How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study

Scott A. Baldwin  
*Brigham Young University - Provo*

Peter E. Clayson

Michael J. Larson

Follow this and additional works at: https://scholarsarchive.byu.edu/facpub

Part of the Psychology Commons

**BYU ScholarsArchive Citation**

Baldwin, Scott A.; Clayson, Peter E.; and Larson, Michael J., "How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study" (2012). *Faculty Publications*. 6062.  
https://scholarsarchive.byu.edu/facpub/6062

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study

Abstract
There is considerable variability in the quantification of event-related potential (ERP) amplitudes and latencies. We examined susceptibility of ERP quantification measures to incremental increases in background noise through published ERP data and simulations. Measures included mean amplitude, adaptive mean, peak amplitude, peak latency, and centroid latency. Results indicated mean amplitude was the most robust against increases in background noise. The adaptive mean measure was more biased, but represented an efficient estimator of the true ERP signal particularly for individual-subject latency variability. Strong evidence is provided against using peak amplitude. For latency measures, the peak latency measure was less biased and less efficient than the centroid latency measurement. Results emphasize the prudence in reporting the number of trials retained for averaging as well as noise estimates for groups and conditions when comparing ERPs.

Descriptors: Event-related potentials (ERPs), Amplitude, Latency, Statistical extraction, Measurement, Simulation

Background EEG Noise
Nonsystematic noise arises in ERP waveforms from artifact sources, such as muscle activity, movement, electrocardiographic activity, and other artifactual signals, such as skin potentials, equipment-related artifact, and electrical noise in the environment. Additionally, high electrode impedance recordings are more susceptible to influence from certain sources of noise, such as skin potentials (Kappenman & Luck, 2010). To minimize the effects of noise, ERPs are averaged at the trial level, such that all trials of a given type are averaged for a particular subject, then the ERP statistical extraction methods to measure amplitude and latency are subsequently performed on the single-subject averages. In theory, nonsystematic noise is random and should be minimized when numerous signals are averaged together; however, following the averaging of trials for a subject, residual noise remains in the averaged ERP waveform. Researchers hope to maximize the amount of signal relative to noise to obtain the best estimate of the true ERP signal. The signal-to-noise ratio (SNR) is generally understood using the following formula: \( \frac{1}{\sqrt{N}} \times R \), where \( N \) represents the number of trials and \( R \) represents noise (Luck, 2005). Thus, the SNR present in the single-subject average increases as a function of the inverse square root of the number of trials included in the average.

The accuracy of ERP measurements is closely associated with the level of noise in ERP averages, such that as noise increases in the average waveform the accuracy of an ERP measurement should decrease. In other words, ERP measurements are reliable only...
insofar as noise has been removed from the average waveform (Glaser & Ruchkin, 1976; Perry, 1966). Although publications frequently include the baseline time period of the averaged ERP waveforms in order for the reader to visually evaluate the presence of noise in the ERP, this approach is insufficient as noise may not necessarily be constant across the entire waveform (Gratton, Kramer, Coles, & Donchin, 1989; Handy, 2005; Spencer, 2005). In a method proposed by Schimmel (1967), every other trial for a given subject is inverted then averaged, which averages out the true signal while leaving in the remaining background EEG noise. After implementing this averaging procedure, the root mean square can be calculated on the remaining noise to quantify the noise present in the single-subject average; thus, in the absence of noise the expected root mean square should be zero. This method provides a measure for quantifying the amount of noise present in an ERP average.

SNR is of particular concern in studies of ERPs where few trials may be retained for single-subject averaging. For example, performance-monitoring research frequently examines ERP activity associated with erroneous responses. The error-related negativity (ERN) reflects evaluative aspects related to performance monitoring and is a negative deflection in a response-locked ERP waveform occurring within 100 ms after the commission of an error (Falkenstein, Hohnsbein, Hoormann, & Banke, 1991; Gehring, Goss, Coles, Meyer, & Donchin, 1993). The correct-related negativity (CRN) is the correct-trial counterpart of the ERN with identical temporal characteristics and scalp topography as the ERN but occurs following correct responses (Falkenstein, Hoormann, Christ, & Hohnsbein, 2000). Previous findings indicate that amplitude of the ERN can differ based on a variety of individual differences including differences in affect, psychopathology, motivation, and demographic variables (see Larson, Fair, Good, & Baldwin, 2010; Larson, South, & Clayson, 2011; Olvet & Hajcak, 2008; van Noordt & Segalowitz, 2012). Research on the internal consistency of the ERN as the number of trials averaged increases suggests that as few as six to eight trials or as many as 14 or more trials are necessary for an adequate SNR (Larson, Baldwin, Good, & Fair, 2010; Olvet & Hajcak, 2009). The ERN provides one example when the SNR is critically important as few trials are retained for single-subject averaging. In this case, background EEG noise has a significant impact on the amplitude and latency measures utilized on the ERN.

**Extraction Methods**

Two common types of ERP measurement approaches are the peak and area measures; however, in order to understand the advantages of the area measurement approach over the peak measurement approach, it is critical to remember that single-trial epochs are averaged. That is, the ERP peak in a single-subject average represents the mode rather than the average of the single-trial ERP waveforms (see Luck, 2005). In order to ameliorate the problems inherent in measuring components with considerable trial-by-trial latency variability, area-based ERP measurements are commonly employed. Luck (2005) notes that the area under the curve in a single-subject average is equivalent to the average of the area under the curve in each single trial. However, this is not necessarily the case for peak measures that are deleteriously biased by noise (see Discussion below). To evaluate the extent to which common ERP amplitude and latency measurements are affected by noise, we chose to evaluate both area- and peak-based measures.

