Best Practices for Repeated Measures ANOVAs of ERP Data: Reference, Regional Channels, and Robust ANOVAs

Joseph Dien[1]

[1] Maryland Neuroimaging Center, University of Maryland, 8077 Greenmead, College Park, MD 20742, USA.

Short Title: ANOVA power

Dien, J. (2017). Best Practices for Repeated Measures ANOVAs of ERP Data: Reference, Regional Channels, and Robust ANOVAs. International Journal of Psychophysiology, 111(1)42-56. (DOI: 10.1016/j.ijpsycho.2016.09.006)

Address for correspondence: Joseph Dien, Maryland Neuroimaging Center, University of Maryland, 8077 Greenmead, College Park, MD 20742, USA.

Phone: 202-297-8117. E-mail: jdien07@mac.com. URL: http://joedien.com.

Abstract

Analysis of variance (ANOVA) is a fundamental procedure for event-related potential (ERP) research and yet there is very little guidance for best practices. It is important for the field to develop evidence-based best practices: 1) to minimize the Type II error rate by maximizing statistical power, 2) to minimize the Type I error rate by reducing the latitude for varying procedures, and 3) to identify areas for further methodological improvements. While generic treatments of ANOVA methodology are available, ERP datasets have many unique characteristics that must be considered. In the present report, a novelty oddball dataset was utilized as a test case to determine whether three aspects of ANOVA procedures as applied to ERPs make a real-world difference: the effects of reference site, regional channels, and robust ANOVAs. Recommendations are provided for best practices in each of these areas.

Key Terms: ERP, ANOVA, Statistics

1.0 General Introduction

Event-related potentials or ERPs are scalp-recorded electrical brain activity that are time-locked to an event of interest such as the presentation of a sound. This method generates very rich datasets that present many choices for even basic analyses such as repeated measures analysis of variances (ANOVAs). In a typical analysis, the researcher applies an ANOVA with within-subject factors for condition and electrodes and possibly time. Between-subject factors may also be used to test group differences but will not be considered in the present report. The advent of high-density recording montages has exacerbated this situation, providing researchers with a plethora of channels.

It is important for the field to develop evidence-based best practices for three reasons. First of all, to minimize the Type II error (missing a true effect) rate by identifying the most effective analytic approach. For example, many common procedures have not been fully re-examined in light of technological advances such as high-density montages (approaches that were effective for three electrodes may not be the optimal choice for 256 electrodes). Second, to minimize the Type I error (falsely detecting an effect) rate by reducing the latitude for varying procedures. While multiple comparison procedures are widely used to control the number of tests performed, there are no such common procedures for potential Type I error rate inflation due to choosing amongst multiple analytic procedures (i.e., shopping around for the best p-value). Establishing guidelines for best practices could help constrain such analytic choices. Third, to identify areas for further methodological improvements. The identification of weaknesses in current procedures can help guide further research into better ways to analyze the data.

While generic treatments of repeated measures ANOVA methodology are available, ERP datasets have many unique characteristics that must be considered. A previous review of the average reference (Dien, 1998b) discussed how the choice of the reference site has a critical effect,

usually ignored, on repeated measures ANOVAs. A later tutorial (Dien & Santuzzi, 2005) reviewed the logic behind repeated measures ANOVA and how it relates to ERP data, especially highlighting its statistical assumptions and the use of regional channels (virtual channels formed by averaging together a cluster of electrodes). Both papers provided recommendations based on a detailed examination of the underlying principles and their ramifications but did not provide empirical evaluations. This present study builds on these prior reports by empirically-based evaluation of the recommendations to develop evidence-based best practice guidelines. While this report is intended for both beginning and experienced investigators, it will be assumed that readers have taken graduate level statistics.

This report evaluates three recommendations with regards to repeated measures ANOVAs of ERPs: the choice of reference site, utilization of regional channels, and application of robust ANOVAs. It will be demonstrated that these recommendations are worthy of consideration because they concern statistical choices that can have substantive effects on results.

