Fz: $F(1, 9) = 6.54$, $p = .03$; Cz: $F(1, 9) = 21.07$, $p = .001$; Pz: $F(1, 9) = 10.23$, $p = .01$—with larger amplitudes for the uncued condition (Figures 2c–2e). The bootstrap analysis showed that at Fz, 4 out of 10 subjects show a significant effect, and 3 additional subjects show a difference in the expected direction without reaching significance (Figure 5). At Cz, 6 out of 10 subjects show a significant effect, and 2 additional subjects show insignificant differences in the expected direction (Figure 6), while at Pz, 5 out of 10 subjects show a significant effect, 1 subject showed a trend ($p < .1$), and 3 additional subjects show insignificant differences in the expected direction (Figure 7).

An analysis of the frequency of significant results at the single-subject level as a function of the group ANOVA F- and p-values showed a relationship (Figure 8): Stronger group ANOVA results were associated with a larger proportion of single subjects with significant results. The percentage of single-subject significance was negatively correlated with group ANOVA p-values ($r = -.74$, $p = .03$, based on bootstrap) and positively correlated with group ANOVA F-value ($r = .75$, $p = .04$).

DISCUSSION

We examined three distinct ERP phenomena—N170, ERN, and P3—in which cognitive processing is reflected primarily as changes in the amplitude of potentials, using both a conventional group-level analysis based on parametric statistics and a single-subject analysis based on nonparametric percentile bootstrap simulations. Our purpose was to assess whether the bootstrap method can be used to evaluate statistical significance at the single-subject level. While the group analysis showed significant amplitude effects for all three phenomena, consistent with the prior literature, the bootstrap analysis of single subjects produced a range of results. For the N170, the