Three commonly used methods for ERP amplitude extraction are the mean, peak, and adaptive mean measures (see Figure 1).

![Figure 1](image.png)

**Figure 1.** A: Example of a mean amplitude window and centroid latency area measurement. B: Example of the points that would be chosen for peak latency and peak amplitude measurements. C: Example of an adaptive mean measurement.
When computing a mean amplitude, the average ERP activity between two fixed time points is extracted to quantify the amplitude of the peak. The time window for the mean amplitude is commonly chosen based on the grand average waveform for the group (i.e., the average of the single-subject averages) and previous research. The peak amplitude method quantifies the amplitude as the maximum or minimum amplitude (i.e., most negative or positive point) at the latency of the minimum or maximum of the amplitude of the ERP. The adaptive mean method is similar to the area-based mean amplitude method with the addition that the mean amplitude is centered on the peak latency of the individual-subject ERP. The adaptive mean measure first locates the peak latency within a specified time window for a single subject, and then the average activity for a predefined mean-amplitude time window is extracted around the identified peak latency (Electrical Geodesics Inc., 2006). For example, the average of activity 15 ms prepeak to 15 ms postpeak negative amplitude may be extracted around the most positive peak of an ERP component with a specified time window. As a result, the adaptive mean is thus more sensitive to individual-subject variability in peak latency and represents a fusion of area-based and peak-based amplitude measurement.

For latency measures, two common measures are the peak latency and centroid latency (see Figure 1). The peak latency measure quantifies the latency as the time of the peak most negative-going or positive-going amplitude (i.e., the time at which the peak reaches minimum or maximum amplitude). The centroid latency measure is an area-based approach that is analogous to the mean latency and a form of a center-of-mass measurement (see Dien, Spencer, & Donchin, 2004). The centroid latency measure is a type of fractional area latency that finds the time at which the area under the curve is divided into equal halves. Whether an area-under-the-curve or area-over-the-curve calculation is used depends on the polarity of the peak; the centroid latency measurement is only dependent on the baseline insofar as the polarity of the peak is concerned. Considering that an ERP peak represents the mode rather than the average of the single-trial ERP peaks, the centroid latency measurement extracts the average area under the waveform for a positive peak (positive centroid) or over the waveform for a negative peak (negative centroid; see Dien et al., 2004). Thus, the centroid latency measure additionally captures differences in the overall shape of the ERP component.

To estimate how much these amplitude and latency measures are used in recent ERP research, we reviewed the last 3 years of published articles employing ERP amplitude or latency measurement methods from three leading ERP journals: *Psychophysiology, International Journal of Psychophysiology, and Neuropsychologia*. We searched PubMed for articles using the key terms “ERP” or “event-related potential” from January 1, 2010, to May 8, 2012, for the mentioned journals. Articles that used factor analytic approaches to extract amplitude or latency were excluded from the measurement frequency counts but were counted toward the total number of articles evaluated. Some manuscripts included multiple approaches for amplitude quantification, such as a peak amplitude for one component with a prominent peak and a mean amplitude for a more tonic ERP component; in this instance, both methods were counted toward the respective frequencies of each approach but only counted once toward the total number of articles evaluated (i.e., total percentages may exceed 100% for this reason).

For measurement of ERP amplitude, 75.6% (337 of 446) of the articles reviewed employed a mean amplitude measurement, 42.4% (189 of 446) used a peak amplitude approach, and 4.9% (22 of 446) used an adaptive mean measurement method to extract amplitude data. For latency measurements, 80.9% (157 of 194) used a peak latency approach and 0.5% (1 of 194) used a centroid latency measurement. Considering the wide use of ERPs to examine the neural time course of various cognitive, affective, sensory, and motor processes and the variability in the ERP amplitude and latency measurement methods used, a rigorous examination of the various measurement methods is warranted to determine the least biased and most efficient statistical extraction methods.

### Noise and ERP Measurement

High levels of noise relative to low levels of noise will have deleterious effects on all statistical extraction measures; however, the respective robustness of each statistical extraction method against increases in noise levels has received little critical examination. Previous work indicates that the amount of noise present in ERP waveforms affects the extent to which ERP subject averages represent reliable waveforms given a fixed number of trials (Turetsky, Raz, & Fein, 1988). That is, given an identical number of trials retained for averaging, higher levels of noise will reduce the reliability of the average. Although previous work has examined how noise affects ERP waveforms, the purpose of the current investigation was to assess how noise affects ERP measurements themselves. Furthermore, emphasis has been placed on assessing the accuracy of ERP measurement approaches when scoring manually or with a computer (e.g., Callaway, Halliday, & Herning, 1983) and between multiple methods of peak-amplitude measurement (e.g., Fein & Turetsky, 1989; Gratton et al., 1989); however, the present study focused on comparing area-based and peak-based measures of both amplitude and latency to examine the effects of noise on these measurements.

With regard to the amplitude measurement methods, the peak amplitude is thought to be the most compromised measure as noise in the waveforms increases (Luck, 2005, 2012). The peak amplitude approach chooses the most extreme value of a peak within a defined time window. Noise is superimposed on the true signal. The peak amplitude by definition chooses the noise as the noise exaggerates the amplitude of the peak (see Luck, 2005, 2012). Indeed, previous work indicates that the peak amplitude extracts an ERP amplitude that exaggerates the true peak measurement without noise (McGillem, Aunon, & Yu, 1985). Thus, in the presence of increasing levels of noise, the peak amplitude measure should be more biased than either a mean amplitude or adaptive mean measurement method.

