PSYCHOPHYSIOLOGY

Psychophysiology, 54 (2017), 123–138. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/psyp.12629

Assessing the internal consistency of the event-related potential: An example analysis

NINA N. THIGPEN,[a] EMILY S. KAPPENMAN,[b] AND ANDREAS KEIL[a]

[a]Center for the Study of Emotion & Attention, University of Florida, Gainesville, Florida, USA
[b]UC Davis Center for Mind & Brain, University of California, Davis, California, USA

Abstract

ERPs are widely and increasingly used to address questions in psychophysiological research. As discussed in this special issue, a renewed focus on questions of reliability and stability marks the need for intuitive, quantitative descriptors that allow researchers to communicate the robustness of ERP measures used in a given study. This report argues that well-established indices of internal consistency and effect size meet this need and can be easily extracted from most ERP datasets, as demonstrated with example analyses using a representative dataset from a feature-based visual selective attention task. We demonstrate how to measure the internal consistency of three aspects commonly considered in ERP studies: voltage measurements for specific time ranges at selected sensors, voltage dynamics across all time points of the ERP waveform, and the distribution of voltages across the scalp. We illustrate methods for quantifying the robustness of experimental condition differences, by calculating effect size for different indices derived from the ERP. The number of trials contributing to the ERP waveform was manipulated to examine the relationship between signal-to-noise ratio (SNR), internal consistency, and effect size. In the present example dataset, satisfactory consistency (Cronbach's alpha > 0.7) of individual voltage measurements was reached at lower trial counts than were required to reach satisfactory effect sizes for differences between experimental conditions. Comparing different metrics of robustness, we conclude that the internal consistency and effect size of ERP findings greatly depend on the quantification strategy, the comparisons and analyses performed, and the SNR.

Descriptors: Reliability, Event-related potentials, Cronbach's alpha, Internal consistency, Effect size, Signal-to-noise ratio

ERPs represent large-scale brain electric fields that are time-locked to an event. They are noninvasively recorded from the scalp and have been used to investigate brain processes for more than half a century (Luck, 2014). ERPs have also been discussed as potential biomarkers for a variety of psychiatric and neurological disorders (Foti, Kotov, & Hajcak, 2013; Light & Swerdlow, 2015; Luck et al., 2011; Perez, Swerdlow, Braff, Näätänen, & Light, 2014) and as indices of individual differences in nonclinical samples (Anokhin et al., 2001; Cohen & Polich, 1997). An ERP can be regarded as a spatiotemporal matrix, often recorded from many scalp locations, and containing time-varying voltage information at high temporal resolution. Numerous indices can be extracted from this spatiotemporal matrix using different quantification methods. Some indices are univariate in nature, such as the latency of a given component, the mean amplitude across a time window, or area measurements of the amplitude for a given component at a given sensor location (Kappenman & Luck, 2012). Others are multivariate, such as the topographical distribution of voltages across the scalp or the temporal sequence of components in a waveform (Dien, Spencer, & Donchin, 2004; Spencer, Dien, & Donchin, 1999). Given their rich potential for answering questions in psychophysiology, potential clinical applications, and the myriad techniques used to quantify ERP indices, a discussion of their psychometric properties is becoming increasingly important.

Recently, discussions about replicability in the cognitive and neural sciences have arisen, particularly regarding the reliability of ERP measures (Keil et al., 2014). The need for establishing the psychometric properties of ERP measures (such as reliability) is obvious when authors are interested in quantifying interindividual differences, especially in the context of clinical and translational work. Quantitative indices of robustness and consistency, however, are also desirable for experimental studies comparing ERP metrics under different conditions, typically using within-participant comparisons. The reliability of a given ERP effect depends on a number of factors, including the recording hardware and sensors (affecting

This research was supported by National Institute of Mental Health Grant R01MH097320 and by a grant from the Office of Naval Research N00014-14-1-0542.

Address correspondence to: Nina Thigpen, Center for the Study of Emotion & Attention, University of Florida, PO Box 112766, Gainesville, FL 32611, USA. E-mail: nthigpen@ufl.edu

the overall signal quality in the raw EEG), how the dependent variable was derived from the spatiotemporal ERP matrix (the quantification method), and how much error variance (noise) affected the desired ERP signal, which can be measured by computing the signal-to-noise ratio (SNR). Another point affecting replicability of experimental reports is the sensitivity of a given ERP index to differences between conditions, readily quantified by computing effect size. The range of acceptable SNRs, effect sizes, and reliability indices will vary by study goals, experimental paradigm, and the specific ERP component examined in the study.

In the present report, we address the issue of reliability of ERP measures not by suggesting recommended parameters for ERP studies, but by providing example analyses of SNR, internal consistency, and effect size that can be readily applied to any ERP dataset. We discuss how authors of within-participant studies may quantify and document the within-study reproducibility of selected ERP metrics using the variability across experimental conditions. We also illustrate the consequences of different quantification methods, and compare the reliability and robustness (measured by the effect size) of different types of dependent variables.

Considerations Regarding SNR

SNR quantifies the strength of a signal of interest in the presence of noise (Teplan, 2002). Often, SNR is defined as a function of signal (S) and noise (sigma) in a single trial, modulated by the square root of the number of trials *n*: SNR = sqrt(*n*) * S/sigma. Thus, SNR increases logarithmically with the number of trials[1] averaged to produce an ERP (Handy, 2005). SNR is important in determining the robustness and replicability of a given ERP finding, and recommendations regarding the design of ERP studies are therefore often based on SNR. For example, Luck (2014) recommended designing ERP studies to contain a trial count high enough to reach a SNR of 10. A comprehensive discussion of trial count recommendations is outside the scope of the present paper, but it should be noted that many recommendations do not contain systematic quantitative or psychometric analyses in their support, discussed in the paper by Kappenman & Luck (2012). Instead, the present paper focuses on simple methods for quantifying and documenting data quality that can inform researchers regarding the suitability of the ERP signal used as a dependent variable in a given study.

Quantifying ERP Robustness Using Internal Consistency

The present study examines an important facet of reliability known as internal consistency. Internal consistency refers to a measure's ability to quantify the same underlying construct or variable (here, ERP data) with different items or subvariables. ERP measures are considered internally consistent if the rank ordering of subjects remains stable for the extracted variable across different experimental conditions, trials, or sessions (Simons & Miles, 1990). Thus, internal consistency is considered particularly important in studies with between-participants factors, in which researchers aim to characterize individuals by means of the ERP component of interest.

Internal consistency of ERP measures is also desirable in studies interested in the effects of within-participant manipulations on a dependent ERP variable. To assess internal consistency of the ERP in a within-participant design, researchers may use condition-averaged ERPs to serve as "items" for Cronbach's alpha, as condition-averaged ERPs represent distinct samples drawn from the same population of trials. This metric answers questions regarding the consistency/reliability of the ERP that is attainable in a specific design, given the number of trials available: High consistency would indicate that the ERP is reliably seen across different averages obtained from the same participant. Another possibility includes randomly dividing all experimental trials into X number of arbitrary groups (ignoring experimental conditions), and reaveraging to form new ERPs to create items for Cronbach's alpha (Fabiani, Gratton, Karis, & Donchin, 1987; Handy, 2005). Although reaveraging randomly drawn trials is a feasible and informative approach, it does not allow quantification of robustness of experimental effects (as different conditions are averaged together) and may obscure changes of robustness associated with condition-specific ERP modulations.

In the present report, we address the issue of robustness and consistency of variables derived from ERPs by using the condition-averaged ERPs as the items for consistency analysis. The rationale for this is as follows: (a) condition averages are the data used for hypothesis testing, and assessing their reliability thus has higher relevance compared to analyses based on surrogate data; (b) condition averages are readily available and do not require resampling, thus lessening the burden on researchers; (c) similar ERP waveforms across conditions are typical for studies in which conditions differ only by a specific manipulation; (d) the experimental effects (in which the consistency of the ERP across conditions may be reduced) are typically confined to specific temporal regions, allowing the use of other temporal regions for analyzing the consistency of the overall waveform; and (e) the internal consistency among ERP variables derived from different conditions can be considered a necessary condition for treating the variables as indices of the same brain process, which is often assumed in ERP studies focusing on amplitude modulation between experimental conditions.

A potential limitation of this approach arises when determining the consistency of differences (e.g., values obtained by subtracting one experimental condition from another). Here, Cronbach's alpha might underestimate the robustness of the ERP measure, because consistency requires that a participant's ordinal position of the dependent variable is maintained across conditions. This requirement will not be met if participants vary strongly in their sensitivity to the experimental manipulation (i.e., to the conditions). Empirically, this question can be addressed by comparing internal consistency of components that are modulated versus components that are not modulated by the experimental manipulation to determine the consistency of the overall waveform, which is one of the approaches illustrated in the present study.

The Role of Quantification Techniques

Quantification techniques differ in their sensitivity to the quality of the data. For example, peak amplitude measures are more sensitive to high-frequency noise compared to mean amplitude measures (Luck, 2014). Conversely, applying different quantification techniques may also lead to different SNRs of the dependent variable extracted. Thus, quantification techniques such as measuring the

1. It should be noted that a logarithmic relationship between SNR and trial count is based on the assumption that random noise in an ERP waveform decreases as trials are added to an average. In this sense, only random noise (and not systematic noise) decreases as trials are added to an average, leading to a higher SNR. Systematic noise could include an increase in alpha power over the recording session, or effects arising from oculomotor activity that may be temporally correlated with stimulus onset or offset, and would not necessarily be diminished with higher trial counts.

peak voltage, or the mean voltage of a given component, may also affect the reproducibility of the findings of a given ERP effect. For instance, averaging time points and sensors for the ERP signal of interest is often thought to increase within-participant SNR and decrease error variance across participants (Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011; Pontifex et al., 2010). However, the choice of the temporal length and spatial extent of voltages to be averaged (or otherwise integrated) into dependent variables may sometimes seem arbitrary, which has led to discussions of peak picking and averaging as potential causes contributing to false positive findings (Dien, 2010; Keil et al., 2014). The contribution by **Luck and Gaspelin (2017)** illustrates this problem in greater detail.

In addition, an extensive discussion in the ERP literature has identified challenges associated with measures that focus on finding component peaks in general (Dien, Spencer, & Donchin, 2003; Donchin et al., 1977; Fabiani, Gratton, Corballis, Cheng, & Friedman, 1998; Spencer et al., 1999). For example, peaks may not capture the signal of interest, and may instead reflect brain processes that are shared between conditions and/or groups. Using difference waveforms is another common approach to address some of these problems (Kappenman & Luck, 2012). In this approach, replicability may be affected by the fact that SNR tends to be lower for difference waveforms. That is, the SNR of the difference waveform is typically lower than the parent waveforms. Thus, the present example analysis examines the effect size between experimental conditions as a measure of robustness with various quantification strategies, and discusses the relationship between these effect sizes and the SNR of the difference waveforms.