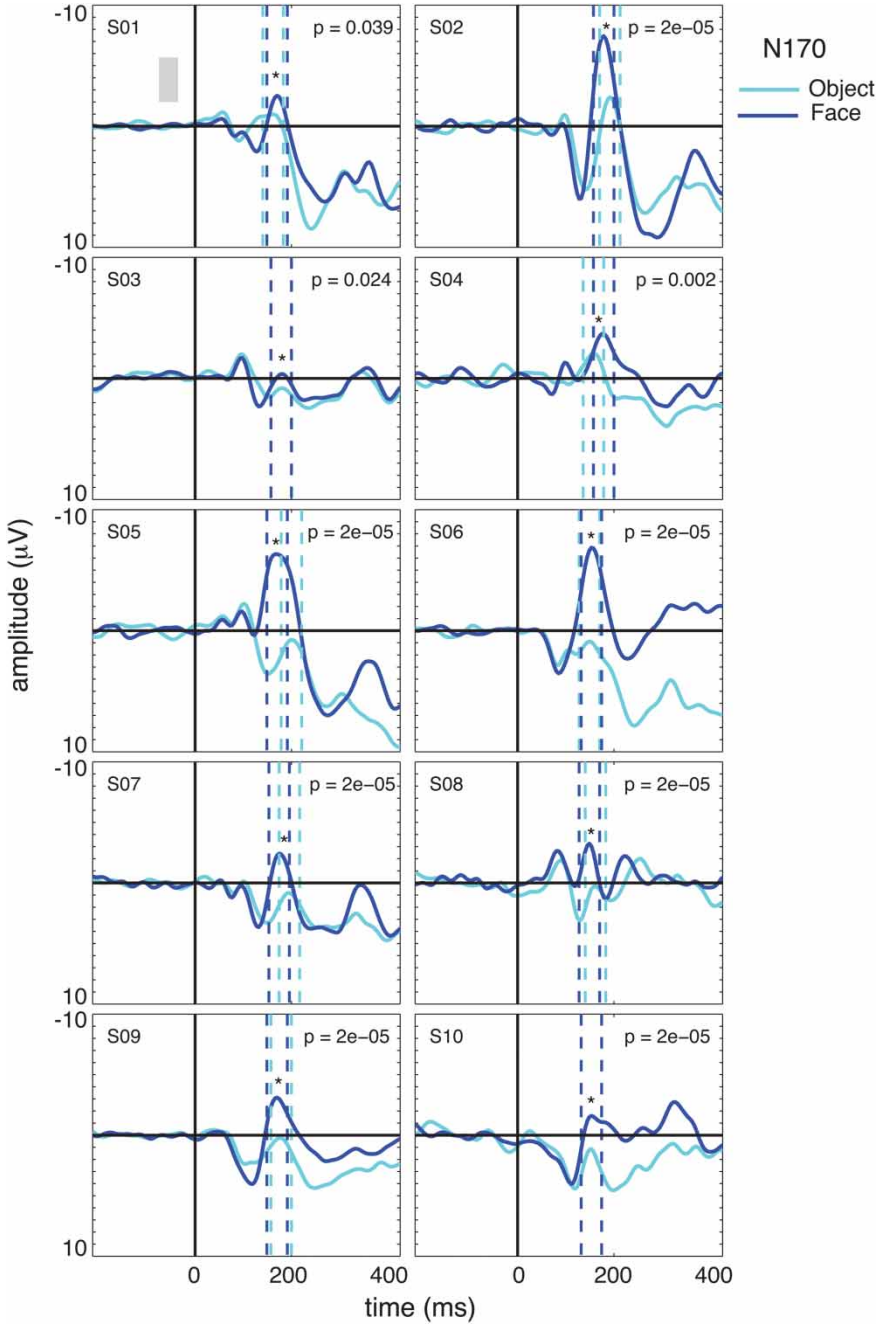

Figure 3. *Experiment 1, N170, P8 electrode, single-subject data. Data for 10 subjects are plotted in separate panels. Solid curves indicate face (blue) and the object (cyan) conditions. Dashed lines indicate the 40-ms time window around the corresponding peak for the two conditions for each subject. The group amplitude difference between the two conditions is shown in light grey on the first (top-left) panel for comparison. An asterisk indicates a significant face–object contrast, with the corresponding* p*-value for each subject plotted on each panel. To view a colour version of this figure, please see the online issue of the Journal.*

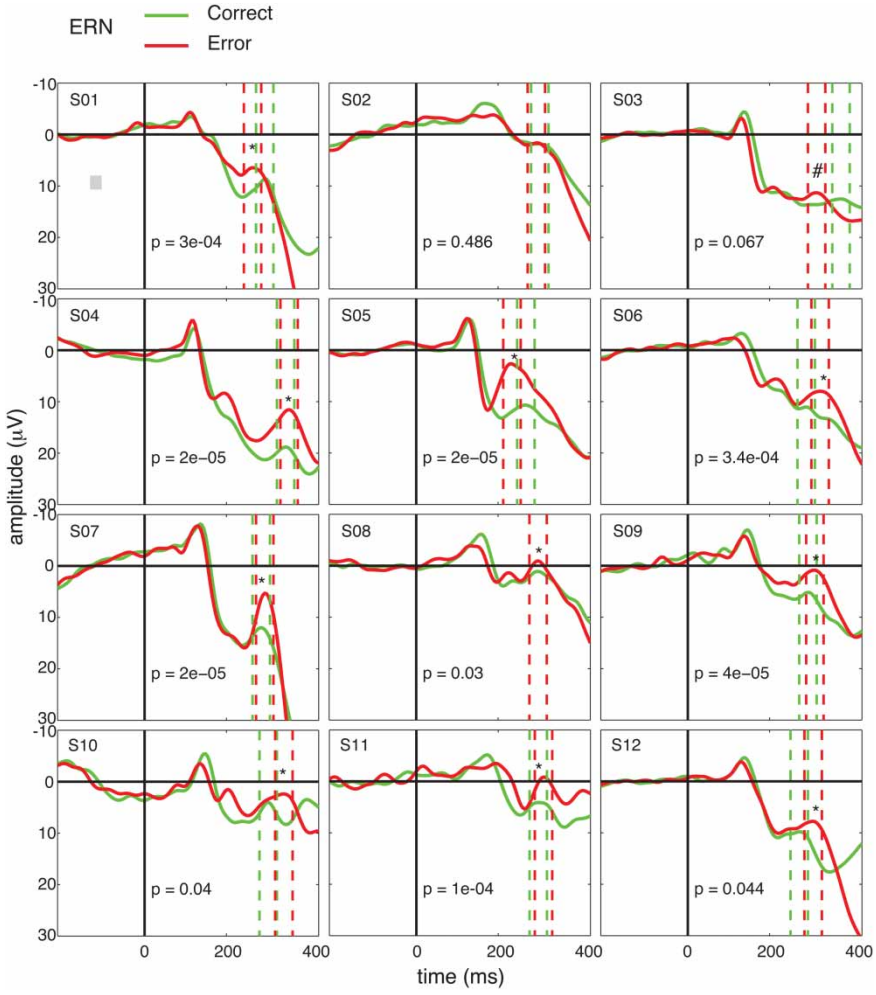

Figure 4. *Experiment 2, ERN (error-related negativity), FCz electrode, single-subject data. Data for 12 subjects are plotted in separate panels. Solid curves show the data for the error (red) and the correct (green). Dashed lines indicate the 40-ms time window around the corresponding peak for the two conditions (also shown bottom-left of each panel) for each subject. The group amplitude difference between the two conditions is shown in light grey on the first (top-left) panel for comparison. An asterisk indicates a significant error–correct contrast, and a hash sign (#) indicates a trend, with the corresponding* p*-value for each subject plotted on each panel. To view a colour version of this figure, please see the online issue of the Journal.*

bootstrap analysis showed individually significant data for all subjects. For the ERN, individual effects were significant for 9 out of the 12 subjects. For the P3, the bootstrap analysis was less successful, with only about half of the subjects' data significant at the individual level.