Because background EEG noise is considered to be random, the noise should “average out” when using the mean amplitude measure.

---

1. A peak-to-peak amplitude measure was used in 4.7% (21 of 446) of articles. Considering that the peak-to-peak amplitude approach is an extension of the peak amplitude measure (see Discussion section), these articles were evaluated as using peak amplitude measures when reported in the body of the manuscript.

2. A limitation of the current investigation is the use of the absolute peak amplitude measurement rather than a local peak amplitude measurement (see Luck, 2005). The local peak amplitude approach chooses the maximum or minimum amplitude value within a specified time window that is surrounded on either side by three or so samples with smaller amplitude values. The use of the local peak amplitude should reduce the likelihood of identifying spurious peaks in the averaged ERP waveform. Shortcomings of the local peak amplitude aside (see Luck, 2005), this measure may prove to be less biased and more efficient than the absolute peak amplitude. However, this measure was not evaluated in the current study.
measure (see Luck, 2005, 2012). Any nonsystematic noise added to the true ERP waveform is equally likely to increase or decrease the amplitude of an ERP peak. Although in any one ERP peak the noise may exaggerate or attenuate the peak, when averaging numerous single-trial peaks the noise should average out. Thus, the mean amplitude measure should average out the noise and act as an unbiased and efficient estimator of the ERP.

Similarly, the adaptive mean should be robust against increases in noise. Considering that the adaptive mean centers around the peak of the ERP, in the presence of high noise levels the adaptive mean-amplitude time window may shift to a larger degree relative to low noise levels to center on the peak and subsequently be distorted by noise. However, averaging pre- and postpeak amplitude should reduce the likelihood that noise will bias the amplitude measurement as noise should average out when extracting mean ERP activity. We anticipate that the adaptive mean measure would be less biased than the peak amplitude measure and similarly efficient as the mean amplitude measure, but this possibility has not been formally tested.

For the latency measures, the peak latency measure will likely not be robust against increases in noise because noise will distort the latency of the true ERP peak; the noise will be superimposed on the true peak thus altering the minimum or maximum amplitude of the peak and subsequently biasing the peak latency measure. The centroid latency measure will likely be robust against increases in noise because the total area under/over the waveform will vary only slightly with increases in random noise. One factor that may significantly affect the latency measures above and beyond increases in noise is the sampling rate of the data. Considering that sampling at 2000 Hz provides higher temporal resolution than sampling at lower rates such as 250 Hz, latency measures should be less biased as sampling rate increases because the measurement of latency itself becomes more precise. As a result, the current simulation examined the effects of sampling rates on ERP measures.

Bias and Efficiency

To evaluate and compare statistical extraction measures for various levels of noise, we computed bias and efficiency for each amplitude and latency measurement method on a simulated ERP peak. Bias was quantified as the difference between the statistical extraction measure on a positive-going peak (arbitrarily referred to simply as a P1) without simulated noise and the P1 peak with simulated noise at various levels of background EEG noise. Bias represents a measure of the systematic deviation between the measures with (estimate of population parameter) and without noise (true population parameter). Efficiency represents a measure of stability or precision against noise fluctuations in the current examination. To examine efficiency, the square root of the average squared deviation (root mean square error; RMSE) between a statistical extraction measure on the P1 peak without noise and the P1 peak with noise was calculated. Thus, examining the bias and efficiency of statistical extraction methods provides measurements of systematic imprecision and variability, respectively, in the presence of background EEG noise for each method.

To illustrate the concepts of bias and efficiency, an excellent metaphor used in previous research is that of examining shots from three different rifles shooting a target (Greenland, 2000). For the target, the bull’s-eye represents the true population parameter or, in the case of the present simulation, the true ERP amplitude or latency measurement without noise. Shots within the inner ring of the target would represent accurate measurements of the population parameter. The separate rifles represent different methods of estimating the population parameter, such as mean, adaptive mean, and peak amplitude measurement methods. Shots from the first rifle scatter around the bull’s-eye, but only 20% of the shots are within the inner ring of the target. The first rifle would represent an estimator that is unbiased but inefficient, as the shots were clustered around the true population parameter but highly variable. Shots from a second rifle are to the left of the bull’s-eye, but 75% of the shots are in the inner ring. This second rifle represents an estimator that is biased, but highly efficient. The second rifle is biased because shots are not centered on the bull’s-eye but is efficient because the shots are clustered together (i.e., consistent low variability). Lastly, shots from a third rifle are outside of the inner ring and are highly variable. The third rifle represents an estimator that is both biased and inefficient. Taken together, bias is a tendency for shots to miss the bull’s-eye (true population parameter) whereas inefficiency is the tendency for shots to be consistently variable.

The purpose of the current investigation was to assess the statistical bias and efficiency of popular measures for ERP statistical extraction as a function of the noise present in ERP waveforms. The mean amplitude, adaptive mean, peak amplitude, peak latency, and centroid measurement methods were computed on the simulated peak. Then, realistic background EEG noise was simulated and superimposed on the true signal—the P1. The statistical measures were subsequently employed on the P1 waveform with noise. Simulations consisted of 5,000 sets of 1,000 “subject” simulations of 30 trials for 250 Hz, 500 Hz, 1000 Hz, and 2000 Hz sampling rates. In this manner, the measurement of the peak without noise was compared to the measurement of the peak with noise. Noise amplitude was incrementally increased by a step of 0.001 μV in each simulation to model higher levels of noise. Furthermore, ERN data from a published study were presented to characterize the levels of noise found in ERP experiments and examine the relationship between the number of trials averaged and the level of noise present in the single-subject average (Larson, Clayson, & Farrer, 2012).