1.1 The Example Dataset

To conduct this empirical evaluation of best practices, a real dataset was employed as a test case. While simulations have the advantage that the exact nature of the signal is known, they are not good at answering whether a procedure makes real world difference (see Beauducel & Debener, 2003) or whether unexpected aspects of the data have an effect. Instead, a real dataset will be used utilizing a very simple and well-understood paradigm where it can be said that the correct results are already known, based on the existing extensive literature. The dataset will be an already published novelty oddball dataset (Spencer, Dien, & Donchin, 1999a; Spencer, Dien, & Donchin, 2001), which is already well-described. While it is desirable to present fresh data for experimental papers (where it can at least provide additional information about replicability), for a methodological paper like this it is advantageous to utilize an existing dataset so that the effects of changes in methodology between papers are not confounded with changes in the datasets.

In this particular dataset, a combination of frequent standards, rare targets, and rare novel sounds like dog barks were presented. It is generally accepted that this paradigm will produce both a P300 (Donchin & Coles, 1988; Sutton, Braren, Zubin, & John, 1965), which is a parietal positivity that is maximal to rare targets and peaks at 300 ms or later depending on the decision time, and a P3a (Courchesne, Hillyard, & Galambos, 1975; Polich, 2007), which is a central positivity that is maximal to stimuli outside the attentional channel and peaks at about 300 ms. As revealed (Simons, Graham, Miles, & Chen, 2001; Spencer, Goldstein, & Donchin, 1999b; Spencer et al., 2001) by principal components analysis (PCA), rare targets produce a strong P300 and a weak P3a whereas rare novel sounds produce a weaker P300 and a strong P3a (Figure One). This dataset was collected with fifteen participants with a 129-channel high density Electrical Geodesics Incorporated electrode net.

1.2 Quantifying Statistical Power

For the most part, statistical power will be measured in terms of p-values, as the most familiar metric of effect size. Formally speaking, statistical power (the probability of detecting an effect), the significance criterion (alpha, the probability of a false alarm), the sample size, and the effect size (the magnitude of the condition effect scaled in some manner by its variability) are interconnected, with three of the parameters determining the fourth (Cohen, 1988, p. 14). When one holds the sample size and the alpha criterion constant, as in the present case, effect size determines statistical power. One can therefore utilize effect size measures, of which there are a variety (Lakens, 2013). In experimental reports it is recommended to provide a measure of effect size in addition to p-values because it removes the influence of sample size, facilitating comparisons of studies with differing sample sizes. From this standpoint, a p-value can be thought of as being a combined measure of effect size, sample size, and assumed sample mean distribution (one computes an effect size measure then generates a p-value based on degrees of freedom and assumed sample mean distribution).

For this methodological study where one is only comparing between methods using the same dataset there is no need to separate out the effect of sample size. In addition, effect sizes (and hence power) are not yet available for the robust ANOVA procedure utilized in Study Three (a robust ANOVA procedure is an ANOVA calculation that is intended to be less susceptible to certain weaknesses of the standard ANOVA calculation and will be fully defined in the Study Three introduction). In any case, the influence of some aspects of the robust ANOVA procedure (i.e., bootstrapping, Welch-James statistic) will be seen in the p-value (and in statistical power) but not in a typical effect size measure. Finally, p-values are more familiar to many readers than effect sizes. Thus, p-value is a more useful metric than effect size for the present methodological report (although experimental reports, where researchers will definitely be concerned with comparing across studies with different samples sizes, should indeed present effect sizes where possible).

2.0 Study One - Reference Choice

2.1 Study One - Introduction

One important but underappreciated influence on ANOVA results is the reference site. Voltages, as in electroencephalography (EEG) recordings, are measures of electrical potential (that is, the potential for current flow). As such, they are relative measures that require a comparison between two points. For example, a nine-volt battery has a nine volt difference between its terminals. Between one terminal and some other location the voltage difference would be something different. It is not meaningful to speak of the terminal as having a voltage value in isolation.