A number of reasons may account for the variability in the single-subject results between these three ERP phenomena. First and most importantly, it may reflect the degree of significance of each phenomenon. Significance at the group level does not necessarily predict significance at the single-subject level, since group-level statistics are affected mainly by intersubject variance while single-subject analyses reflect intrasubject variance alone. Nevertheless, both would be greater for

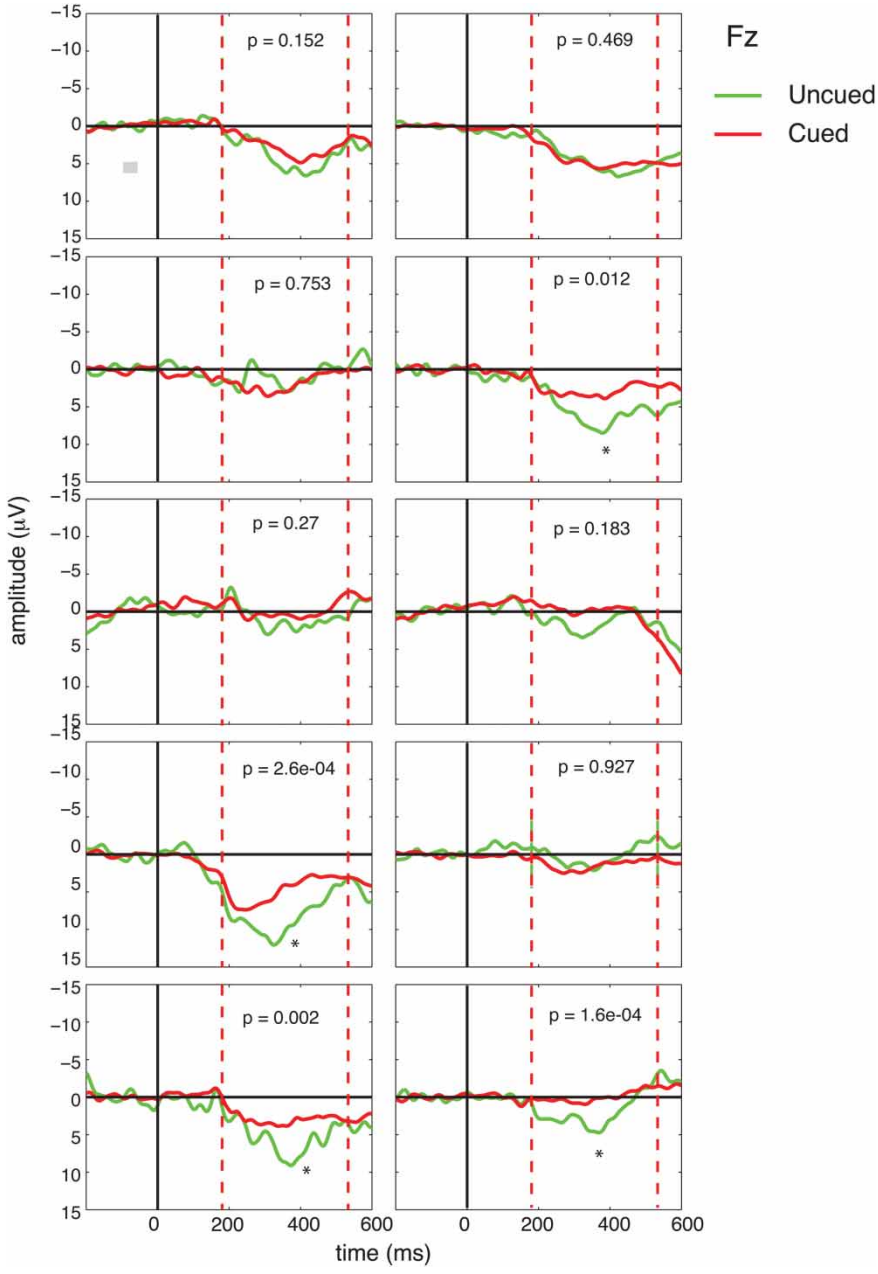

Figure 5. *Experiment 3, P3, Fz, single-subject data. Data for 10 subjects are plotted in separate panels. Solid curves indicate the cued (red) and the uncued (green) conditions. Dashed lines delineate the fixed temporal windows that were determined based on the group data. The group amplitude difference between the two conditions is shown in light grey on the first (top-left) panel for comparison. An asterisk indicates a significant contrast (uncued > cued), and a hash sign (#) indicates a trend. To view a colour version of this figure, please see the online issue of the Journal.*