Method

Peak Simulation

In order to empirically evaluate the susceptibility of each ERP statistical extraction method to EEG noise, a positive-going ERP peak was simulated to provide a “true” peak without noise for a given statistical extraction measure (the P1 described above). This ERP peak consisted of a half-cycle 2.5-Hz sinusoid with peak amplitude of 5 μV (see Figure 2). In order to reflect realistic single-subject ERP data, the latency was varied between 85 and 115 ms. A 200-ms baseline and 200-ms postpeak line were added to the ERP peak to allow for the latency jitter to vary while maintaining a smooth sinusoidal peak (total waveform length: ~200 ms to 400 ms). EEG noise was simulated similarly to previously published ERP simulation studies in which the ERP peak was added to phase-randomized (co)sinusoids (Gratton et al., 1989; Yeung, Bogacz, Holroyd, & Cohen, 2004; Yeung, Bogacz, Holroyd, Nieuwenhuis, & Cohen, 2007). Five phase-randomized sinusoids with randomized amplitude and frequency similar to empirical EEG data were generated to simulate background EEG noise. The generated noise was subsequently added to the simulated ERP peak (see Figure 2). Amplitude of the noise randomly varied within...
maximum amplitude parameters that increased by a step of 0.001 mV to evaluate the effects of increased noise on statistical extraction methods.

Study Simulation

A “subject” consisted of 30 single-trial simulated waveforms, which were constructed following the above-mentioned procedure. Statistical extraction was conducted on the subject averages of the 30 trials. This process was repeated 1,000 times to provide a robust measure of the difference between the ERP peak without noise and the ERP peak with noise. Maximum amplitude of noise simulations began at 0 mV and incrementally increased by a step of 0.001 mV; the amplitude of the noise was allowed to vary randomly within the defined maximum amplitude of noise and increased for each subsequent 1,000-subject simulation up to 5.0 mV. For example, in the 2-mV noise simulation, the noise amplitude randomly varied for each trial between 0 and 2 mV (peak-to-peak amplitude of 0.0 to 4.0 mV). Frequency of the noise varied randomly for each noise sinusoid between 0.1 and 50 Hz and was applied to the entire simulated epoch. This wide range of frequency for noise was narrower than other simulation research (Yeung et al., 2004, 2007) and was chosen so as to reflect frequencies of noise that may remain after applying a band-pass filter to actual EEG data. Taken together, 5,000 1,000-subject simulations (each subject simulation contained 30 trials) were conducted for each sampling rate. Simulation sets were run for four commonly used sampling rates by downsampling from a simulated sinusoid: 250, 500, 1000, and 2000 Hz. We include different sampling rates in an effort to investigate the effects of temporal precision on latency measurements. All simulations and statistical analyses were implemented in the statistical software program R (R Development Core Team, 2011).

Data Analysis

To quantify noise, a method proposed by Schimmel (1967), which is readily accessible in open-source MATLAB toolboxes such as the ERP PCA Toolkit (Dien, 2010), was adapted to the current simulated ERP data. For the noise present in each subject average, a RMSE approach was implemented on the noise in isolation (i.e., before imposing the noise onto the simulated P1 peak). In order to calculate the RMSE for noise estimation, the amplitude of noise at every time point (sample) was first squared, then subsequently averaged. The square root was then taken of the averaged noise amplitude. The bias was calculated as the difference between the measurement without noise and the measurement with noise for each 1,000-study simulation (measurement with noise minus measurement without noise). The 95% confidence interval was ascertained to determine whether the confidence interval contained zero at steps of 0.5 μV noise estimates, such as 0.0 to 0.5 μV, 0.5 to 1.0 μV, etc. (i.e., no difference between the measurement without noise and the measurement with noise). The RMSE was used as a measurement of efficiency for each statistical extraction method; the difference between the measurement with and without noise was squared, averaged, and then square rooted. Taken together, the bias, efficiency, and noise estimate was calculated separately for each 1,000-participant simulation, resulting in 5,000 total estimations, to examine the effect of incremental increases in noise across a wide range of noise severity.

The peak-amplitude measure of the simulated P1 was extracted as the absolute amplitude at the time point corresponding to the maximum amplitude between 0 and 200 ms. The mean amplitude was calculated as the arithmetic average of the amplitude at every sample between 80 and 120 ms (the likely window that would have been chosen based on the grand average waveform of each

Figure 2. A: Example of five phase-randomized noise (co)sinusoids of various amplitudes and frequencies. B: Average of noise (co)sinusoids. Noise estimate = 0.31. C: The solid black line depicts the simulated half-cycle 2.5-Hz sinusoid with a peak amplitude of 5 μV (this activity was not depicted in [A]). This waveform was considered the “true” peak referred to as the P1. The gray line depicts the summation of the activity without noise combined with the average of noise (co)sinusoids that were shown in (A) and (B).
1,000-subject simulation as the mean of the P1 latency was 100 ms). An adaptive mean measure was also conducted, wherein the peak positive amplitude between 0 and 200 ms was identified, and then the average amplitude of 20 ms prior to the peak and 20 ms after the peak was extracted.