The conventional procedure in EEG research is to specify an electrode as the reference site, which is to say all the other electrodes will be contrasted against it. Since the "true" voltage is not known, this reference site is arbitrarily defined as being zero voltage. Any voltage activity at the reference site ends up being mathematically attributed to the electrodes being compared to it (so a negativity at the reference site shows up as a positivity at all the other electrodes). One can also mathematically rereference a dataset by the simple procedure of subtracting the waveform (X) at the new reference site from all the sites including itself; as a result, the new reference site will have zero voltage (X-X=0) and the old reference site will be an opposite sign version of the new reference site's original waveform (0-X=-X).

Typically the mastoids are used as the reference site. It is common practice to mathematically average the two mastoids together (i.e., mean mastoid reference); the usual justification is that doing so avoids hemispheric bias. A mean mastoid rereference cannot reduce both of them to zero and so they will split their difference, as mirror images of each other (to ensure they sum to zero) so any asymmetries have not been eliminated. The mastoids were originally chosen because they were distant from the recording electrodes but with today's high-density montages this is no longer true. This author has heard it claimed that the mastoids are a good reference site because it can be seen that electrodes near the reference site are largely flat but in truth this is a false inference - these electrodes appear flat because they are near the designated reference site and are therefore similar to the location that has been arbitrarily defined as being zero voltage.

One approach to dealing with this voltage ambiguity is the average reference (Bertrand, Perrin, & Pernier, 1985). The average reference is based on the biophysical principle that a voltage source in an enclosed conductive surface will result in a voltage distribution that sums to zero across its surface. One advantage of the average reference is that it is more true to the basic principle that all ERP components are (by biophysical necessity) bipolar with both a positive and a negative pole. The polarity labels of ERP components are determined by which pole is more prominent (higher magnitude and topographically focal) at the commonly used electrode locations on the top of the head and do not mean that the ERP components are in fact "positive" or "negative" as a label like "P300" or "N400" seems to imply. A vertically oriented (meaning that the generator site or sites is positioned such that the maximum voltage from one pole is near the top of the head and the maximum voltage from the other pole is at the bottom of the head, as seen in the overall scalp topography) ERP component like the P300 can appear to be wholly positive using mean mastoids reference simply because its negative pole is near the mastoids and so all the other electrodes are relatively more positive compared to it (its negative waveform has been subtracted from all the other waveforms, resulting in an added positivity) but in truth it too is bipolar. The average reference thus provides a more veridical depiction of the nature of an ERP component.

These reference issues are relevant to ANOVAs of windowed measures because, again, voltages cannot exist in isolation. They are always relative to another location. Thus, when one performs an ANOVA on windowed voltages, they implicitly contrast the measured electrode with the reference site (Dien, 1998b). This can benefit the ANOVA (improve the p-value) when one is measuring the voltage at one pole and the other pole is near the reference site. If, for example, the experimental effect voltage at the recording electrode is +x and the voltage at the reference site is actually -x, the rereferencing procedure has increased the effect at the recording electrode to +2x and one is getting twice the effect, increasing statistical power all things being equal. Conversely, if the recording electrode is near the reference site then one will lose statistical power.

It is likely that the common choice of the mastoids as the reference site has had the unintended consequence of favoring ERP components that are vertically oriented like the P300 or the N400. For example, if the reality is that the P300 is +5 microvolts at Pz and it is -4 microvolts at the base of the head (including the mastoids) then under mean mastoids reference it will register (in the waveform and thus in the windowed measure) as being +9 microvolts at Pz because rereferencing will add +4 microvolts to all the electrodes. Noise at the mastoids will also be reallocated but, to the extent that the noise is random noise, it will cancel out with noise at Pz, resulting in a doubled signal-to-noise. In contrast, a horizontal ERP component like the Positive Slow Wave where the positive pole is in the back and the negative pole is in the front has very little presence at the mean mastoids and hence if the reality is +5 microvolts at Oz and -4 microvolts at Fpz and zero microvolts at the mastoids then with mean mastoids it would still only be +5 at Oz and therefore not get as much of a boost as the P300.