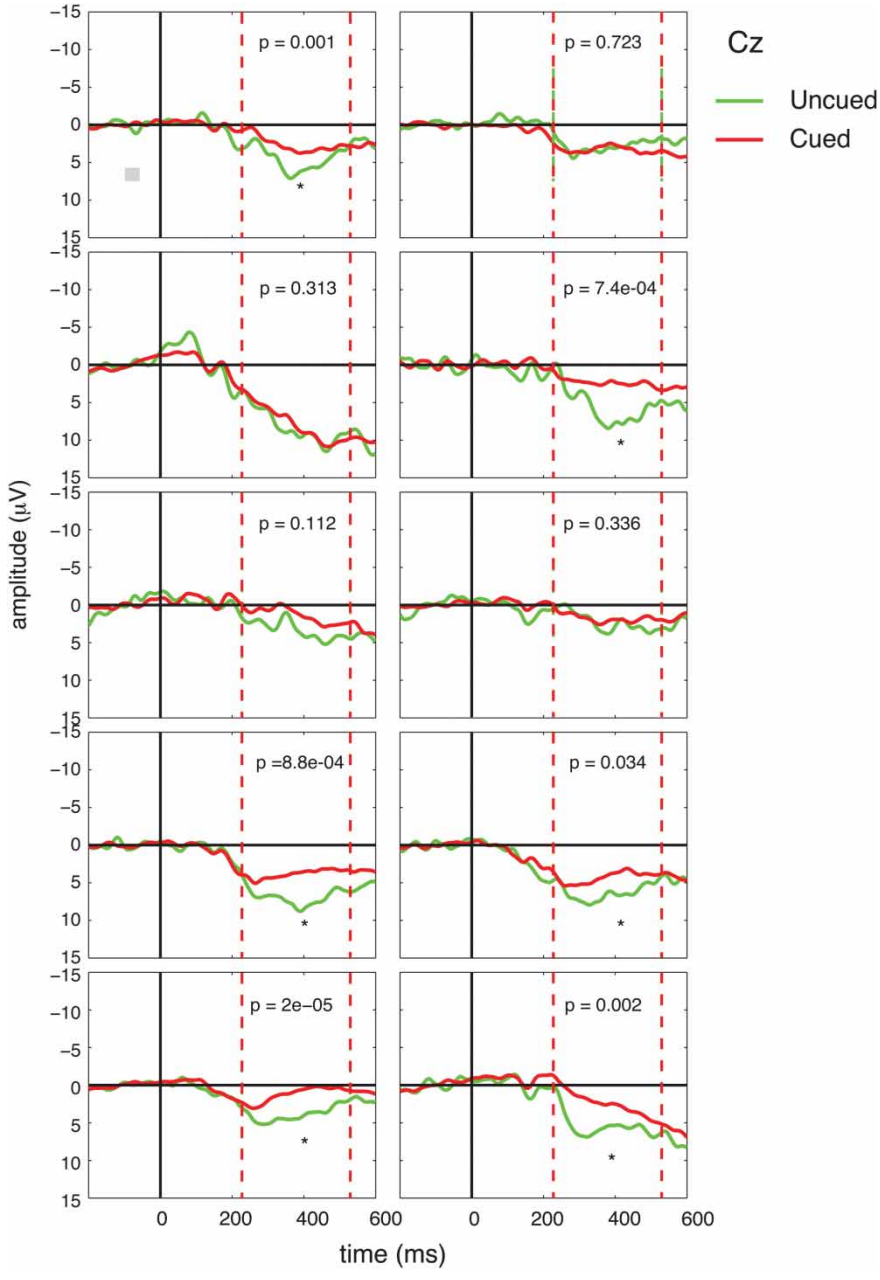

Figure 6. *Experiment 3, P3, Cz, single-subject data. Data for 10 subjects are plotted in separate panels. Solid curves indicate the cued (red) and the uncued (green) conditions. Dashed lines delineate the fixed temporal windows that were determined based on the group data. The group amplitude difference between the two conditions is shown in light grey on the first (top-left) panel for comparison. An asterisk indicates a significant contrast (uncued > cued), and a hash sign (#) indicates a trend. To view a colour version of this figure, please see the online issue of the Journal.*