For latency measurements, peak latency was extracted as the time point at which the peak amplitude reached the maximum. A positive centroid latency analysis was also conducted using a previously proposed center-of-mass measurement from 0 ms to 200 ms (Dien et al., 2004; using corrected formula outlined in the ERP PCA Toolkit, see Dien, 2010). The formula used was \( \frac{\sum(t \cdot [v_t - \text{max}])}{\sum(v_t - \text{max})} \). In this formula, \( t \) represents the time point in the defined window and \( v_t \) is the voltage at time \( t \). For a minimum centroid (area under waveform; positive component), the max, maximum voltage in the time window, is used. For a maximum centroid (area over the waveform; negative component), the minimum voltage in the time window should occupy the place of max in the equation: \( \frac{\sum(t \cdot [v_t - \text{min}])}{\sum(v_t - \text{min})} \). The centroid measurement characterizes the central tendency of the ERP component latency by using the area under the curve. For each simulated participant, the average of the difference between the measurement of the P1 peak without noise and the P1 peak with noise was calculated to quantify the accuracy of the extraction measure.

Further, data from a previously published study from our lab were summarized to illustrate the noise levels encountered in real ERP data and demonstrate the relationship between the number of segments averaged and background EEG noise (Larson et al., 2012). The Larson et al. study examined the CRN and ERN components of the scalp-recorded ERP. Thirty-six individuals with mild traumatic brain injury and 46 neurologically healthy controls completed a color-naming Stroop task while EEG was recorded using a 250-Hz sampling rate. Participants responded via button press to the color of the font the word was printed in using the index, middle, and ring fingers of their right hand. Data were high-pass filtered at 0.1 Hz and low-pass filtered at 30 Hz. Trials consisted of words presented in their same color of font (congruent trial; RED written in red) or in a different color of font (incongruent trial, RED written in blue). Stimuli were presented for 1,500 ms followed by a 1,500-ms duration fixation cross. Equiprobable congruent and incongruent trials were presented in five blocks of 100 trials (500 total trials).

For current purposes, data were collapsed across the mild traumatic brain injury and control groups to include a healthy population and a neurologic population, which may potentially have increased noise (i.e., those that have experienced a traumatic brain injury). Noise was calculated by following a procedure similar to the one proposed by Schimmel (1967) described above using the ERP PCA Toolkit (Dien, 2010). Every other trial was inverted before single-subject averaging. The trials were then averaged to obtain an estimate of noise with the true nonrandom ERP signal averaging out. The RMSE was then calculated on the remaining ERP activity to quantify noise. Linear and quadratic regression equations were evaluated between numbers of trials of noise estimates separately for the CRN and ERN. Cook’s D was used to identify the presence of influential outliers.

**Results**

**Amplitude Measures**

Figure 3 contains scatter plots of the noise estimates and bias measurements of the P1 as a function of sampling rate for amplitude measures. Figure 4 contains scatter plots of the noise estimates and RMSE measures for P1 amplitude at each sampling rate. Figures 3 and 4 provide visual portrayals of the deleterious impact of increases in noise on ERP amplitude measurements.

The mean amplitude measure demonstrated the least bias for all sampling rates. Furthermore, the confidence interval of the mean amplitude measure contained zero for each 0.5 \( \mu \)V step indicating no difference between the mean amplitude on the P1 with noise and the mean amplitude on the P1 without noise, suggesting that the mean amplitude measure is an unbiased measure of amplitude. The RMSE for the mean amplitude was smallest relative to the RMSE for the peak amplitude and adaptive mean measures at each 0.5 \( \mu \)V noise-estimate step, suggesting that the mean amplitude is the most efficient amplitude measure.

For the adaptive mean measure, bias was increased relative to the mean amplitude measure, although bias was decreased compared to the peak amplitude approach. For each sampling rate, the adaptive mean was biased, which was exacerbated as noise estimates increased. The confidence intervals for the adaptive mean only contained zero for the 1.5 to 2.0 \( \mu \)V noise estimates for the 1000 and 2000 Hz sampling rates. Although the adaptive mean measure was indeed biased relative to the mean amplitude measure, the RMSE was comparable to the mean amplitude measure indicating that the adaptive mean exhibits similar efficiency to the mean amplitude.

The peak amplitude demonstrated high levels of bias and poor efficiency. The confidence interval for the differences between the peak amplitude measure with noise and the peak amplitude measure without noise never contained zero, suggesting that the peak amplitude never provided an accurate measure of the P1-amplitude peak measure without noise. Furthermore, the peak-amplitude measurement was consistently the least efficient measure. Indeed, across all sampling rates, bias and inefficiency linearly and rapidly increased with increases in levels of noise when using the peak amplitude measure.

**Latency Measures**

Figure 5 contains scatter plots of the noise estimates and bias measures of the P1 latency measures for each sampling rate. Figure 6 contains scatter plots of the noise estimates and RMSE measures of the P1 latency measurements.

The peak latency measure relative to the centroid latency measure was less biased and less efficient. The confidence intervals for the peak latency measure always contained zero, and the mean differences between measures with noise and without noise were minimal. However, the peak latency measure was consistently less efficient than the centroid latency measure as supported by higher RMSEs for each 0.5 \( \mu \)V noise-estimate step. As the sampling rate increased, the peak latency measure became less biased and more efficient.