The Present Research

The goal of the present study is to provide a set of example analyses of SNR, internal consistency, and effect size in a typical ERP study, using metrics that are rapidly and easily computed. The ultimate goal of this approach is to stimulate more extensive use of quantitative reports of reliability and replicability in empirical reports (Keil et al., 2014). We address this question using a dataset involving pattern-onset ERPs in a feature-based visual selective attention task, containing four experimental conditions, each producing an ERP containing five well-known components (i.e., the P1, N1, P2, N2, P3). Based on a number of studies with this paradigm (see Harter & Aine, 1984; Hopf, Boelmans, Schoenfeld, Luck, & Heinze, 2004; McGinnis & Keil, 2011; Müller & Keil, 2004; Schoenfeld et al., 2007), we expected the spatiotemporal properties of the ERPs elicited in each experimental condition to be similar. Specifically, experimental condition-averaged ERPs are expected to differ during a brief time window from 190 to 220 ms poststimulus onset, over parietooccipital electrode locations (Anllo-Vento & Hillyard, 1996; Harter & Aine, 1984). In the literature, this time window is referred to as the selection negativity (SN), and has been shown to contain an amplitude enhancement for attended when compared to nonattended features. Thus, ERPs derived from the conditions in a feature-based attention paradigm are particularly suitable for demonstrating the use of Cronbach's alpha for quantifying internal consistency.

We quantified the internal consistency of selected variables derived from the pattern-onset ERP by measuring Cronbach's alpha, and the robustness of experimental effects (differences between conditions) by measuring effect sizes, expressed as R^2. In an attempt to illustrate the sensitivity of these approaches to SNR, we also varied the number of trials entering the ERPs used in the analyses. Solutions and example outcomes are presented for different quantification techniques, notably, averaging in the time domain (i.e., across time points) and averaging in the spatial domain (across electrodes), prior to statistical analysis. Finally, when using ERP data for hypothesis testing, researchers may be interested in three different types of variables: (1) the ERP topography at a given point in time may be relevant to testing a hypothesis regarding spatial extent of a voltage change, (2) the shape of the ERP waveform at a given sensor or sensor group may be of interest when testing hypotheses regarding the temporal evolution of neurocognitive processes, and (3) the voltage amplitude at specific time points and electrodes may be used to examine hypotheses regarding the differences in neural population activity between experimental conditions. Accordingly, for each of these variable types, we demonstrate simple methods for estimating consistency and effect size, and examine their relation for different trials counts and quantification techniques.

Method

Participants

Nineteen healthy volunteers (12 females; mean age 18.6, *SD* 0.9; 1 left-handed) participated in the experiment in exchange for course credit. Participants were excluded if their response accuracy fell below two standard deviations from the mean, which applied to two participants. All participants gave written informed consent prior to participating. The Institutional Review Board of the University of Florida, in line with the Declaration of Helsinki, approved all procedures.

Stimuli and Procedure

Stimuli consisted of four sinusoidal gratings filtered with a Gaussian envelope (i.e., Gabor patches). All Gabor patches were composed of grayscale gratings and were presented against a gray background with the same mean luminance (31 cd/m^2) as the Gabor patches (see Figure 1). The four stimuli varied with respect to two task-relevant features: orientation and spatial frequency. Stimulus orientation was manipulated by rotating the Gabor patch grating relative to a vertical axis (1.5° or 358.5°). Stimulus spatial frequency was either 1.33 or 1.78 cycles per degree, at a visual angle of 4.5°, achieved by seating participants 140 cm from a 23" 3D LED monitor (Samsung LS23A950) set to a vertical refresh rate of 120 Hz.

Participants performed a feature-based visual selective attention task, which involved discriminating target stimuli (a Gabor patch defined by a combination of orientation and spatial frequency) from nontarget stimuli. The experimental session was organized into 12 experimental blocks, with the target patch varying between blocks. Prior to a given trial block, participants were presented with a target stimulus in the middle of the screen (e.g., Stimulus A in Figure 1), and asked to memorize the two features defining this particular target, which included a specific orientation and spatial frequency. Once participants reported familiarization with the target stimulus, they completed a block containing 40 trials. Each trial contained one of the four stimuli (shown in Figure 1), presented at the center of the screen for 66.7 ms. Participants were instructed to indicate whether the presented stimulus matched or did not match the target stimulus for that block. Participants responded with their dominant hand by pressing one arrow key of a standard keyboard when a stimulus identical to the target stimulus appeared, and the

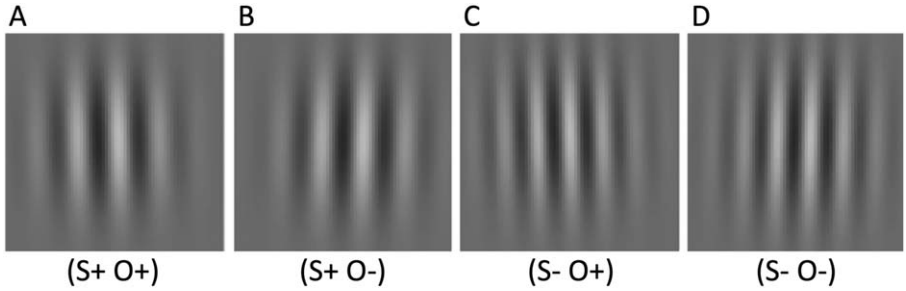

Figure 1. The stimuli used in the present study. Gabor gratings were shown against a gray background, and thus onset of a grating did not change the overall luminance of the display. The four stimuli differed on two feature dimensions: orientation (O) and spatial-frequency (S). Labels on the bottom refer to the example in the text, illustrating the changing role of each stimulus in different experimental blocks:

other arrow key when any nontarget stimulus appeared. The keyboard was placed in a comfortable location and could easily be operated by all participants, and the mapping of the arrow keys to target/nontarget conditions was counterbalanced across participants. Between stimulus presentations, a fixation circle occupying 0.5° of visual angle was present for an interval varying between 1.5–2.1 s. If participants did not press either response button during this interval, it was counted as an incorrect response. A new target stimulus was assigned at the start of each block. Participants were allowed breaks as needed in between blocks. Both the order of stimuli presented within a block and the order of blocks was fully randomized. Participants were instructed to avoid head movements and to maintain gaze on the central fixation circle.

After data collection, each trial was assigned to one of four experimental conditions, contingent on the block's target stimulus: trials containing (1) stimuli that matched the target's spatial frequency and orientation (S+O+); (2) stimuli that matched the target's spatial frequency, but not orientation (S+O−); (3) stimuli that matched the target's orientation, but not spatial frequency (S−O+); and (4) stimuli that did not match the target in either spatial frequency or orientation (S−O−). A total of 120 trials was presented in each of the four conditions.

Behavioral Data

Participants' accuracy and response time was calculated across blocks separately for each condition. This included the percentage of correctly identified targets (hits), incorrect responses to targets (misses), incorrect responses to a nontarget (false alarms), and correct responses to a nontarget (correct rejections). To ensure the four stimuli were comparable in their discriminability, a 2 × 2 repeated measures analysis of variance (ANOVA) was conducted on both the hit rates and response times observed with the four different stimuli, with factors of spatial frequency and orientation.

Data Acquisition

EEG data were recorded continuously with a 129-channel Geodesic Sensor Net (Electrical Geodesic, Eugene, OR) connected to a high-input impedance (>200 MOhms) amplifier. Electrodes were evenly spaced across large areas of the head, including facial and neck regions (see Figure 2). Impedance for each electrode was kept below 60 kOhms, and the vertex electrode (Cz) was used as the recording reference. All channels were digitized at a rate of 500 Hz and filtered online using a Butterworth low-pass filter with a 3 dB

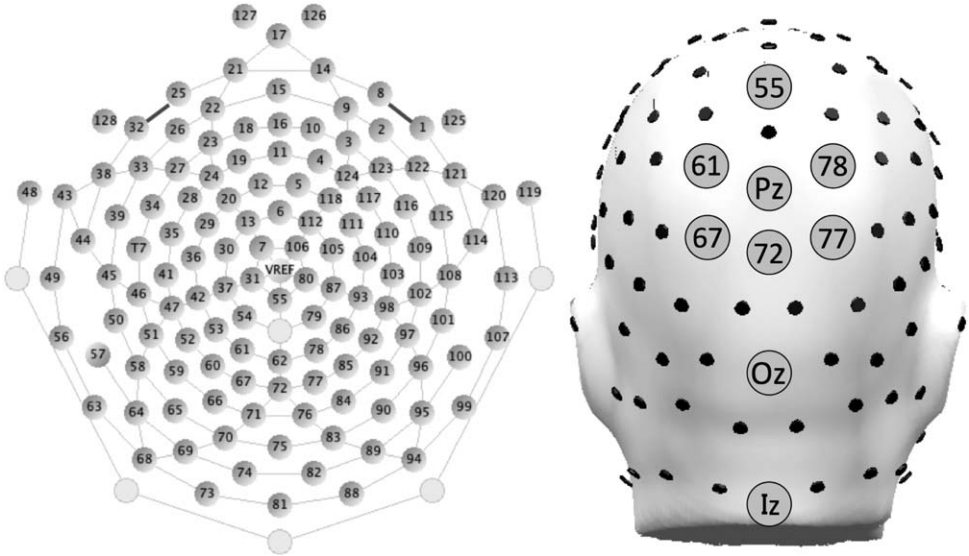

Figure 2. The 128-channel HydroCel Geodesic Sensor Net used in the present study. Left: Projection of the full electrode layout (VREF marks the location of the common reference sensor, located at site Cz of the International 10-20 system). ERP reliability was assessed at all sensors. In addition, specific clustered subsets of sensors were examined (right). For analyses of the effects of spatial averaging, sensor clusters expanded radially from either Pz (sensor 62) or Oz (sensor 75).

point (cutoff) at 200 Hz. All further data processing was done offline.

Trial Segmentation, Filtering, and Artifact Handling

Continuous EEG data were digitally filtered offline using a 2nd order Butterworth high-pass filter having a 3 dB point at .15 Hz[2], as well as a 12th order Butterworth low-pass filter with a 3 dB point at 40 Hz. Eye movement artifacts were detected and corrected using an artifact correction method based on linear regression performed on residuals, implemented in the BioSig suite of MATLAB functions (Schlögl et al., 2007; Vidaurre, Sander, & Schlögl, 2011). It creates a linear model of the data based on representative ocular events, in which the contribution of electrooocular processes to the EEG measured at each time point is estimated and removed through subtraction of the weighted electrooculogram (EOG). This procedure bears the risk that brain-related activity is removed if it shares spatial and temporal variance with EOG events (compare Gratton, Coles, & Donchin, 1983, for a different approach). In the present dataset, however, rerunning the preprocessing without eye correction resulted in suppressed, not in augmented, ERP amplitude. Following EOG correction, segments were extracted from the continuous EEG, with each segment having a duration of 1,000 ms (200 ms before and 800 ms after stimulus onset).