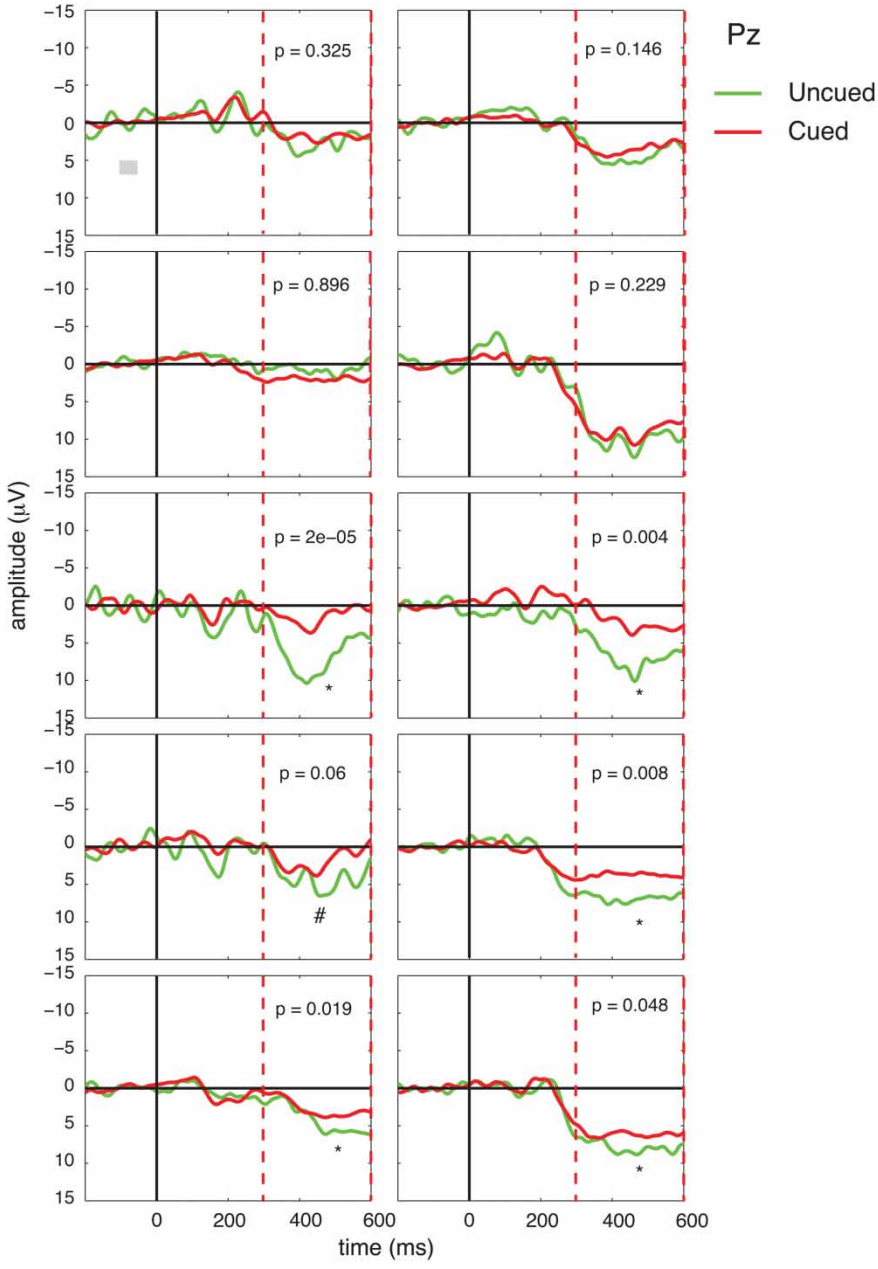

Figure 7. *Experiment 3, P3, Pz, single-subject data. Data for 10 subjects are plotted in separate panels. Solid curves indicate the cued (red) and the uncued (green) conditions. Dashed lines delineate the fixed temporal windows that were determined based on the group data. The group amplitude difference between the two conditions is shown in light grey on the first (top-left) panel for comparison. An asterisk indicates a significant contrast (uncued > cued), and a hash sign (#) indicates a trend. To view a colour version of this figure, please see the online issue of the Journal.*