Although the centroid latency measure was more efficient than the peak latency measure, the centroid latency measure was markedly more biased. Furthermore, the centroid latency measure was always biased. The RMSE was consistently smaller relative to the peak latency measure. Notably the bias of the centroid latency measure was closely associated with sampling rate such that bias was increased when the sampling rate was lower.

**Noise and Trial Counts**

Data from a published study were presented to characterize the levels of noise found in ERP experiments and examine the
relationship between the number of trials averaged and the level of noise present (Larson et al., 2012). Four individuals with Cook’s D estimates greater than 3 SDs from the mean for the model including noise estimates and CRN amplitude were excluded from the following CRN and ERN analyses to ensure that findings are not primarily the result of outliers. Thus, final analyses included 78 individuals. Error trials included an average standard deviation of 17 (range: 6 to 68); correct trials contained 478 (range: 430 to 498). Noise estimates averaged 2.37 (range: 0.90 to 5.20) for error trials and 0.39 (range: 0.24 to 0.69) for correct trials. Higher numbers of error trials were associated with a decrease in noise estimates as supported by a significant linear relationship, $F(1,76) = 58.77, p < .0001, R^2 = 43.6\%$ (see Figure 7); however, the quadratic trend was a better fit for the current data and accounted for more variance, $F(1,75) = 67.30, p < .0001, R^2 = 64.2\%$. Neither a linear nor a quadratic fit were significant for the number of correct trials and correct-trial noise estimates ($F_s < 1.3, p_s > .28, R^2$’s < 4%). These findings indicate that noise estimates quadratically decreased as the number of trials retained for averaging increased. Furthermore, the range of noise estimates from the abovementioned ERN study indicate that the current investigation examined noise estimates that are plausibly encountered in common ERP research.

**Discussion**

The primary purpose of the current investigation was to examine the robustness of area- and peak-based methods for ERP statistical extraction to the presence of simulated background EEG noise. Based upon the simulations, the mean amplitude and adaptive mean measurement methods clearly outperformed the peak amplitude measurement on indices of bias and efficiency. For latency measures, the peak latency measure was less biased than the centroid latency measure; however, the centroid latency measure was more efficient suggesting that the centroid latency measure is less variable than the peak latency measure. Although ERP researchers do their best to minimize noise during EEG recording and remove nonsystematic and systematic sources of noise during data collection and preprocessing, using appropriate ERP statistical extraction methods is another avenue to minimize the deleterious effects of noise.

A salient, although unsurprising, finding from the current investigation is that as noise increases ERP measurement accuracy decreases. As demonstrated by the presentation of ERN data from the Larson et al. (2012) study, the number of trials retained for averaging is closely associated with amount of noise present in the ERP waveform, such that increases in the number of trials retained for averaging was associated with decreases in noise estimates. This finding empirically demonstrates what is commonly assumed among ERP researchers. In studies of the ERN, it is commonplace to report the number of trials retained for averaging after artifact rejection and correction due to the potentially low number of trials. However, it would be informative to include noise estimates for ERP waveforms when comparing groups or conditions regardless of the ERP component being studied.
The importance of reporting noise estimates can be inferred from the current simulations. Each subject in the current examination contained an identical number of trials; however, the trials retained for averaging contained varying levels of background noise. Comparing the number of trials retained for averaging between individuals with a high-noise estimate to individuals with a low-noise estimate would show no group differences despite clear group differences in noise estimates. Considering that levels of noise affected the bias and efficiency of ERP statistical extraction methods in the present simulations, comparing noise estimates for groups would yield important information about possible alternative explanations for group differences (i.e., measurement bias as a result of noise) above and beyond what would be gleaned from reporting only the number of trials retained for averaging. Although it may be expected that background EEG noise would be comparable for each participant when tested in identical environments with the same data acquisition equipment, differences in impedance measurements or background electrical noise between individuals or across time, for example, may result in differences in the consistency of measured signal (see Kappenman & Luck, 2010). There are various proposed measures of noise available, and the reader is directed elsewhere for a more thorough discussion of the measures (Glaser & Ruchkin, 1976; Handy, 2005). The current examination did not compare methods for noise quantification; however, the present findings suggest that it is critical to quantify and compare noise estimates between groups and conditions. As recommended numerous times before (Fein & Turetsky, 1989; Perry, 1966; Turetsky et al., 1988), we suggest that, in addition to evaluating the number of trials retained for averaging examining group or condition, differences in background EEG noise would be beneficial and informative to ensure that findings are not primarily the result of noise.

Previous research demonstrating differences between conditions or groups using the peak amplitude measurement method may be primarily the result of differences in noise levels rather than true ERP signal differences. The peak amplitude measurement was consistently positively biased; that is, the peak amplitude measurement overestimated the true measurement value. As Luck (2012) cogently argued, the peak amplitude measure is noise-prone due to the peak amplitude approach choosing the most extreme value in a specified time window, and is inappropriate to use when comparing ERP waveforms from different conditions with averages containing different numbers of trials. The current investigation empirically supports this conclusion by evidencing how drastically increases in noise affect the peak amplitude measure. To reiterate the rifle example of bias and efficiency, the peak amplitude measurement is like a rifle that shoots consistently outside the inner ring of the target with high variability in shots. Considering that the peak amplitude method was biased and inefficient, it seems that the peak amplitude measure results in a quantification of the ERP amplitude that does not accurately represent the true population parameter being estimated (i.e., the true ERP amplitude without noise). Thus,

Figure 4. Scatter plots of noise estimate and the root means square error (RMSE) of each 1,000-subject simulation for amplitude measures: (A) 250 Hz, (B) 500 Hz, (C) 1000 Hz, and (D) 2000 Hz sampling rates.
the current simulations indicate that findings drawn from ERP peak amplitude measurement methods are unacceptably biased when high levels of noise are present in averaged ERP waveforms. This finding is particularly disconcerting, as the peak amplitude remains a widely used statistical extraction method for ERP amplitude measurement (42% of studies in our brief sample of the literature).