These segments were submitted to a semiautomated artifact detection procedure designed for multichannel electrophysiology, which is based on distributions of trial and channel statistics (Junghöfer, Elbert, Tucker, & Rockstroh, 2000). First, specific channels that were bad throughout the experimental session were detected with voltage data given relative to the original recording reference (i.e., Cz). That is, channels that fell above a 2.5 *SD* threshold with respect to the median of three distributions calculated across all trials (amplitude, standard deviation, and gradient) were interpolated across all time points using spherical spline functions (Junghöfer et al., 2000). Data at eliminated channels were replaced with a statistically weighted spherical spline interpolation from the full channel set (Junghöfer, Elbert, Leiderer, Berg, & Rockstroh, 1997).

In a next step, based on the offline average reference, distinct sensors from individual trials were also excluded and interpolated when located in the tails (2.5 *SD* above the median) of the distribution of their absolute amplitude, maximum standard deviation, and gradient, calculated by integrating across the time points in each trial. Trials in which interpolated channels were clustered in one scalp region (quantified as described in Peyk, DeCesarei, & Junghöfer, 2011) and trials with fewer than 103 good channels were excluded entirely. Only trials with correct responses were retained for final ERP averaging, leading to an overall mean of 78.6 trials included per condition (*SD* = 14.3, range = 60–103). On average, 24% of trials were rejected due to artifact, and 11% of trials were not used for ERP analysis due to incorrect behavioral responses. The target condition (S+O+) included a mean of 77.2 trials across participants (*SD* = 15.5, range = 61–99); condition S+O- included a mean of 78.1 (*SD* = 14.9, range = 66–102); condition S-O+ included a mean of 79.2 trials (*SD* = 14.0, range = 61–103); and condition S-O- included a mean of 80.1 trials (*SD* = 16.7, range = 65–103).

Analysis of Experimental Effects: SN

To ensure that the dataset used for reliability analyses was representative, we established the extent to which the ERP waveforms in the present study replicated a large body of earlier work in this area (e.g., Anllo-Vento & Hillyard, 1996). Specifically, we expected to observe a greater posterior negativity for stimuli with target features compared to stimuli with no target features, during the time window of the selection negativity (typically 160–280 ms). To this end, a posterior cluster of electrodes was formed around Pz and its superior and inferior nearest neighbors (containing electrodes 54, 55, 61, 62, 72, 75, 78, 79, 81, as shown in Figure 2), chosen based on earlier research with this paradigm (e.g., Keil & Müller, 2010). Then, considering the waveform differences seen in the grand mean as well as the previous studies discussed above (McGinnis & Keil, 2011; Müller & Keil, 2004), the mean voltage amplitude was extracted across this sensor cluster and across time windows representing early and late selection negativity (178–234 ms and 236–292 ms, respectively, in line with the studies cited above). A repeated measures ANOVA was conducted on each of the early and late mean amplitudes with factors of time (early SN, late SN), spatial frequency (match vs. nonmatch with the target), and orientation (match vs. nonmatch with the target; Keil & Müller, 2010).

SNR

SNR was determined at each sensor for the components P1, N1, P3, and selection negativity using averages based on varying numbers of trials (10 through 80 trials in steps of 10). Specifically, SNRs were calculated for each participant by dividing the voltage measurement (peak amplitude) obtained for each component by the peak across the baseline variance (from −100 to 0 ms). The peak amplitude of each component was determined by taking either the maximum or minimum voltage amplitude (for positive or negative components, respectively), across the time window centered around the grand mean component peak (defined as P1: 100–130 ms, N1: 160–190 ms, P3: 300–330 ms, and SN: 190–220). The mean amplitude was calculated as the average voltage within these same time windows. SNR values were then averaged across participants. If not otherwise indicated, figures display SNRs for the target condition.

Reliability Analyses

Reflective of the spatiotemporal nature of ERP data, reliability was assessed for various combinations of time windows and sensor clusters extracted from the ERP matrices. Many empirical studies do not form dependent variables based on peak amplitude measures at single sensors, but use voltage averages across multiple electrodes and time points for hypothesis testing (Fabiani, Gratton, Karis, & Donchin, 1987). The effects of this strategy were examined here by systematically averaging across increasing numbers of time points around component peaks and across increasing numbers of electrodes, prior to calculating internal consistency. Internal consistency was quantified with Cronbach's alpha, a coefficient representing the consistency of items (variables or repetitions) across observations (e.g., participants). The formation of items is described for each example analysis in the results. Cronbach's

2. The P3 ERP component can be distorted by high-pass filters (Duncan-Johnson & Donchin, 1979). To ensure that the present filter sitting did not significantly distort the P3 component in the current study, we reanalyzed our data after preprocessing with a 2nd order Butterworth high-pass filter having a 3 dB point set at .1 Hz. As expected, amplitude of either component examined here was unaffected.

Table 1. *Behavioral Data*

	Accuracy		Response time	
Category	Mean	*SD*	Mean	*SD*
Hit	.89	.9	764	161
Miss	.11	.9	805	288
False alarm	.6	.8	933	351
Correct rejection	.96	.8	714	158

$N = 19$.
Note. Mean and standard deviation for accuracy and response time in milliseconds.

alpha has been widely used to test the response similarities of items on a questionnaire across observations, and is considered to represent acceptable reliability when the coefficient is above .70 (Cronbach, 1951). More recently, Hinton, McMurray, and Brownlow, 2004 have suggested that Cronbach's alpha exceeding .90 indicates excellent internal consistency, alphas between .70 and .90 indicate high internal consistency, and alphas from .50 to .70 indicate moderate internal consistency, whereas a coefficient below .50 is considered poor.

Effect Size Analyses

Cronbach's alpha could not be used to examine the internal consistency of difference waveforms in the present design, because the difference waveforms of interest contained the same (target) condition and thus represented linear combinations of each other, which violates the independence-of-items assumption required for calculating Cronbach's alpha. Instead, to quantify effect size of the well-established selection negativity difference, a trend (F contrast) analysis was performed using a general linear model procedure (Rosnow, Rosnow, & Rosenthal, 1996), with weights based on the hypothesis that feature-based attention increases the SN amplitude with the number of attended features (Harter & Aine, 1984; Hopf et al., 2004; McGinnis & Keil, 2011; Müller & Keil, 2004; Schoenfeld et al., 2007): Across the selection negativity time window (i.e., 160–280 ms), the four conditions were weighted according to their overlap with the target condition. The target condition itself was weighted the lowest (expected to show greatest selection negativity), the two conditions with one feature in common with the target (either orientation or spatial frequency) were weighted intermediate, and the no-matching features condition was weighted the highest (resulting in condition weights of −2, .5, .5, and 1, respectively). The effect size of the resulting F contrast is readily expressed as R^2, which reflects the proportion of trend-related variance relative to the total variance (i.e., trend variance plus unique error variance). Traditionally, R^2 estimates of effect size are assigned to three levels: .14, a small effect; .39, a medium effect; and .59, a large effect (Cohen, 1992). The further computational steps taken to address different aspects of effect size are detailed below, in the Results section.

Results

Behavioral Data

Participants performed the task with high accuracy ($M = 89\%$ correct across all trials, $SD = 9\%$), and response times as expected for this task ($M = 760$ ms, $SD = 160$ ms). The accuracy and response time data are shown in Table 1. The repeated measures ANOVA for both hit rate and reaction time showed no significant effect for

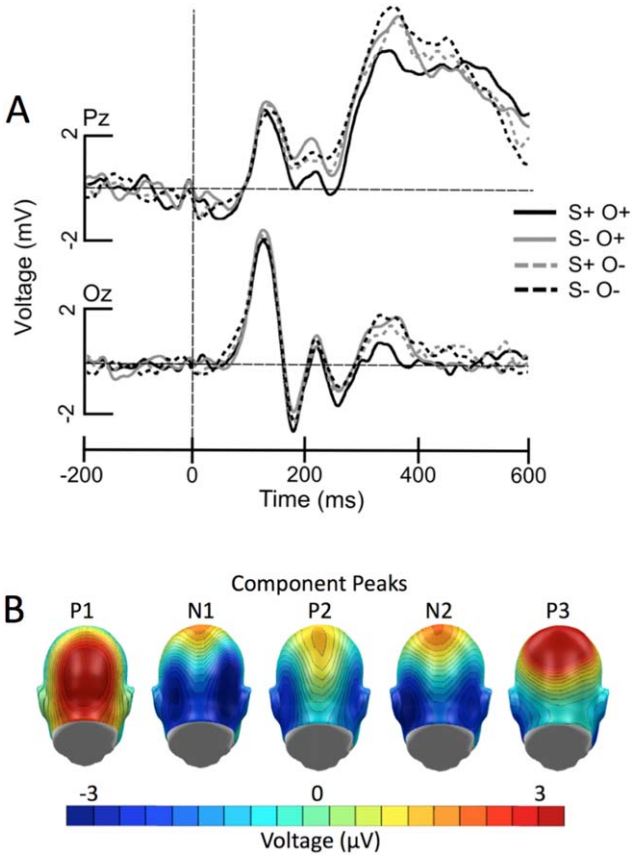

Figure 3. Pattern-onset visual ERPs: A: Grand mean ($N = 19$) voltage time course at sensors Pz and Oz (see Figure 2) for the four experimental conditions, (S+O+), (S+O-), (S-O+), (S-O-). B: Grand mean voltage topographical distributions shown for the five major component peaks: P1 (120 ms), N1 (176 ms), P2 (216 ms), N2 (256 ms), P3 (320 ms).

spatial frequency or orientation, suggesting that the four different Gabor patches did not differ in their discriminability or saliency as target stimuli.

ERP Morphology and Condition Differences

Five well-known ERP components (i.e., the P1, N1, P2, N2, P3) showed latencies and topographies typically consistent with previous studies of pattern-onset ERPs (see Figure 3), with component peaks in the grand mean centered at 120, 176, 210, 256, and 340 ms poststimulus, respectively. A standard analysis of experimental effects (differences between voltage amplitudes) obtained in the different conditions was conducted to document the extent to which the present dataset replicates known effects of feature-based attention.

Condition differences were only prominent during the selection negativity time window, paralleling previous work on feature-based attention with multifeature stimuli (Anllo-Vento & Hillyard, 1996; Martinez et al., 1999; Keil & Muller, 2010; McGinnis & Keil, 2011): Selection negativity was observed for attended features, over parietooccipital sensors, at latencies between 178 and 292 ms poststimulus. As shown in Figure 3, the selection negativity was most pronounced when comparing the target condition (S+O+) to the condition with no target features (S-O-). That is, stimuli containing target features evoked larger negative deflections compared to stimuli with fewer target features, during the selection negativity time window (which encompasses the N1, P2,

and N2 components). As expected, an ANOVA showed main effects of orientation, $F(1,18) = 6.95$, $p < .05$, $\eta_p^2 = .27$, and spatial frequency, $F(1,18) = 8.96$, $p < 0.01$, $\eta_p^2 = .33$, which both reflected larger negative deflections for stimuli with target features. There was no main effect or interaction effect involving time (i.e., early vs. late selection negativity). Furthermore, the two factors corresponding to attended features (orientation and spatial frequency) did not interact, which replicates previous work interpreting this finding to indicate additive effects of feature dimensions on the selection negativity. In summary, analyses of condition differences using traditional ANOVA suggested that the present dataset is consistent with previous work in terms of direction and size of experimental effects. Because condition differences were confined to a specific temporal region and absent during the remainder of the ERP epoch, this paradigm was considered particularly suitable for the purpose of internal consistency analysis, using conditions as items.