If one needs an unbiased reference scheme, as when a number of ERP components of different orientations are of interest, then the average reference may be preferable. In the average reference, the effective reference point is the zero voltage isopotential line halfway between the recorded positive and negative poles. It therefore always splits the difference. In principle, it should not provide as much statistical power for vertically oriented ERP components like the P300 or N400 but it should provide more to those with other orientations.

The choice of reference site is therefore important because it directly impacts the statistical power of ERP analyses and yet this author cannot recall seeing an ERP paper in which the authors explicitly considered the possible effect of reference choice on whether their study achieved statistical significance. In general, researchers (including the present author) seem to have made an ongoing choice based on general considerations such as established practice or biophysical plausibility in the absence of systematic evaluation of the effect of reference choice. Who knows how many studies have been dubbed failures due to lack of evidence-based guidance on this matter? Are researchers who use average reference unnecessarily handicapping themselves? Are researchers who use mean mastoids reference unnecessarily biasing themselves to ERP components with a vertical orientation like the P300 or the N400?

The mean mastoid reference and the average reference are not the only reference options. A possible concern with the average reference is that the underside of the head is generally undersampled and so taking the mean of the recording electrodes will not in fact sum to zero. The degree to which it will be inaccurate is presently unclear due to the greater complexity of the bottom side of the head. In any case, one proposed solution is the polar average reference effect, or PARE, -corrected average reference (Junghöfer, Elbert, Tucker, & Braun, 1999). In this variant, the surface of the entire head is interpolated based on the measurement sites and then the average of this surface is used as the estimate of zero. One possible concern with this approach is that it necessarily gives greater weight to the electrodes at the fringe of the electrode montage, which will have the greatest influence on the interpolation of the underside of the head; these fringe electrodes tend to be less well-anchored and thus noisier. Whether this is a substantive concern will be examined as part of this report.

An alternative that has generally been ignored in the ERP literature is the bipolar reference. Technically a bipolar reference is obtained by using a dedicated reference site paired to each recording site rather than a single common reference site. One could in principle obtain the same effect by explicitly contrasting the positive and negative poles by windowing at both sites and subtracting them from each other. This procedure would result in a reference-independent solution that would potentially provide optimal statistical power (Dien, 1998b). One would just need to know the electrode locations of both poles in advance to do so *a priori*. A bipolar reference is already in effect used for the mismatch negativity or MMN studies, where it has become customary to use a nose reference (which is near the positive pole of the MMN). In principle, the Positive Slow Wave would probably also get a boost from using a nose reference since the pole opposite the usual measurement site (at the back of the head) is near the nose.

As already noted, the example will be a novelty oddball dataset previously published (Spencer et al., 1999b; Spencer et al., 2001) and then used as an example in a more recent tutorial with an updated analytic procedure (Dien, 2012). Unlike the 2012 tutorial (Dien, 2012), this report will use the average reference as the starting point. Doing so proved to be a good instructional example of the nuances of reference choice and the importance of careful attention to scalp topography. Looking at both reference results, it became clear that the "Frontal Negativity" (Dien, 2012; Spencer et al., 2001) was actually the Positive Slow Wave (Ruchkin, Munson, & Sutton, 1982; Squires, Squires, & Hillyard, 1975). As it turns out, one effect of using a mean mastoid reference rather than an average reference with a high-density dataset is that a positivity at the mastoids is subtracted from the rest of the electrodes, resulting in their being more negative; as a result, the Positive Slow Wave appears to be a strong frontal negativity (with a weaker posterior positive pole) whereas under average reference it appears to be a strong posterior positivity (with a weaker frontal negative pole). The putative Positive Slow Wave in the 2012 report was actually the latter portion of the P3b, split in the PCA from the early portion of the P3b by the presence of the P3a.