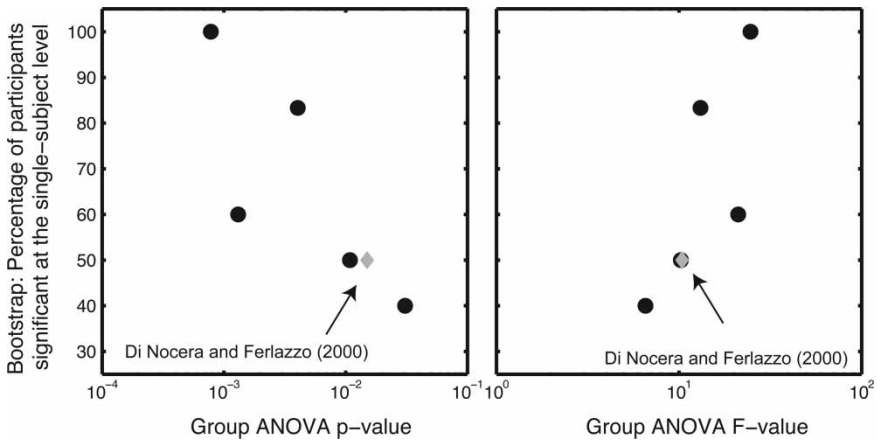

Figure 8. *Meta-analysis comparing bootstrap at the single-subject level and conventional ANOVAs (analyses of variance) at the group level. Percentage of subjects that showed a significant effect at the single-subject level using bootstrap analysis is plotted as a function of the* p*-value (left panel) and* F*-value (right panel), obtained with a conventional group-level ANOVA on the same data sets. Diamond marker shows related result from Di Nocera and Ferlazzo (2000).*

phenomena with larger differences in the underlying true signals. Our comparison of the bootstrap analysis with the results of the group-level ANOVAs (Figure 8) confirms that there is an effective relationship between the two: There are more consistently significant bootstrap outcomes in single subjects when the group analysis is more significant (i.e., lower p-values, higher F-values). This result provides a useful empiric rule of thumb for judging which ERP phenomena would be suitable candidates for successful application of a bootstrap method for single-subject analysis: It suggests that to obtain a 95% rate of significance in single subjects, a group ANOVA significance of $p < .001$ and $F > 20$ is required for sample sizes of 10 to 15 subjects.

The variations in the consistency of the signal in different ERP phenomena is probably related in part to the dynamics and nature of the underlying cognitive process responsible for its generation. For any given amplitude difference in potentials, narrow peaks with stereotyped temporal dynamics are more likely to yield more consistent statistical analyses within and between individuals. Such peaks are more typical of earlier, more perceptual ERP components, such as the N170, whereas later, more cognitive ERP phenomena tend to show broad peaks with more variable waveforms between subjects, as shown in our P3 data. Just as our P3 analysis showed greater variability of this phenomenon at the single-subject level, an earlier work by Di Nocera and Ferlazzo (2000) showed that a significant group-level difference in a memory task for ERPs in the 400–800-ms latency range was only found for half of the subjects in a single-subject bootstrap (also depicted in Figure 8). This is highly consistent with our results: The correlation between group-level significance and significance frequency in single subjects is virtually unchanged when we include the values reported in Di Nocera and Ferlazzo (2000) with ours ($r = -.75$, $p = .007$, and $r = .77$, $p = .01$, for p and F values, respectively, based on bootstrap). Hence, single-subject bootstrap methods for amplitude effects in later cognitive potentials may only be useful if the amplitude differences are large and highly significant at a group level.

Although analyses of ERP amplitude effects at the group level are most commonly done using conventional statistics, there are a few prior studies that have used the bootstrap method in a group-based approach. Rousselet, Husk, Bennett, and Sekuler (2005) investigated the effect of stimulus eccentricity on the amplitude of face-selective N170, and whether such effects can be

explained by low-level visual factors. Instead of focusing on specific electrodes and latencies determined a priori, they resampled across subjects at each electrode and time point to obtain a time series of group-level statistical significance values for the ERP amplitude using bootstrap. Based on this analysis, they were able to demonstrate a significant eccentricity effect on the amplitude of the face-selective N170, which was eliminated when stimulus sizes were matched according to the appropriate cortical magnification factor. In another study, Vizioli and colleagues (2010) examined whether the face inversion effect observed in the amplitude of N170 (larger amplitudes for inverted than for upright faces) was modulated by the race of the face stimuli. They performed a bootstrap analysis by resampling the N170 face inversion effect (FIE) across subjects and demonstrated that FIE was significantly larger for same-race face stimuli than for other-race faces. While their bootstrap results generally agreed with the conventional parametric tests they performed, their motivation for using the bootstrap analysis was the robustness of this method against outlier subjects and in settings with only a small number of participants.