SNR

As expected, the SNR increased as a function of the number of trials included in the average ERP waveform. With 10 averaged trials, SNR for component peaks relative to the peak of the baseline variance tended to be around three, and increased linearly as trial count increased logarithmically (see Figure 4). Recommended SNRs (10 and above, Luck, 2014) for component peaks of P1, N1, and P3 were reached with 40 trials. Additional averaging led to SNRs around 20, showing topographical distributions consistent with the distribution of voltage maxima. As expected, the selection negativity difference waveform was associated with significantly lower SNR, as shown in Figure 4.

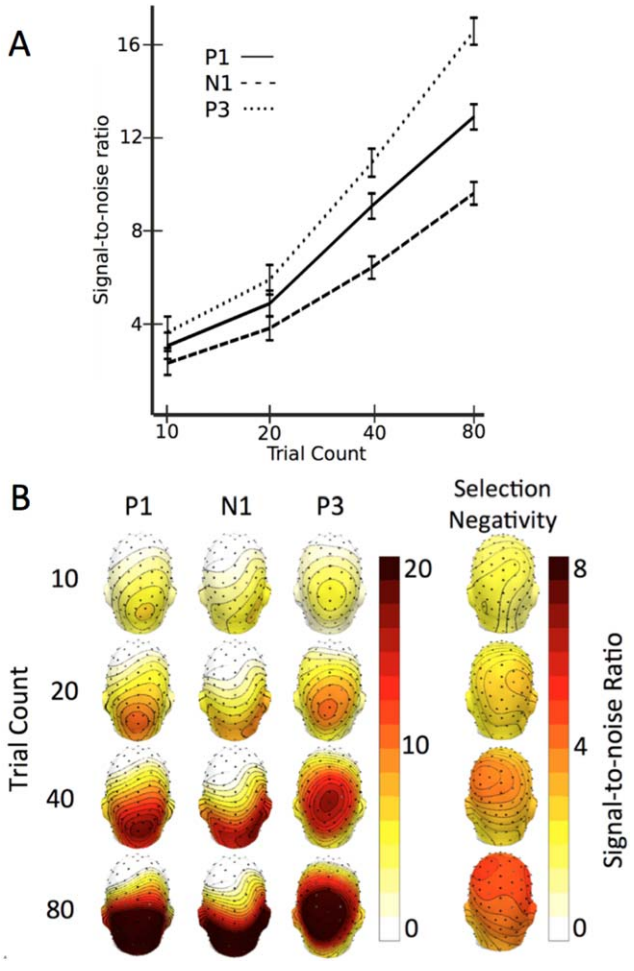

Figure 4. A: Signal-to-noise ratio (SNR) as a function of trial count for three ERP components. Error bars represent the standard error of the mean. B: Topographical distribution of the SNR of three example ERP components, and the selection negativity (right panel). SNR was calculated for each participant and sensor for varying trial counts.

Reliability Analyses

Reliability of peak voltage at individual time points and sensors. One of the most common forms of ERP analysis is the statistical comparison of voltage measurements taken at a given sensor and time point. To illustrate how internal consistency of these measurements can be easily assessed, and to document the range of possible outcomes of such an analysis, Cronbach's alpha was calculated for individual ERP voltages at each sensor and time point, using the four conditions as items and 19 participants as observations. This analysis yielded consistency estimates for each of the 129 sensors at all 501 epoch sample points (i.e., 1,002 ms), for a total of 64,629 alphas. These calculations were repeated with varying numbers of trials included in the averaged ERP. The first six trial counts were based on subsets of 10, 20, 30, 40, 50, and 60 trials per participant, respectively, in each condition. The 7th calculation included all artifact-free trials, which included a median of 80 trials per participant (range: 60–103 trials).

Figure 5 shows the topographical distribution of Cronbach's alpha for the peak across the baseline variance (−100 to 0) and each component peak voltage, calculated for different trial counts. Peak latencies were determined on the basis of the grand mean ERP waveform, a widely used practice in ERP studies. For all five ERP component peaks analyzed, high Cronbach's alpha values (i.e., exceeding an alpha of .7) were observed with relatively low trial counts (20 trials), but only at scalp locations at which the respective signal was pronounced. For instance, peak voltages of P1 and N1 displayed excellent (> .9) internal consistency with as few as 20 trials in the posterior portion of the scalp, at sensors surrounding site Oz of the International 10-20 system. Later, components P2, N2, and P3 similarly displayed high internal consistency with few trials at scalp locations associated with their voltage maximum. Including more trials into the average was associated with more widespread internal consistency of peak voltages, across all component peaks examined. At a trial count of 40, internal consistency reached levels of .8 or greater at 127 out of 129 sensors, for all component peak voltages examined, representing high internal consistency. Thus, experimental conditions used as replications (items) displayed high consistency in estimating the underlying dimension (here, the peak voltage at a given sensor) at trial counts exceeding 40, over widespread scalp regions. Note that Cronbach's alpha as used above is easily determined for voltage scores extracted from individual participants (n) and a given number of conditions (k), arranged in one or more $n \times k$ matrices, using a wide range of statistics or computing software packages.

Reliability of the voltage topography for each time point. Researchers interested in the internal consistency of the voltage topography (the distribution of voltages across the electrode array) may also employ Cronbach's alpha. In the present example, observations were voltages for all 129 sensors for 19 participants, and items were the four experimental conditions, resulting in a

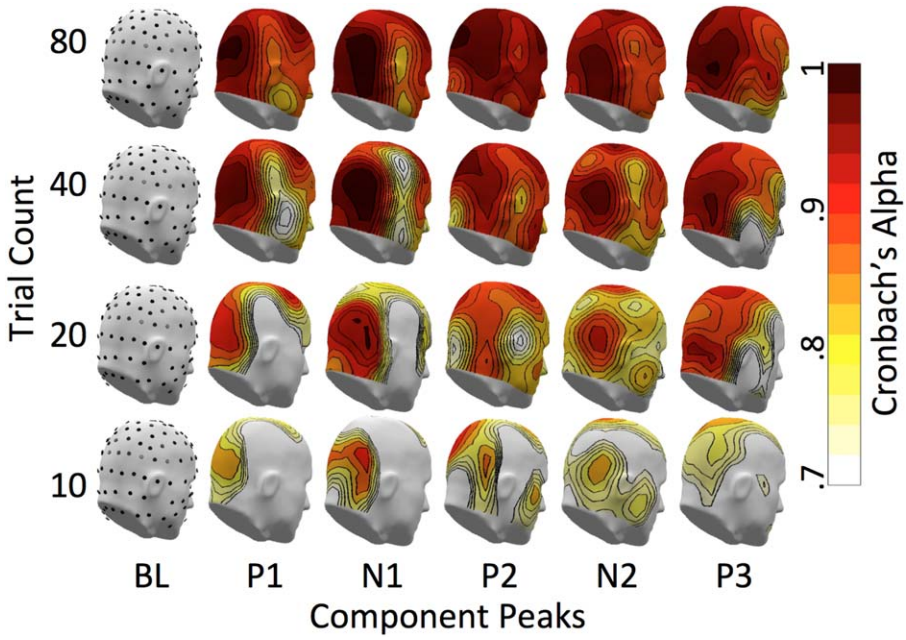

Figure 5. Reliability of peak measurements for five ERP components at each electrode location, projected to the scalp. Baseline reliability (1*st* column) was measured as the peak amplitude across 100 ms of baseline. Cronbach's alpha values are color coded, with red indicating greater consistency. Note the greater spatial spread of high Cronbach's alpha values with increasing trial count.

2,451 × 4 matrix. One such matrix was created for each ERP time point, and each matrix produced one Cronbach's alpha value. Thus, a time series of internal consistency estimates resulted, reflective of the consistency of the individual topographies across the four experimental conditions (see Figure 6). Across all trial counts, internal consistency was low during the baseline segment, as expected. The internal consistency of the topographical distribution increased with the rising slope of the P1 (at 110 ms), and again strongly varied with trial count. Cronbach's alpha values exceeded the threshold for high internal consistency (i.e., Cronbach's alpha $> .7$) when averaging 30 trials. Excellent internal consistency (Cronbach's alpha $> .9$) was observed between 115 and 620 ms poststimulus for 40 or more trials. For the duration of this time window, moderate reliability (i.e., values near .68) was observed with 20 averaged trials, and low reliability was found with 10 trials (i.e., Cronbach's alpha $< .5$). The time range of the selection negativity (178–292 ms) was characterized by a sharp transient decrease in cross-condition consistency. Experimental conditions (used as variables) systematically differed in amplitude and topography during this time range in the present task. This added variability as associated with decreased internal consistency across conditions, while still being at levels of satisfactory to excellent consistency. In combination with the analysis of individual voltage scores obtained from individual electrodes, this result highlights that a comparison of consistency indices across components may yield converging information about the reproducibility of the ERP measures of interest, across conditions.

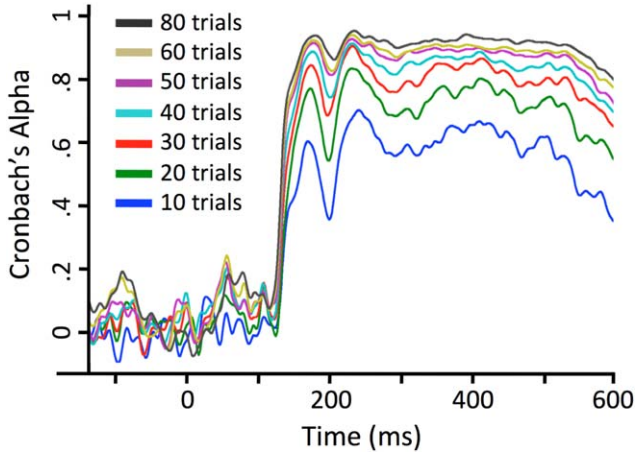

Figure 6. Reliability of the topographical distribution of pattern-evoked ERPs. Cronbach's alpha values represent the internal consistency of the voltage distribution across the scalp, computed at each individual time point, shown for subsets of trials. A sharp increase in internal consistency of the voltage topography is visible with the onset of the P1 component, around 120 ms.