An extension of the peak amplitude measure is the peak-to-peak amplitude measure. In this measure, the amplitude is quantified as the difference between the peak amplitude of interest and the amplitude of the preceding opposite-polarity peak. Considering that the peak-to-peak amplitude measure simply takes the difference between the peak of interest and the previous opposite-polarity peak, the peak-to-peak amplitude measurement would be even more susceptible to bias from increases in noise than a simple peak amplitude measurement. Thus, findings from the peak amplitude measures could be extended to the peak-to-peak amplitude measure.

The peak amplitude approach may be used more frequently in studies examining more distinct components, such as the conflict N2 (e.g., Clayson & Larson, 2011a, 2011b; Forster, Carter, Cohen, & Cho, 2011), error-related negativity (e.g., Falkenstein et al., 1991; Gehring et al., 1993), or conflict P3 (e.g., Falkenstein, Stoerig, & Hohnsbein, 2008), relative to slower, tonic components, such as the late positive potential (e.g., De Cesarei & Codispoti, 2011; Hajcak & Nieuwenhuis, 2006) and post-error positivity (e.g., Overbeek, Nieuwenhuis, & Ridderinkhof, 2005), for which there is no distinct peak. When evaluating the percentage of articles using the respective ERP measures, the frequency of each measure based on the component of interest was not examined. As a result, the percentage of studies using the peak amplitude measure may be an underestimate for those studies for which it is possible to use a peak amplitude measurement approach.

With regard to noise and artifact correction, ERP waveforms from developmental and neurologic populations are typically contaminated with a large amount of ocular and movement artifact (e.g., Jung et al., 2000; Luck, 2005; Romo-Vazquez, Ranta, Louis-Dorr, & Maquin, 2007). Despite advances in removing this artifact, such as the use of independent component analysis (Delorme & Makeig, 2004), residual noise is likely increased in these populations even after correcting the systematic artifact. Indeed, if a measure was taken to reject trials that contained artifact altogether, at the very least a small number of trials retained for single-subject averaging would result in increased noise estimates. The use of the mean amplitude approach would be particularly beneficial when examining ERPs in these populations as there will likely be increased noise present in the ERP averages; however, some of these populations may show distinct differences in latencies relative to control populations. Current findings support the use of the adaptive mean measure over a peak amplitude measurement when individual-subject variability in latency affects the measurement of the ERP of interest.

Figure 5. Scatter plots of noise estimate and the bias measurement for each 1,000-subject simulation for latency measures: (A) 250 Hz, (B) 500 Hz, (C) 1000 Hz, and (D) 2000 Hz sampling rates.
Although the mean amplitude measure is the preferred amplitude measurement method (shots would be tightly clustered and consistently in the inner ring of the target), in the case of individual-subject variability in latency an adaptive mean measure is likely more appropriate to capture the true mode of the ERP signal (the adaptive mean measurement would represent shots from a rifle that would likely be in or just outside of the inner ring of the target but would have a low-to-medium amount of scatter). For example, in the current simulations, the latency of the P1 randomly varied between 85 and 115 ms. A mean amplitude window chosen based on the grand average waveforms of 80 to 120 ms may inaccurately capture the individual-subject peak as it may only partially capture the mode of the ERP peak that would occur whenever the P1 peak latency without noise was any value other than 100 ms. When individual-subject latency is considerable, such as when comparing neurologic or developmental populations, a mean amplitude approach may not appropriately extract the ERP peak. Indeed, a mean amplitude measure may spuriously bias results for one group that has the peak amplitude within the window of interest if the other group has a strong latency shift. Although the adaptive mean measure was slightly positively biased in the present simulations, the adaptive mean measurement was an efficient measure of the true P1—indicating that, despite the measurement being consistently different than the true P1 amplitude value without noise, the adaptive mean is consistently close to the true P1 amplitude. Thus, after demonstrating a similar number of trials retained for averaging between conditions (and/or groups) and similar noise estimates, the adaptive mean measure would likely be a better measure of the true P1 peak relative to the mean amplitude measure in instances of wide latency variability.

Another point to be considered when using an adaptive mean amplitude or peak amplitude approach is that the peak amplitude and adaptive mean amplitude measurements assume that the peak latency is known; however, the peak latency is an unknown quantity that must be estimated. As a result, this assumption may result in measurement imprecision. For example, when defining a time window in which to identify a peak latency based on a grand average waveform, an individual-subject ERP average waveform may not contain a definitive peak. When extracting the peak latency, a computer-automated approach may identify the point at which the ERP activity reaches a minimum or maximum in the absence of a definite ERP peak. Another possibility for imprecision when using this approach is when overlapping components may obscure the peak of the component of interest in a specified time window. For example, the peak latency within the specified time window may not actually correspond to the target component. These assumptions are not typically acknowledged and can result in increased measurement error above and beyond what the contribution of outside noise sources may contribute.