While general consensus positive peak electrodes are available from the literature for the P3a and P3b, the positive peak channel for the Positive Slow Wave and the negative peak channels for these three components are not available, so PCA will be used to provide candidate positive and negative peak channels for all three ERP components (keeping in mind that going forward one would then use these channels for *a priori* peak channel determination). For consistency's sake, these peak latencies and channels will also be used to determine the traditional windowed measures.

This first study will therefore examine the effects of reference on statistical power with the example oddball dataset. The goal will be to determine whether reference choice actually makes a difference when applied to a real dataset and, if so, what choice provides the best statistical power. By the reasoning outlined earlier, it is hypothesized that the mean mastoid reference should provide stronger results than the average reference when the latter yields activity at the mastoid that is in the opposite direction as the peak channel and weaker results when there is mastoid activity that is in the same direction as the peak channel. Based on the activity at the mastoids as characterized by the average reference (Figure Two), it is hypothesized that the mean mastoid reference will provide stronger results than the average reference for the P3a, modestly stronger results for the P3b, and weaker results for the Positive Slow Wave. It is also hypothesized (Dien, 1998b) that the bipolar reference should provide the strongest results for all the ERP components.

2.2 Study One - Methods

The P300 dataset was presented in a prior report (Spencer et al., 1999a). In the portion of the dataset used for the present report, fifteen participants (no participants were rejected as bad data) performed 300 trials of an active novelty oddball task. They were presented with frequent standard tones (12%), rare target tones (12%), and novel environmental sounds (76%). They were instructed to press a button to the targets and were not given instructions regarding the novel sounds. The data were recorded using a 129-channel GSN200 Geodesic Sensor Net (Tucker, 1993) with Cz as

reference and were then low-pass filtered at 20 Hz, and baseline corrected using the 150 ms prestimulus period. Further information is available from the original report (Spencer et al., 1999a).

Further processing was conducted using the EP Toolkit 2.5.1 (Dien, 2010b) running on Matlab 2015b under OS X 10.10.5 and 10.11.1. The data were rereferenced to mean mastoids (the average of e57 and e101), to average reference (the average of all 129 channels), and to PARE-corrected average reference. Bipolar referencing was accomplished by subtracting the opposite sign peak channel from the data prior to the t-test.

A two-step PCA procedure (Dien, 2010a; Spencer et al., 1999a; Spencer et al., 2001) was utilized to identify the peak channels for the ERP components. The first step was a temporal PCA using a Promax rotation with a kappa of 3 and utilizing a variance-covariance matrix (Dien, Beal, & Berg, 2005). Following a previous example analysis of this dataset (Dien, 2012), nine factors were retained based on the results of a Parallel Analysis (Dien, 1998a; Horn, 1965). For the second step, a spatial PCA using the EEGlab (Delorme & Makeig, 2004) Infomax rotation (Bell & Sejnowski, 1995) implementation was utilized and three factors were retained. In order to ensure replicability of the Infomax results, the random number generator (using the Twister option) was seeded with a zero before each independent components analysis (ICA) procedure and the line in runica that reseeds the random number generator was commented out. The factors corresponding to the ERP components of interest were identified based on *a priori* information about their characteristic latencies and scalp topographies. The factors are presented in Figure One so that readers may judge for themselves whether they agree. The PCA factors were not otherwise used except to determine windows and peak channels.

For the windowed measures, the windows of interest were centered at the peak latency sample and extended to the six samples before and afterwards (52 ms total). The two-tailed dependent measures t-tests were conducted using Matlab to compare the target cell to the standard cell for all but the P3a, where the comparison was between the novel cell and the target cell. The power analyses were conducted using G*Power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007), for

a matched pairs t-test with two-tails, alpha of .05, and power of .95. While no standards are available for setting an appropriate Type II error rate other than Cohen's suggestion that .80 be the minimum acceptable, .05 (.95 power) was chosen to be consistent with the accepted Type I error rate (which is .05).