Reliability of the voltage time course for each sensor. A final example analysis quantified the internal consistency of the ERP time course (the entire voltage time series representing the poststimulus ERP waveform), for each sensor. For each Cronbach's alpha calculation, observations were 400 time points (the entire epoch, excluding the baseline) for 19 participants, and items were the four experimental conditions, resulting in a 7,600 × 4 matrix. Each matrix yielded one Cronbach's alpha value, and this value was determined for each EEG sensor. Thus, a topographical map of internal consistency estimates resulted, indicating the consistency of the voltage time courses across the four experimental conditions. Calculations were repeated for ERPs based on 10, 20, 40, and 80 averaged trials, in the same manner as the analyses described above. Paralleling other ERP measures, the internal consistency of the ERP time course at individual sensors increased with trial count (see Figure 7). As expected given the visual stimulus used in the present study, waveform consistency was highest at posterior sensors. High reliability of the ERP waveform across conditions was reached after 40 trials, with a majority of sensors displaying

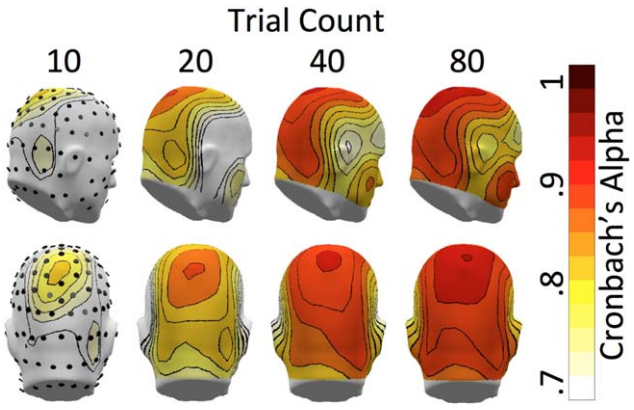

Figure 7. Reliability of the ERP voltage time course. Cronbach's alpha values represent the internal consistency of the temporal sequence of pattern-evoked ERP components across the epoch at each individual sensor, shown for different trial counts. Cronbach's alpha values are projected to the scalp for illustration. Top: right lateral view. Bottom: back view. Red colors indicate greater internal consistency.

Cronbach's alpha values exceeding .8. With all trials, Cronbach's alpha values exceeded .9 (excellent internal consistency) at 110 (of 129) sensors, suggesting that waveforms were consistent across the four experimental conditions.

Reliability after averaging across time points and sensors. In many empirical ERP studies, averaging voltage across time points and/or sensors prior to statistical analysis, assumed to reduce error variance and noise, forms the dependent variable. Thus, we also assessed internal consistency of pattern-onset ERPs using the more common approach in which dependent variables were formed by voltage averaging prior to statistical analysis. In this analysis, Cronbach's alphas were calculated after parametrically increasing the number of sensors and time points included in an average. In terms of sensors, this procedure started with the midline electrode at which the grand mean pattern-evoked potential tended to be largest (Oz for P1 and N1, Pz for N2, P2, P3; see Figure 2 and 3), and then included increasing numbers of additional electrodes, added in sequence of proximity to the first electrode. Averaging across time started with the peak within a time window for each component: P1 (100–140 ms), N1 (160–190 ms), P2 (200–230 ms), N2 (240–270 ms), and P3 (300–380 ms). Once the peak was found, each successive average included adjacent time points in both directions in a time-domain average until reaching the borders of the time windows encompassing the major pattern-onset ERP components. Cronbach's alpha was calculated for each component separately, for all sensor cluster sizes and time window durations. Again, Cronbach's alpha values were calculated for ERPs with different trial counts.

Quantifying Cronbach's alpha for spatiotemporal voltage averages partly supported the notion that averaging across sensors and time points may heighten reliability. Figure 8 shows the increase in internal consistency associated with averaging across time points and sensors, for the N1 time window: Moderate consistency was observed for a measure of the N1 amplitude that was based on averaging across a 30-ms time window around the N1 peak (176 ms), at sensor Oz. High values (>.7) were obtained for ERPs based on 20 trials, when averaging across 2–10 posterior sensors irrespective of temporal averaging. In the same set of averages (20 trials), Cronbach's alpha values were excellent (>.9) after averaging 15 sensors in a cluster around Oz, and 40 time points in the window average around the peak of the N1. Importantly, including additional sensors was associated with a sharp decrease in internal consistency, in line with the consistency analysis of peak voltages at individual sensors presented above. When using all trials, no additional benefit in terms of consistency emerged from across additional time points and sensors, but including sensors beyond the parietooccipital region again led to a decrease in

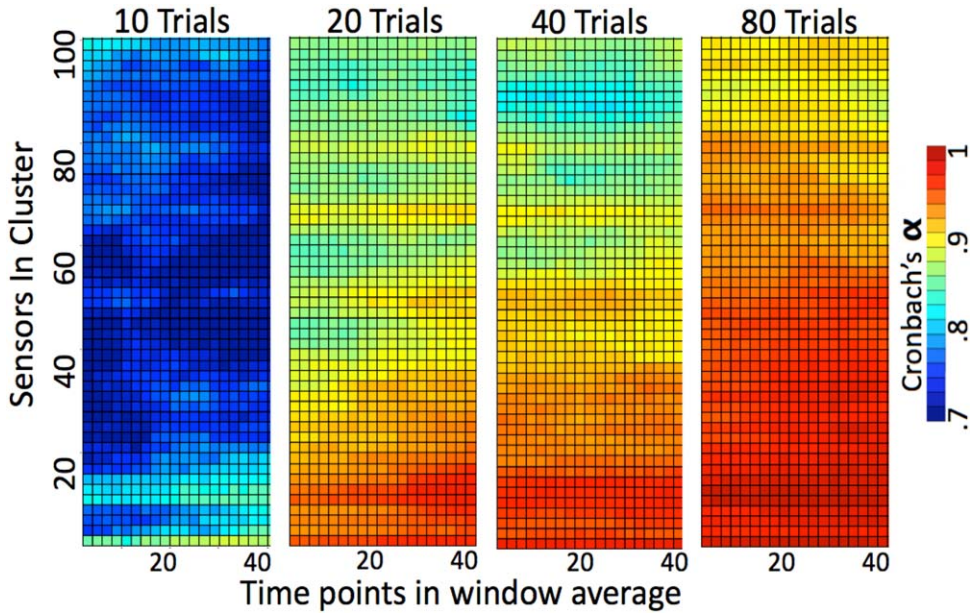

Figure 8. Effects of measuring mean voltage across time points or sensors on internal consistency during the N1 time window (160–190 ms). Raster plots show color-coded Cronbach's alpha as a function of increasing the number of time points (*x* axis) and sensors (*y* axis) included in the mean voltage measure used as a dependent variable. Sensors were added radially, starting with Oz (sensor 1). Time points were added bilaterally, starting at the N1 peak temporal peak at 176 ms. The four panels represent raster plots containing Cronbach's alphas for different trial counts. High internal consistency across the four experimental conditions is displayed in red.

Table 2. *Percentage of Sensors and Time Points with Excellent Internal Consistency*

	P1		P3	
Trial count	Time	Sensors	Time	Sensors
10	45	1	0	0
20	100	40	100	25
40	100	98	100	43
80	100	100	100	98

Note. Excellent internal consistency was defined as Cronbach's alpha > .9. Here, P1 and P3 amplitude measurements are shown after averaging across the temporal and spatial domains. Paralleling the strategy described for Figure 8 and 10, Cronbach's alpha was calculated for increasing numbers of sample points within each component's time range and for growing numbers of sensors included in the voltage measurement, for different trial counts (rows). After averaging and reliability analyses, the percentage of time windows and sensor clusters yielding high-reliability values (Cronbach's alpha > .9) are shown for each component. Data from the N1 and SN components are shown in more detail in Figure 8 and Figure 10, respectively.

consistency. The same conclusions are suggested by analyses of the P1 and P3 components, as shown in Table 2.

Robustness of differences between experimental conditions (effect size of the SN). Beyond internal consistency, replicability of ERP results in many cases depends on the robustness of differences between experimental conditions. In ERP research, these differences are often visualized using difference waveforms, such as the selection negativity in the current study. *F*-contrast analyses were used to quantify the effect size of the predicted voltage difference between attention conditions. Paralleling reliability analyses with Cronbach's alpha, an initial effect size analysis was conducted at each electrode and time point within the selection negativity time range, for different trial counts.

As illustrated by a comparison of Figure 9 and 10, difference waveforms tended to have significantly lower SNR. We first quantified the effect size of the attention-related differences during the selection negativity time window (178 to 292 ms) using planned contrasts, calculated for each sensor location. As shown in Figure 9, medium (.39) to large (.59) effect sizes were reached only when including all available trials, and were confined to a parietooccipital sensor cluster. The greatest effect size of .47 was observed at sensor Pz, where the SN displayed greatest SNR.

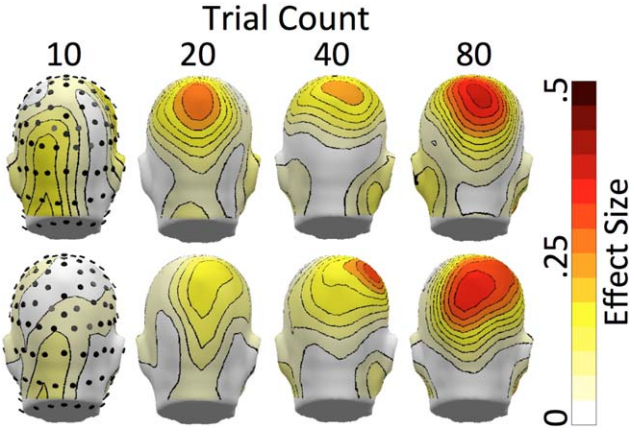

Figure 9. Effect size of condition differences. The topographical distribution of the robustness (effect size) measured as the R^2 of the linear contrast across four attention conditions (S+O+, S+O-, S-O+, S-O-), for the peak during early (178–234 ms; top row) and late (236–292 ms; bottom row) part of the selection negativity component. R^2 values were determined for each sensor and projected to the scalp for illustration. Note that satisfactory (medium) effect size of the predicted effect emerges only when including all trials, corresponding to SNRs around 20 in the present study.

Robustness of condition differences after averaging across time points and sensors. Paralleling the approach of the reliability analyses above, we modeled the predicted differences between conditions after averaging time windows and sensor clusters surrounding Pz. The analysis started by calculating the effect size at the two difference wave peaks (230 and 256 ms, respectively) for sensor Pz. The analysis then grew to include larger time windows surrounding the peak, and larger sensor clusters expanding radially from Pz. Again, this analysis was conducted for all subsets of trials.

As shown in Figure 10, effect sizes of condition differences were affected by averaging across time points and sensors, and varied from early (178–234 ms) to late (236–292 ms) selection negativity (see Figure 10). In both early and late SN time windows, a moderate effect size was observed after averaging all trials, when averaging between 3 and 60 sensors around Pz, and for any group of time points within the respective early or late SN window. Effect sizes differed between the early and late SN time window. The early time window showed highest effect sizes (maximum of .52) after averaging across 20 ms of time around the temporal peak of the SN (i.e., 206 ms) and a cluster of 12 sensors surrounding Pz. The late SN displayed highest effect sizes (up to .41) when averaging across 60 ms of time around the peak and 15 sensors surrounding Pz. Importantly, including time points or sensors that were not consistent with the voltage topography and time course of the selection negativity component tended to dramatically decrease effect size estimates. To compare these results with previous internal consistency analyses, see confidence intervals at various trial counts in Figure 11.