In the present study, we investigated the utility of the bootstrap method for a different purpose, but one that is highly relevant to neuropsychology, the ability to determine the presence or absence of a cognitive phenomenon in one person. We examined a number of ERP components ranging from early compact perceptual waveforms to later broad cognitive ones to assess the general applicability of bootstrap to amplitude effects by resampling individual trials in one person. The statistical evaluation of single subjects is critical in the neuropsychological field, because some conditions are rare while others are heterogeneous in either pathophysiology or lesion anatomy. Being able to determine whether an ERP component is missing or present in an individual patient requires the demonstration that the component is consistently and uniformly present in healthy subjects. Only then can the absence of an ERP component in a lesioned patient be considered relevant for structure–function correlations.

While we focused upon the analysis of amplitude effects for neuropsychological purposes, the bootstrap technique could also be adopted easily to analyse other effects. Given the high temporal resolution of ERP data, it may be of greater theoretical interest in some studies to evaluate latency differences rather than amplitude differences. However, while our estimations of amplitude rely on averaging within a temporal sampling window around a peak, latency calculations depend on the precision of defining the single data point that represents the peak in the waveform. Hence, as with our amplitude analyses, latency analyses will also be more robust for large sharp peaks and may be more vulnerable to intrasubject noise. Inspection of Figures 3–5 illustrates some of the challenges that would be faced by a latency analysis for the P3.

Compared to group studies, single-subject research is more challenging: It requires modifications to design and analyses because of the statistical issues related to assessing within-subject trial-to-trial variance rather than between-subject differences in mean performance (see Crawford & Garthwaite, 2006, 2007; Crawford, Garthwaite, & Porter, 2010, for a comprehensive treatment of statistical assessment of single-subject data using a matched control sample as comparison). Our study was aimed at determining how confidently one could state whether a particular phenomenon was present or absent in a given subject. However, single-subject research can also be directed at other goals. Besides assessing for the presence of an ERP component, it may also be important to assess for the normalcy of that component. It is possible for the data of a patient to show significant amplitude differences between two experimental conditions, but these differences may be either heightened or reduced compared to the magnitude of those differences in a normal population. Future work can examine whether the bootstrap can be used as an effective tool to provide estimates and significance of effect sizes for ERP component magnitudes or amplitude differences at the single-subject level. Investigations of other ERP effects such as steady-state responses may also be amenable to a bootstrap approach, but, as with our work,

studies will need to be done to establish for each effect the consistency of the effects within subjects.

Bootstrap methods could also be applied to other imaging data such as functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG). McIntosh and Lobaugh (2004) demonstrate the utility of resampling techniques (including the bootstrap) to assess the relative importance and reliability of latent variables in a partial least squares analysis of neuroimaging data. McCubbin et al. (2008) have devised a nonparametric hypothesis test based on bootstrap that aims to balance power and significance to improve signal detection performance when signal-to-noise ratio is inherently low, as in the case of prenatal MEG. Thus, in a variety of settings where conventional statistics are not appropriate or useful, the bootstrap is a simple and versatile technique that can be used to devise special-purpose statistical tests for the immediate application.

Our results suggest that in ERP, a percentile bootstrap can be a useful means of evaluating the status of amplitude effects as markers of cognitive processes in an individual patient, providing certain conditions apply. The technique may be more suited to ERP phenomena with clearly defined, stereotyped peaks, which are probably more characteristic of early perceptual processing. The degree of significance (i.e., *p*-value) at the group-level analysis may be a useful guideline to deciding a priori whether a certain ERP amplitude effect will permit conclusions in an individual patient by use of the bootstrap method of analysis.

Manuscript received 11 November 2010
Revised manuscript received 18 November 2011
Revised manuscript accepted 19 November 2011

REFERENCES

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, *8*(6), 551–565.

Botzel, K., Schulze, S., & Stodieck, S. R. (1995). Scalp topography and analysis of intracranial sources of face-evoked potentials. *Experimental Brain Research*, *104*(1), 135–143.

Caryl, P. G., Golding, S. J. J., & Hall, B. J. D. (1995). Interrelationships among auditory and visual cognitive tasks: An event-related potential (ERP) study. *Intelligence*, *21*(3), 297–326.

Charest, I., Pernet, C. R., Rousselet, G. A., Quinones, I., Latinus, M., Fillion-Bilodeau, S., et al. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, *10,* 127.

Crawford, J. R., & Garthwaite, P. H. (2006). Methods of testing for a deficit in single-case studies: Evaluation of statistical power by Monte Carlo simulation. *Cognitive Neuropsychology*, *23*(6), 877–904.

Crawford, J. R., & Garthwaite, P. H. (2007). Comparison of a single case to a control or normative sample in neuropsychology: Development of a Bayesian approach. *Cognitive Neuropsychology*, *24*(4), 343–372.