As a further aid to interpret the results, power calculations were used to determine minimum required sample size.

2.3 Study One - Results

The average reference was used as the starting place as it is in principle more unbiased with respect to scalp topography than mean mastoids. The three relevant temporal factors (Figure One) had latencies of 292, 356, and 480 ms. The 292 ms temporal factor was split into a P3a factor with a positive peak channel of e6 (FCz) and a negative peak channel of e70 (just to the left of O1) and an early P3b factor with a positive peak channel of e54 (to the left of Pz) and a negative peak channel of e14 (just to the right of Fp2). The 356 ms temporal factor contained the late (majority) portion of the P3b with a positive peak channel of e55 (CPz) and a negative peak channel of e26 (to the left of Fp1). The 480 ms temporal factor contained the Positive Slow Wave with a positive peak channel of e79 (to the left of P4) and a negative peak channel of e17 (nasion).

For the windowed measures, based on these factor results, four windows of interest were specified: P3a, early P3b, late P3b, and Positive Slow Wave. Although there is no reason to think that the early P3b window is of theoretical interest (it should show the same effects as the late P3b) it provides a useful example of a weak effect. The results can be seen for all reference types in Table One. Table Two provides the information on the effect sizes, as well as the mean condition difference and the standard deviation of the condition differences used to compute them.

A PCA based on the mean mastoids reference was also run and while, as expected (Dien, 2012), the reference scheme did affect the factor solution somewhat, the decision was made to use the exact same windows used for the average reference measures to maintain consistency. When

run with windows derived from the mean mastoids PCA, the t-test results were comparable and did not change conclusions.

Reference	P3a		Early P3b		Late P3b		Positive Slow Wave	
	p	n	p	n	p	n	p	n
Average	.000062	9	.038	39	.000051	9	.0000083	7
PARE	.000064	9	.035	39	.000052	9	.0000084	7
Mean Mastoids	.0000080	7	.0049	20	.00060	13	.000015	7
Bipolar	.00027	11	.14	81	.00020	10	.0031	18

Table One. Dependent T-test Results for Different Reference Schemes. The p-value (p) and the sample size (n) required to achieve a power of .95 (in parentheses) are listed for each of the reference schemes. In all cases there were 14 degrees of freedom. The windowed measures were 52 ms windows centered on the each latency. See Section 2.3 for more information about the measures and how they were derived. PARE = polar average reference effect corrected average reference.

Reference	P3a			Early P3b			Late P3b			Positive Slow Wave		
	m	s	Dz	m	s	Dz	m	s	Dz	m	s	Dz
Average	5.44	3.74	1.45	2.21	3.73	0.59	6.46	4.35	1.49	6.50	3.69	1.76
PARE	5.45	3.8	1.45	2.24	3.71	0.60	6.47	4.37	1.48	6.48	3.69	1.76
Mean Mastoids	6.52	3.69	1.77	4.13	4.79	0.86	8.20	7.21	1.14	5.75	4.06	1.66
Bipolar	6.10	4.90	1.24	3.36	8.25	0.41	13.47	10.47	1.29	12.60	13.68	0.92

Table Two. Effect Sizes for Different Reference Schemes. The mean (m) of the difference scores, their standard deviation (s), and the Cohen's D_z score calculated from them are presented. Conditions are Novel minus Target for the P3a and Target minus Standard for the others. The windowed measures were 52 ms windows centered on the each latency. See Section 2.3 for more

information about the measures and how they were derived. PARE = polar average reference effect corrected average reference.