Discussion

The goal of this study was to provide an example analysis for how SNR, internal consistency, and robustness may be established for dependent variables derived from ERPs in an experimental design with within-participant manipulations. To illustrate possible ways toward quantifying (and maximizing) the internal consistency of ERP results, we systematically examined the relation between the trial count (and thus the SNR) and internal consistency, while using effect size as a measure of ERP robustness. In addition, the effects of several commonly used quantification techniques on reliability were investigated, such as measuring the peak voltage or the mean voltage across time points and/or electrodes. Given the spatiotemporal nature of ERP data, different types of dependent variables may be extracted from the Electrode × Time matrices available for each participant and condition. Of these different variables, we examined the internal consistency of (a) peak and mean voltage at selected sensors, (b) the entire voltage topography at selected time points, and (c) the entire waveform at selected electrodes. The findings have implications for a series of questions that are of theoretical and practical relevance for ERP researchers, discussed below.

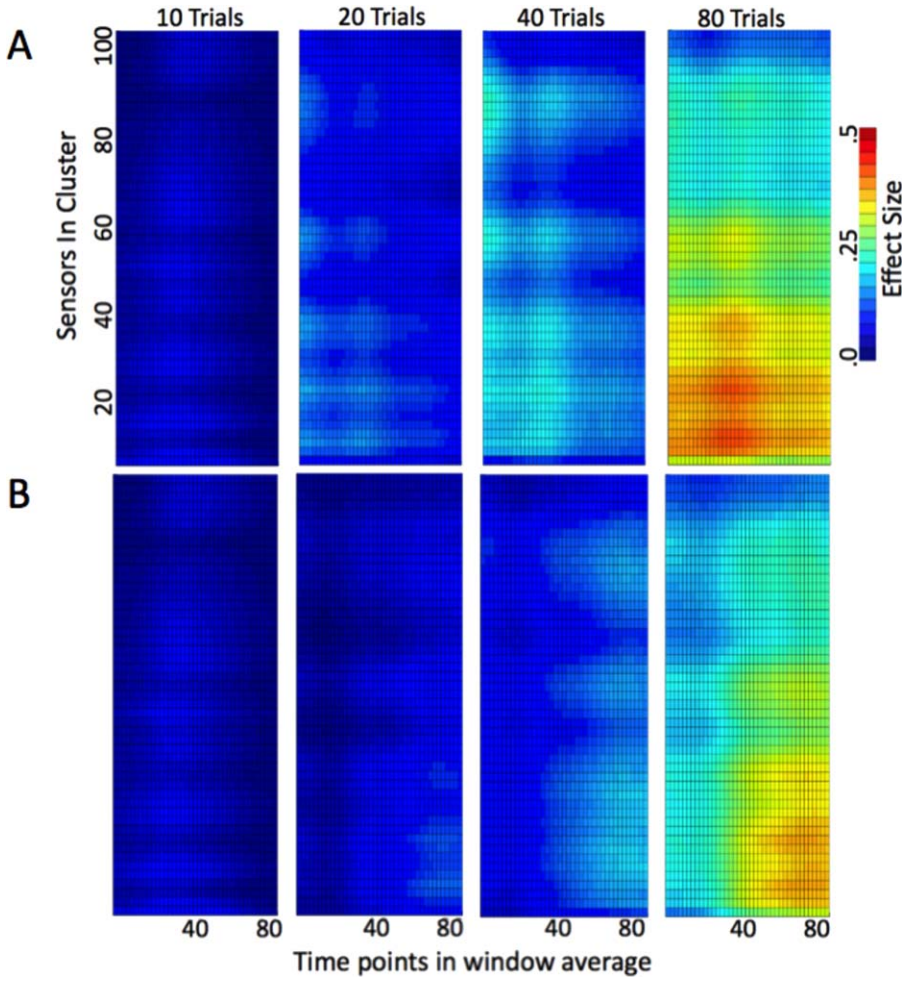

Figure 10. Effects of measuring mean voltage across time points and sensors on the effect size of condition differences. Raster plots show color-coded effect size (R^2) as a function of increasing the number of time points (x axis) and sensors (y axis) included in the mean voltage measure used as a dependent variable. Values were computed during early (178–234 ms; top row) and late (236–292 ms; bottom row) part of the selection negativity component. Sensors were added radially, starting with Oz. Time points were added symmetrically, starting at the N1 temporal peak at 176 ms. The four panels represent raster plots containing Cronbach's alphas for different trial counts. High internal consistency across the four experimental conditions is displayed in red.

Does calculating internal consistency metrics in studies of within-participant effects have practical value?. The present analysis found that the reproducibility of all variables examined across repeated measurements in the same participant was readily captured by calculating internal consistency using the experimental conditions as items. Notably, this approach was sensitive to several properties of ERP data known to affect reproducibility. For example, high SNR strongly predicted high consistency, and consistency also displayed spatial and temporal specificity reflective of the known time course and topography of pattern-evoked ERPs. An important question is how reactivity to the experimental manipulation will affect internal consistency, compared to consistency of components that are not modulated by the experimental manipulations. In the current study, Cronbach's alpha for occipital voltage amplitude was relatively reduced during a narrow time window (the selection negativity time window: 160–280 ms), although still being at satisfactory to very good levels (see Figure 5). Thus, reactivity (the change of the ERP variable in response to the experimental manipulations) in the present study did not drastically alter the ranking of participants across conditions, again supporting the use of conditions as items for estimating consistency. In practice, to document the consistency of the ERP in a given study, researchers may compare consistency of aspects of the time-varying ERP signal that are outside versus inside the temporal region of interest. Relatively lower consistency accompanied by satisfactory effect size during the time window of interest then would point to an effect that is built on consistent, robust individual ERPs, as opposed to noisy and irregular waveforms.

Are there differences in internal consistency between spatial and temporal properties of the ERP?. The ERP is given as a two-dimensional matrix with temporal and spatial properties. Thus, researchers often use portions of the temporal and/or spatial information to quantify the latency (Miller, Patterson, & Ulrich, 1998) and topographical distribution (McCarthy & Wood, 1985) of ERP components. Latency and topographical distribution of an ERP component can be utilized to compare amplitude differences across time, locations on the scalp, and experimental conditions (Cuthbert, Schupp, Bradley, Birbaumer, & Lang, 2000; Dien, Spencer, & Donchin, 2004; Foxe & Simpson, 2002; Kappenman & Luck, 2012).

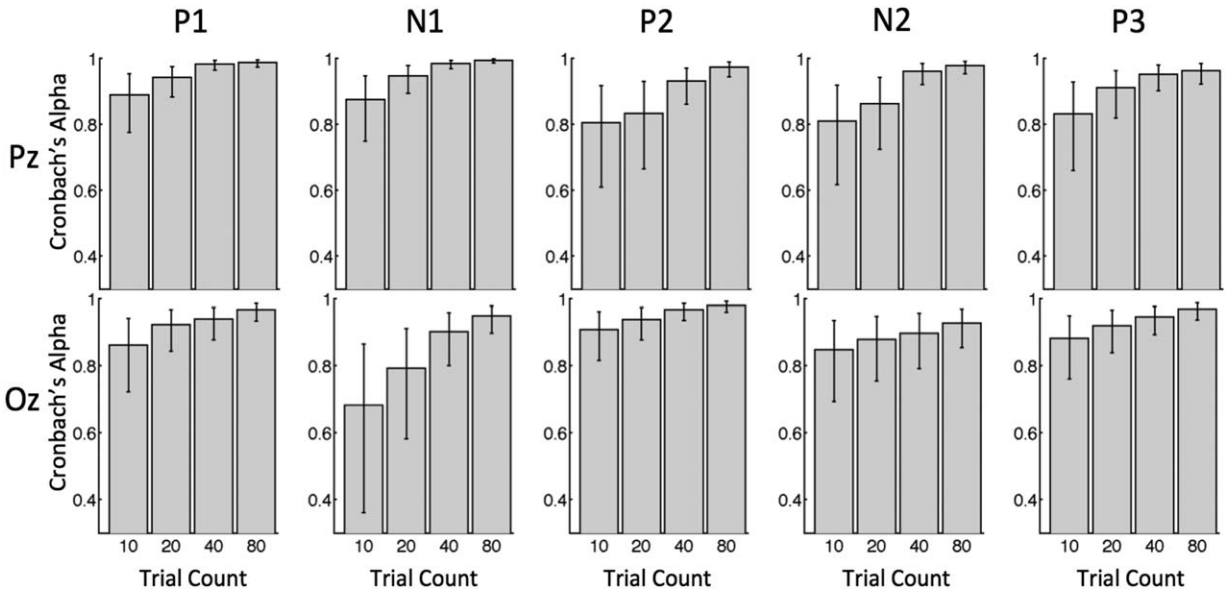

Figure 11. Selected Cronbach's alpha values with confidence intervals: These bar graphs represent the Cronbach's alpha values conducted on the 19 × 4 matrices, each representing the internal consistency of the four experimental conditions across the 19 participants calculated at every time point and sensor. Shown here are 10 most relevant plots from this analysis: the peak time points for the five ERP components at two midline sensors, each with incremental trial counts. (Notably, components P1 and N1 had their peak amplitude at sensor Pz, while components P2, N2, and P3 peaked at sensor Oz).

The current study found both the topographical distribution of voltages (at individual time points) and the ERP waveform (the sequence of voltage changes over time at individual sensors) internally consistent under certain conditions: The reliable quantification (Cronbach's alpha $> .7$) of the voltage topography at a given time point was restricted to the time range in which clear ERP components were seen, and required averages containing 40 to 50 trials, which corresponded to SNRs above 10 in the present analysis. The time-varying internal consistency of the voltage topography was most sensitive to differences between the experimental conditions. Since conditions were used as items here, a significant drop in consistency marked the time range of the selection negativity, in which the differences between the four experimental conditions were most pronounced. Thus, although the present approach of using experimental conditions as items or repetitions for consistency analysis is convenient and widely applicable in most empirical studies, caution is warranted in situations in which between-conditions differences are expected to lead to qualitative changes in the ERP topography. Conversely, the time-varying consistency analysis demonstrated here may provide a sensitive, quantitative data mining tool for detecting time periods of systematic topographical differences between conditions. Future work may build on existing work by systematically examining the effects of electrode density on voltage measurements (Junghöfer et al., 1997) and source estimation (Hauk, Keil, Elbert, & Müller, 2002), and include the aspect of reliability of experimental differences.

Reliably measuring the temporal sequence of ERP components across the entire epoch was possible when using averages comprising 30 or more trials, across major portions of the posterior scalp. At these trial counts, SNRs of the main ERP components varied between approximately 4–10. Although the findings of the present report will necessarily be paradigm specific, this may be taken to indicate that the dynamics of the waveform are replicable at lower SNR than is necessary to reliably capture voltage topographies. To fully harness the internal consistency available at relatively low trial counts, however, it is crucial to measure the voltage waveform at time points and sensors associated with the maximum SNR across components. In the present dataset, high SNR across components was seen at parietooccipital sensor locations only. Accordingly, reliable waveform estimation at anterior electrode sites requires substantially higher trial counts, compared to posterior sensors. This relation highlights the important role of SNR for reliable measurement of ERP voltages, discussed in greater detail in the next paragraph.