Figure 6. Scatter plots of noise estimate and the root means square error (RMSE) of each 1,000-subject simulation for latency measures: (A) 250 Hz, (B) 500 Hz, (C) 1000 Hz, and (D) 2000 Hz sampling rates.
For the latency measures, the peak latency measure was less biased than the centroid latency measurement; however, the centroid latency measure was more efficient than the peak latency measure. Thus, the peak latency measurement was like shots from a rifle that were in the inner ring of the target but were highly variable, whereas for the centroid latency measurement shots were likely outside of the inner ring but remained tightly clustered at higher sampling rates. To clarify, the centroid latency was more susceptible to bias as noise increased but was consistently less variable than the peak latency measure. The peak latency measure was consistently scattered around the true P1 peak latency measure with noise despite being more spread in terms of variability around the true P1 peak latency measure. However, statistical bias and efficiency were closely tied to sampling rates, and it is clear that when using a centroid latency measure it is more appropriate to use higher sampling rates.

Considering that the centroid latency measure was more efficient than the peak latency measure, it seems more appropriate to use the centroid latency despite its greater bias. When analyzing latency measurements, the primary concern is the relative differences in the time course of an ERP component between conditions or groups. Although the centroid latency measure may be off by a certain number of samples (i.e., more biased than the peak latency measure) in each estimation, it is likely a more accurate estimator of the relative differences between conditions given its higher efficiency when the same amount of bias is present in both conditions. Thus, when comparing latencies between conditions and groups, the centroid latency measurement is the recommended approach.

The mean and adaptive mean amplitude measures may be differentially biased by the frequency of noise in the ERP waveform. In the presence of high-frequency noise, these measures may show reduced bias than in the presence of low-frequency noise. For example, in order for the noise in the mean amplitude measurement to average out, a long enough time window needs to be selected to capture sufficient random noise activity around the true signal. In the presence of low-amplitude noise, when using a short time window the average of noise activity may not cancel out. Related to the time window chosen, it should further be noted that the ERP area- and peak-based measures in the current examination rely on a specification of a time window within which the amplitude or latency measurement is applied. Although researchers frequently select the time window for these measures based on the grand average waveforms and previous research guidelines, the fact that researchers are choosing the time window for measuring presents another source of possible bias above and beyond bias from the measurements themselves. Thus, it is important to consider how both the time window itself and the length of the time window selected for mean amplitude or adaptive mean amplitude measurement may affect the bias of the measure. Further, high-pass and low-pass filters may impact the frequency of noise present in the data. Readers are referred to Luck (2005) Chapter 5 and Edgar, Stewart, and Miller (2005) for in-depth chapters on the effects of filtering on ERP data and noise levels.

A remaining consideration would be how much bias or efficiency is acceptable in a given measurement method. It is unclear whether increases in bias between various levels of noise for the mean amplitude approach, for example, are meaningful. The degree of acceptable inaccuracy may depend on the ERP components of interest. For example, in studies where differences between conditions may be small, such as the conflict N2 (e.g., Clayson & Larson, 2011a; Forster et al., 2011), these small differences in ERP amplitude measurement bias may have a more influential impact relative to larger condition-related differences in components, such as the late positive potential (e.g., De Cesarei & Codispoti, 2011; Hajcak & Nieuwenhuis, 2006). Although it would always be better to record EEG with as little background noise as possible, whether the presence of significant background EEG noise compromises ERP measurement depends on the robustness and size of the ERP of interest as well as the expected size of condition or group differences. Thus, appropriately measuring the component of interest and reporting noise-related information are important aspects of any group-related ERP study.

An additional implication of knowing noise-related information is as it relates to measurement error and statistical significance. Any increase in measurement error would decrease the likelihood of obtaining statistical significance when true differences exist, as increases in measurement error result in a reduction of power (Baguley, 2004; Charter, 1997; Williams & Zimmerman, 1989). Measurement error results in an underestimation of the magnitude of effects when using standardized effect sizes; thus, by reducing measurement error, the standardized effect size will increase subsequently increasing power (Baguley, 2004; McClelland, 1997) and producing narrower confidence intervals (Charter, Adkins, Alekoumbides, & Seacat, 1987). Certainly, measurement error cannot be reduced to zero during EEG recordings as this is to some extent

---

**Figure 7.** A: Scatter plot of the relationship between number of trials retained for averaging for the correct-related negativity (CRN) and noise estimates. B: Scatter plot of the relationship between number of trials retained for averaging for the error-related negativity (ERN) and noise estimates. Note the different scales for CRN and ERN scatterplots for the x and y axes.
determined by recording equipment and background neural processes during recording; however, how ERP amplitudes and latencies are quantified greatly influences measurement error as evidenced in the current examination. Choosing those methods that demonstrate the smallest bias would increase the likelihood of obtaining statistical significance in ERP research, as power would be increased relative to when using more biased measures that increase measurement error.

Overall, findings from the simulations indicate that the peak amplitude measure should rarely be used. Although the mean amplitude measure is the preferred measurement for amplitude statistical extraction, an adaptive mean measure better captures individual-subject variability in latency. For the latency measures, the centroid latency measure is a more efficient although more biased measure than the peak latency measure; however, the centroid latency measure should be used only when EEG is recorded with a high sampling rate (approximately 1000 Hz). Lastly, the current examination is a call for greater vigilance in collecting clean EEG data and for transparency in ERP measurement by coupling measurements with trial counts and noise estimates for ERP waveforms. Such a practice would improve confidence that significant findings between different groups or conditions are the result of true ERP signal differences rather than simply differences in background EEG noise.

References


(Received May 25, 2012; Accepted October 25, 2012)