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, *27*(3), 245–260.

Dalrymple, K. A., Oruc, I., Duchaine, B., Pancaroglu, R., Fox, C. J., Iaria, G., et al. (2011). The anatomic basis of the right face-selective N170 IN acquired prosopagnosia: A combined ERP/fMRI study. *Neuropsychologia*, *49*(9), 2553–2563.

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman Hall.

Eimer, M. (1996). ERP modulations indicate the selective processing of visual stimuli as a result of transient and sustained spatial attention. *Psychophysiology*, *33*(1), 13–21.

Eimer, M. (1998). Mechanisms of visuospatial attention: Evidence from event-related brain potentials. *Visual Cognition*, *5*(1–2), 257–286.

Eimer, M., & McCarthy, R. A. (1999). Prosopagnosia and structural encoding of faces: Evidence from event-related potentials. *Neuroreport*, *10*(2), 255–259.

Fabiani, M., Gratton, G., Corballis, P. M., Cheng, J., & Friedman, D. (1998). Bootstrap assessment of the reliability of maxima in surface maps of brain activity of individual subjects derived with electrophysiological and optical methods. *Behavior Research Methods, Instruments, & Computers*, *30*(1), 78–86.

Gehring, W. J., & Fencsik, D. E. (2001). Functions of the medial frontal cortex in the processing of conflict and errors. *Journal of Neuroscience*, *21*(23), 9430–9437.

Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, *44*(6), 913–917.

Itier, R. J., & Taylor, M. J. (2004). N170 or N1? Spatiotemporal differences between object and face processing using ERPs. *Cerebral Cortex*, *14*(2), 132–142.

Jacques, C., d'Arripe, O., & Rossion, B. (2007). The time course of the inversion effect during individual face discrimination. *Journal of Vision*, *7*(8), 3.

Krigolson, O. E., Heinekey, H., Kent, C. M., & Handy, T. C. (2012). Cognitive load impacts error evaluation within medial-frontal cortex. *Brain Research, 1430*, 62–67.

Krigolson, O. E., & Holroyd, C. B. (2006). Evidence for hierarchical error processing in the human brain. *Neuroscience*, *137*(1), 13–17.

McCubbin, J., Yee, T., Vrba, J., Robinson, S. E., Murphy, P., Eswaran, H., et al. (2008). Bootstrap significance of low SNR evoked response. *Journal of Neuroscience Methods*, *168*(1), 265–272.

McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *NeuroImage*, *23*(Suppl. 1), S250–S263.

Miltner, W.H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798.

Nagamatsu, L. S., Liu-Ambrose, T. Y., Carolan, P., & Handy, T. C. (2009). Are impairments in visual–spatial attention a critical factor for increased falls risk in seniors? An event-related potential study. *Neuropsychologia*, *47*(13), 2749–2755.

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *Journal of Neuroscience*, *26*(35), 8965–8975.

Philiastides, M. G., & Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex*, *16*(4), 509–518.

Rousselet, G. A., Gaspar, C. M., Pernet, C. R., Husk, J. S., Bennett, P. J., & Sekuler, A. B. (2010). Healthy aging delays scalp EEG sensitivity to noise in a face discrimination task. *Frontiers in Psychology*, *1,* 19.

Rousselet, G. A., Husk, J. S., Bennett, P. J., & Sekuler, A. B. (2005). Spatial scaling factors explain eccentricity effects on face ERPs. *Journal of Vision*, *5*(10), 755–763.

Rousselet, G. A., Husk, J. S., Bennett, P. J., & Sekuler, A. B. (2008). Time course and robustness of ERP object and face differences. *Journal of Vision*, *8*(12), *3*, 1–18.

Rousselet, G. A., Husk, J. S., Pernet, C. R., Gaspar, C. M., Bennett, P. J., & Sekuler, A. B. (2009). Age-related delay in information accrual for faces: Evidence from a parametric, single-trial EEG approach. *BMC Neuroscience*, *10,* 114.

Vizioli, L., Foreman, K., Rousselet, G. A., & Caldara, R. (2010). Inverting faces elicits sensitivity to race on the N170 component: A cross-cultural study. *Journal of Vision*, *10*(1), *15*, 11–23.

Webb, S. J., Jones, E. J., Merkle, K., Murias, M., Greenson, J., Richards, T., et al. (2010). Response to familiar faces, newly familiar faces, and novel faces as assessed by ERPs is intact in adults with autism spectrum disorders. *International Journal of Psychophysiology*, *77*(2), 106–117.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Amsterdam, The Netherlands: Elsevier/Academic Press.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*(3), 254–274.