How many trials are required for the robust quantification of an ERP effect?. Recommendations for trial counts are often based on experience, tradition, laboratory lore, or estimates of signal-to-noise of a given ERP component (Woodman, 2010). An alternative approach taken here consists of quantitatively assessing different psychometric and quality criteria. Internal consistency of an ERP measure is a minimal condition for its use as an index of a given brain process. Across the present study, high Cronbach's alpha values ($> .7$) were observed for different measures derived from the ERP, at trial counts that may be considered surprisingly low: When considering the cross-condition internal consistency of individual (peak) voltage amplitude measurements, high consistency was observed after averaging 30 or more trials, at posterior scalp regions and across different ERP components. At the same time, SNR for the 30-trial averages was in the range of 4–10 in parietal and occipital clusters, considered not optimal for empirical studies (Luck, 2014). It is important to keep in mind that Cronbach's alpha indexes the extent to which a number of items (here, four experimental conditions) covary across observations, which is often interpreted as evidence of their measuring the same underlying construct (here, the brain process of interest). Thus, internal consistency can be regarded as a minimum necessary, but not sufficient, condition for robust estimation of ERP effects: Authors interested in using an ERP voltage measure (e.g., a component peak such as the P3) as a marker for individual participants may rely on

relatively low trial counts. However, high internal consistency of the nondifference voltages does not imply that any condition differences will be reliably detected.

The effects of SNR and quantification techniques on capturing experimental effects (condition differences) were measured by the effect size of the selection negativity effect across the four conditions. This procedure has the advantage that it takes all experimental conditions into account, thus paralleling the Cronbach's alpha analyses. Deviating strongly from the internal consistency measures, however, the observation of moderate to high effect sizes required using all available trials (i.e., median of 80 artifact-free trials with correct responses). Likewise, SNR of the difference waveform was strongly attenuated compared to nondifference waveforms, with 80-trial averages associated with SNRs ranging around 5–6, at parietooccipital sensors. These findings support trial count recommendations targeting a SNR of 10 (Luck, 2014), which in the present dataset would be expected to lead to greater spatial extent of high effect size measurements, across wide areas of the scalp. In many studies with clinical, pediatric, or aging populations, however, these trial counts may not always be achievable, and explicit measurements of SNR may assist authors wishing to document the data quality available in a given study. Keeping in mind the paradigm specificity of the present results, researchers may expect to obtain reliable findings (at moderate to high effect size) when the SNR of the difference waveform is in the range of 5–6, specifically in studies where the time range and electrode location of the expected effect are known a priori. This a priori knowledge allows further improvement of SNR by using appropriate quantification techniques, discussed next.

How does the measurement technique impact reliability and effect size?. Previous studies examining reliability have focused on a particular ERP component of interest, often measured in many different ways. Measurement techniques widely used in ERP studies include averaging or integrating voltages across time points and sensors, with substantial variability regarding the extent (and type) of averaging in both the time and the spatial domain. Conventions for measuring a given ERP waveform are often grounded in tradition and tend to be flexibly adjusted to changing demands, for example, posed by studying a specific population or using a different paradigm or experimental task.

The current analysis demonstrated, not surprisingly, that averaging across time points prior to analyses improved reliability and effect size. This is of particular relevance for researchers interested in quantifying spatiotemporal dynamics at high temporal resolution (Dien, Spencer, & Donchin, 2004), based on spatial information derived from individual sample points. As illustrated in a recent analysis of the low-amplitude C1 component (Foxe et al., 2008), such approaches should be guided by caution, because important spatiotemporal information may be lost by generous averaging across time points when measuring mean amplitude. The temporal and spatial specificity, and thus the external validity, of ERP measurements may be endangered particularly in situations where mean amplitudes are computed across extended time periods of ERP signals measured at low SNR (Ravden & Polich, 1999). Many strategies have been proposed to address this issue, including combining time points according to their multivariate structure into temporal factors (Dien, 2010) or by capitalizing on the rich information contained in the single trials entering the ERP average (Makeig, Debener, Onton, & Delorme, 2004). In a similar vein, techniques that use the variability in the time course and topography to determine temporally stable "microstates" in the ERP (Pascual-Marqui, Michel, & Lehmann, 1995) may assist in ensuring that the integration of voltages at subsequent time points into one index does not reduce the validity of the measurements.

Averaging across any of the available domains (trials, time, or sensors) may increase both the signal-to-noise ratio and the internal consistency, at different rates for each domain. The present study strongly suggests caution, however, when applying this approach, because SNR, effect size, and internal consistency were all negatively affected by excessive averaging across electrodes and time points. For instance, varying the number of trials averaged together produced internally consistent results after ~40 trials for the entire time course, but only at EEG sensors located over occipital and parietal areas. Including frontal or facial EEG sensors drastically decreased internal consistency and effect size. Thus, the major components of the pattern-evoked visual ERP may be consistently measured based on a 40-trial average at any occipital or parietal sensor, but voltage differences at frontal or facial sensors will not be reliably captured by such an analysis. In a similar vein, measuring individual peaks of the pattern-evoked ERP from 40-trial averages is possible for posterior sensors, but the same 40-trial average would result in unsatisfactory internal consistency when considering anterior sensors. It is highly likely that these specific numerical results will not apply to other ERP studies, given differences in ERP components evoked from different stimuli and in different paradigms, along with varying data quality drawn from different populations and EEG systems. However, analyses of internal consistency are easily implemented and may accompany reports using new analysis techniques, new ERP variables, or ERP measurement techniques, ideally accompanied by reporting the SNR. Communicating quantitative indices of internal consistency such as Cronbach's alpha may assist both the authors and readers in assessing the robustness of effects, thus helping to increase reproducibility in future studies with similar paradigms.

Further highlighting this point, the present study found generally nonlinear relations between SNR, internal consistency, and effect size, for different measurement techniques such as averaging across domains (e.g., trials, time, or sensors). These indices also greatly varied by the scalp location and time segment included in the analyses. As predicted, SNR increased logarithmically as a result of averaging across trials, such that doubling the trial count produced a linear SNR increase, but this relation was specific to scalp locations sensitive to the component under consideration. For example, as shown in Figure 4, SNR for the P1 ERP component measured at sensor Oz increased linearly as the number of trials doubled. Sensors near Oz (the location of the P1 maximum) showed similar increases in SNR, whereas sensors in frontal areas (distal to the location of the P1 maximum) showed low SNR regardless of trial count. By contrast, internal consistency increased with the number of trials averaged across wide areas of the scalp, including at frontal and lateral sensors. Frontal EEG sensors reached excellent reliability with all trials at component peaks, despite small SNR at those locations.

Whereas internal consistency increased logarithmically with added trials, it changed quadratically as a function of averaging across multiple time points within a component. For example, for the N1 component (based on 10 trials) measured at sensor Oz, averaging 10, 20, 30, and 40 time points surrounding the peak yielded Cronbach's alphas of .7, .85, .9, and .8, respectively. Thus, the N1 component was most consistent when using a 20–30 ms window average centered on the N1 peak, but reliability decreased if this window was expanded further. This is consistent with the intuitive notion that measuring the mean amplitude improves internal

consistency only as long as the averaging window includes time points that are part of the component of interest. Including time points with different properties, with polarity being an obvious example, will necessarily reduce SNR and reliability as well as external validity of the measurements. Time domain averaging for other components, including the selection negativity, produced reliability fluctuations in a similar quadratic pattern, with a time window of approximately 30 ms found to maximize internal consistency and effect size. Thus, averaging across trials and averaging across time points within a given component window each increased reliability, but trial domain averaging reached ceiling (Cronbach's alpha values nearing 1) with all trials, while time domain averaging began to decrease internal consistency when extending the window beyond 30 ms.

In ERP studies, spatial averaging is sometimes implemented in a data-driven way, by selecting the EEG sensor with the largest SNR ratio to serve as the center of an electrode cluster containing sensors for spatial averaging. Analogous to the temporal averaging described above, the present study examined effects of this simple technique by averaging across electrode clusters containing increasing numbers of sensors, while comparing quality indices of the data. Paralleling averaging across time points, spatial averaging showed strong nonlinear effects on quality indices; for example, the analysis for the P1 component started with sensor Oz, where the SNR distribution of the P1 component showed a maximum. Additional sensors were then added to the cluster based on spatial proximity. As shown in Figure 8, internal consistency was highest when the cluster was smallest (only Oz), and tended to decrease as sensors were added to the cluster. Thus, cluster sizes of 5, 20, 35, and 50 were associated with internal consistency values of approximately .95, .9, .85, and .8, respectively, for the peak P1 voltage extracted from a 20-trial average. This somewhat unexpected negative relationship was apparent for all components examined (the P1, N1, P3, SN), and for all averaged trial counts (except 80 averaged trials, where internal consistency remained near one for nearly all cluster sizes). Close examination of Figure 8 shows that sensor averaging may result in very modest consistency increases compared to individual sensor measurements. The feature-based attention difference waveform (containing the SN) is, by virtue of being a difference waveform, particularly dependent on the signal-to-noise ratio of the nondifference ERPs on which it is based. In the present study, SN showed modest effect sizes with trial counts of 80 when considering individual time points and sensors. Alphas were substantially greater when averaging across sensors and time points: Pooling voltages for posterior midline sensors in the time range of the N1 and N2 components, where the SN was maximal, resulted in the highest effect sizes, but still only reached values around 0.6. Together, these findings suggest internal consistency is promoted by measuring pattern-evoked ERPs by including time points, but to a lesser extent by including sensors into component scores used as a dependent variable. Given the wide range of practices used in the ERP literature, the increased availability of similar quantitative analyses of quality indices would be desirable, allowing comparison of different quantification approaches.

Conclusions and Outlook

The present study explored ways in which the internal consistency of ERP measurements can be assessed. A representative dataset from a selective attention task was used, involving pattern-evoked visual ERPs recorded by means of dense-array EEG. Main results converged to show high internal consistency of measurements taken from nondifference ERPs, even at surprisingly low trial counts, corresponding to relatively low SNRs. By contrast, robust quantification of voltage differences between experimental conditions, measured by the effect size, required significantly greater SNRs. Overall, consistency as well as effect size varied by SNR, but not in a linear fashion: SNR predicted consistency and effect size at posterior scalp locations where the pattern-evoked ERP signal was pronounced, but not at other sites. A comparison of quantification techniques assessed differences between measuring the peak amplitude and measuring the mean amplitude with varying time points and electrode sites included in the mean. Throughout these analyses, internal consistency and effect size benefited from measuring mean voltage, compared to the peak voltage in situations where (a) SNR of the signal of interest was low, and (b) when including only neurophysiologically plausible time points and sensors into a mean amplitude measurement (i.e., time points and electrode locations that captured the same process). Including additional scalp locations and time points was associated with a sharp decrease in internal consistency and effect size. Thus, the common method of measuring mean amplitude as spatiotemporal averages across a subset of the ERP matrix may be informed by quantitative analyses of consistency, to ensure that a given practice reliably captures the desired aspect of the ERP signal.

It will be an interesting goal for future studies to explore the extent to which the present analyses may be extended to other experimental paradigms. Because the necessary computation efficiency and technical training are now widely available in ERP laboratories, quantitative analyses of internal consistency could easily accompany reports on experimental findings. Future studies may also wish to characterize the reliability using additional paradigms and measurements common in ERP studies, such as a component's temporal peak or metrics extracted from independent or principal component analysis. Overall, given the growing number of methodological developments, novel paradigms, and increased use of sophisticated measurement techniques, extensive practice of reporting internal consistency may be a welcome addition to the psychophysiologist's toolbox.

References

Anllo-Vento, L., & Hillyard, S. A. (1996). Selective attention to the color and direction of moving stimuli: Electrophysiological correlates of hierarchical feature selection. *Perception & Psychophysics*, *58*, 191–206. doi: 10.3758/BF03211875

Anokhin, A. P., van Baal, G. C. M., van Beijsterveldt, C. E. M., de Geus, E. J. C., Grant, J., & Boomsma, D. I. (2001). Genetic correlation between the P300 event-related brain potential and the EEG power spectrum. *Behavior Genetics*, *31*, 545–554. doi: 10.1023/A:1013341310865

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*, 98–101.

Cohen, J., & Polich, J. (1997). On the number of trials needed for P300. *International Journal of Psychophysiology*, *25*, 249–255. doi: 10.1016/S0167-8760(96)00743-X

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi: 10.1007/BF02310555

Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birbaumer, N., & Lang, P. J. (2000). Brain potentials in affective picture processing: Covariation with autonomic arousal and affective report. *Biological Psychology*, *52*, 95–111.

Dien, J. (2010). The ERP PCA toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of*

Neuroscience Methods, *187*, 138–145. doi: 10.1016/j.jneumeth.2009.12.009

Dien, J., Spencer, K. M., & Donchin, E. (2003). Localization of the event-related potential novelty response as defined by principal components analysis. *Brain Research: Cognitive Brain Research*, *17*, 637–650.

Dien, J., Spencer, K. M., & Donchin, E. (2004). Parsing the late positive complex: Mental chronometry and the ERP components that inhabit the neighborhood of the P300. *Psychophysiology*, *41*, 665–678. doi: 10.1111/j.1469-8986.2004.00193.x

Donchin, E., Callaway, E., Cooper, R., Desmedt, J. E., Goff, W. R., Hillyard, S., & Sutton, S. (1977). Publication criteria for studies of evoked potentials in man. In J. E. Desmedt (Ed.), *Attention, voluntary contraction and event-related cerebral potentials* (Vol. *1*, pp. 1–11). Brussels, Belgium: Karger.

Duncan-Johnson, C. C., & Donchin, E. (1979). The time constant in P300 recording. *Psychophysiology*, *16*, 53–55. doi: 10.1111/j.1469-8986.1979.tb01440.x

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*, 78–86. doi: 10.3758/BF03209418

Fabiani, M., Gratton, G., Karis, D., & Donchin, E. (1987). Definition, identification, and reliability of the P300 component of the event-related brain potential. *Advances in Psychophysiology*, *2*, 1–78. Retrieved from http://www.researchgate.net/publication/225304622_Definition_identification_and_reliability_of_the_P300_component_of_the_event-related_brain_potential

Foti, D., Kotov, R., & Hajcak, G. (2013). Psychometric considerations in using error-related brain activity as a biomarker in psychotic disorders. *Journal of Abnormal Psychology*, *122*, 520–531. Retrieved from http://psycnet.apa.org/journals/abn/122/2/520

Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans: A framework for defining "early" visual processing. *Experimental Brain Research*, *142*, 139–150.

Foxe, J. J., Strugstad, E. C., Sehatpour, P., Molholm, S., Pasieka, W., Schroeder, C. E., & McCourt, M. E. (2008). Parvocellular and magnocellular contributions to the initial generators of the visual evoked potential: high-density electrical mapping of the "C1" component. *Brain Topography*, *21*, 11–21. doi: 10.1007/s10548-008-0063-4

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

Handy, T. C. (2005). *Event-related potentials: A methods handbook*. Cambridge, MA: MIT Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=OQyZEfgEzRUC&pgis=1

Harter, R. M., & Aine, C. J. (1984). *Brain mechanisms of visual selective attention*. Retrieved from http://www.researchgate.net/profile/Cheryl_Aine/publication/243784610_Brain_Mechanisms_of_Visual_Selective_Attention/links/53f4d0690cf2888a74912369.pdf

Hauk, O., Keil, A., Elbert, T., & Müller, M. M. (2002). Comparison of data transformation procedures to enhance topographical accuracy in time-series analysis of the human EEG. *Journal of Neuroscience Methods*, *113*, 111–122. doi: 10.1016/S0165-0270(01)00484-8

Hinton, P., McMurray, I., & Brownlow, C. (2004). *SPSS explained* (2nd ed). New York, NY: Routledge.

Hopf, J. M., Boelmans, K., Schoenfeld, M. A., Luck, S. J., & Heinze, H. J. (2004). Attention to features precedes attention to locations in visual search: Evidence from electromagnetic brain responses in humans. *Journal of Neuroscience*, *24*, 1822–1832.

Junghöfer, M., Elbert, T., Leiderer, P., Berg, P., & Rockstroh, B. (1997). Mapping EEG-potentials on the surface of the brain: A strategy for uncovering cortical sources. *Brain Topography*, *9*, 203–217. doi: 10.1007/BF01190389

Junghöfer, M., Elbert, T., Tucker, D. M., & Rockstroh, B. (2000). Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology*, *37*, 523–532. doi: 10.1111/1469-8986.3740523

Kappenman, E. S., & Luck, S. J. (2012). ERP components: The ups and downs of brainwave recordings. In S. J. Luck & E. S. Kappenman (Eds.), *Oxford handbook of ERP components*. New York, NY: Oxford University Press.

Keil, A., Debener, S., Gratton, G., Junghofer, M., Kappenman, E. S., Luck, S. J., … Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*, 1–21. doi: 10.1111/psyp.12147

Keil, A., & Muller, M. M. (2010). Feature selection in the human brain: Electrophysiological correlates of sensory enhancement and feature integration. *Brain Research*, *1313*, 172–184. doi: 10.1016/j.brainres.2009.12.006

Light, G. A., & Swerdlow, N. R. (2015). Future clinical uses of neurophysiological biomarkers to predict and monitor treatment response for schizophrenia. *Annals of the New York Academy of Sciences*, *1344*, 105–119. doi: 10.1111/nyas.12730

Luck, S. J. (2014). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*, 146–157.

Luck, S. J., Mathalon, D. H., O'Donnell, B. F., Hämäläinen, M. S., Spencer, K. M., Javitt, D. C., & Uhlhaas, P. J. (2011). A roadmap for the development and validation of event-related potential biomarkers in schizophrenia research. *Biological Psychiatry*, *70*, 28–34. doi: 10.1016/j.biopsych.2010.09.021

Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in Cognitive Sciences*, *8*, 204–210. doi: 10.1016/j.tics.2004.03.008

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, *48*, 852–860. doi: 10.1111/j.1469-8986.2010.01152.x

Martinez, A., Anllo-Vento, L., Sereno, M. I., Frank, L. R., Buxton, R. B., Dubowitz, D. J., … Hillyard, S. A. (1999). Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nature Neuroscience*, *2*, 364–369.

McCarthy, G., & Wood, C. C. (1985). Scalp distributions of event-related potentials: An ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology*, *62*, 203–208.

McGinnis, E. M., & Keil, A. (2011). Selective processing of multiple features in the human brain: Effects of feature type and salience. *PLoS ONE*, *6*, 1–12. doi: 10.1371/journal.pone.0016824

Miller, J., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology*, *35*, 99–115.

Müller, M. M., & Keil, A. (2004). Neuronal synchronization and selective color processing in the human brain. *Journal of Cognitive Neuroscience*, *16*, 503–522. doi: 10.1162/089892904322926827

Pascual-Marqui, R. D., Michel, C. M., & Lehmann, D. (1995). Segmentation of brain electrical activity into microstates: Model estimation and validation. *IEEE Transactions in Biomedical Engineering*, *42*, 658–665. doi: 10.1109/10.391164

Perez, V. B., Swerdlow, N. R., Braff, D. L., Näätänen, R., & Light, G. A. (2014). Using biomarkers to inform diagnosis, guide treatments and track response to interventions in psychotic illnesses. *Biomarkers in Medicine*, *8*, 9–14. doi: 10.2217/bmm.13.133

Peyk, P., DeCesarei, A., & Junghöfer, M. (2011). Electro magneto encephalograhy software: Overview and integration with other EEG/MEG toolboxes. *Computational Intelligence and Neuroscience*, *2011*, Article ID 861705.

Pontifex, M. B., Scudder, M. R., Brown, M. L., O'Leary, K. C., Wu, C. T., Themanson, J. R., & Hillman, C. H. (2010). On the number of trials necessary for stabilization of error-related brain activity across the life span. *Psychophysiology*, *47*, 767–773. doi: 10.1111/j.1469-8986.2010.00974.x

Ravden, D., & Polich, J. (1999). On P300 measurement stability: Habituation, intra-trial block variation, and ultradian rhythms. *Biological Psychology*, *51*, 59–76.

Rosnow, R. L., Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods*, *1*, 331–340.

Schlögl, A., Keinrath, C., Zimmermann, D., Scherer, R., Leeb, R., & Pfurtscheller, G. (2007). A fully automated correction method of EOG artifacts in EEG recordings. *Clinical Neurophysiology*, *118*, 98–104. doi: 10.1016/j.clinph.2006.09.003

Schoenfeld, M. A., Hopf, J. M., Martinez, A., Mai, H. M., Sattler, C., Gasde, A., … Hillyard, S. A. (2007). Spatio-temporal analysis of feature-based attention. *Cerebral Cortex*, *17*, 2468–2477.

Simons, R. F., & Miles, M. A. (1990). *Nonfamilial strategies for the identification of subjects at risk for severe psychopathology: Issues of reliability in the assessment of event-related potential and other marker variables*. New York, NY: Oxford University Press.

Spencer, K. M., Dien, J., & Donchin, E. (1999). A componential analysis of the ERP elicited by novel events using a dense electrode array. *Psychophysiology*, *36*, 409–414.

Teplan, M. (2002). *Fundamentals of EEG measurement*. Retrieved from http://www.edumed.org.br/cursos/neurociencia/MethodsEEGMeasurement.pdf

Vidaurre, C., Sander, T. H., & Schlögl, A. (2011). BioSig: The free and open source software library for biomedical signal processing. *Computational Intelligence and Neuroscience*, *2011*, 935364. doi: 10.1155/2011/935364

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, *72*, 2031–2046. doi: 10.3758/BF03196680

(Received September 2, 2015; Accepted January 26, 2016)