2.4 Study One - Discussion

As predicted, the P3a was more significant for mean mastoids than for average reference and less significant for the Positive Slow Wave (Table One), keeping in mind the prior caveats in Section 1.2 about the use of the p-value as a stand-in for effect size. Unexpectedly, the late P3b measure was stronger for the average reference. Surprisingly, the bipolar reference results were overall weaker, coming in last for all but the late P3b where it was intermediate between the other two references. Overall, the average reference performed well. The PARE-average reference results did not result in the feared increase in noise levels but also made little difference to the inferential tests.

Examination of Table Two illustrates how the reference effects were mediated not just by the position of the reference site relative to ERP component's scalp topography but also the quality of the channels. The Cohen's D_z effect size score for a dependent measures t-test is calculated by dividing the mean of the difference score by their standard deviation (Cohen, 1988, p. 48). This in turn directly relates to the t-test as the t-value is simply the D_z score multiplied by the square root of the sample size (Rosenthal, 1991). While the reference choices generally had the expected effects on the mean of the difference scores (the numerator), the standard deviations (the denominator) were also greatly affected. So whether a given reference choice overall did better or worse was determined by both parameters.

Looking at these two determinants of the statistical tests, it is apparent that as predicted the mean mastoids reference yielded a stronger numerator for the P3a, early P3b, and late P3b measures and a weaker numerator for the Positive Slow Wave (reflecting the nature of the activity at the mastoids, as characterized by the average reference). However, for the P3b measures and the Positive Slow Wave the denominator was markedly smaller, with the result that the late P3b measure was nonetheless more significant with the average reference. It is not clear why the denominator for

the P3a and the early P3b did not also improve to this same degree; it may be that the presence of both the P3a and the P3b in this early time window handicapped the average reference in some manner. In any case, while the bipolar reference had the strongest mean difference (up to twice as much), increased error variance swamped the improvement, resulting in an overall weaker effect size across the board (compared to the average reference).

The changes in variability largely reflect the effects of the reference channels. As can be seen in Figure Three, the use of the average reference overall reduced the level of noise as seen in the baseline period. The mean mastoid reference increased overall variability because it effectively took the noise in the two mastoid channels and pushed them out to all the other channels. The average reference appears to be less noisy in part because it was at least partly successful in isolating the mastoid electrode noise and in part because it effectively averages the reference noise level across the entire set of electrodes. As a demonstration of this principle, one can take the mean across the region surrounding the two mastoids (twelve total electrodes) and use the result as the reference; in this case one achieves a lower noise level, although not as low as that achieved by the average reference (all 129 electrodes), as seen in Figure Three. In contrast, the bipolar reference especially suffered as it relied on the noisy electrodes near the face (P3b and Positive Slow Wave) and the back of the head (P3a).

Overall, these results confirm the proposition (Dien & Santuzzi, 2005) that the reference choice affects ANOVA statistical power. As predicted based on whether one ERP pole coincided with the reference site, the mean mastoid reference yielded a stronger P3a result but a weaker Positive Slow Wave result than the average reference. A new insight provided by this study is that while the average reference has a general disadvantage of always (in effect) placing the reference halfway between the two poles, it appears to have a general advantage of decreased overall noise levels. Thus, the average reference actually had the advantage over the mean mastoid reference for the late P3b even though the mean mastoid received a modest boost from mastoid site activity.

It could also be seen that, at least for the P3b (the late part of it), a bipolar reference can yield stronger effect sizes compared to the mean mastoids reference, although it was not stronger than the average reference. For specialized applications like mobile EEG or brain-computer interfaces (BCI) where there may only be a handful of electrodes, effect sizes may particularly benefit from strategically placed reference electrodes if they are designed to provide high quality data.

Finally, at least for the present dataset, while the PARE-corrected average reference (Junghöfer et al., 1999) did not suffer from enhanced noise levels as feared, it also did not differ meaningfully from the results with the normal average reference. For this reason, the remainder of this report will use the more commonly utilized normal average reference, although the PARE-correction may indeed be preferable for general use.

3.0 Study Two - Channel Regions

3.1 Study Two - Introduction

The results of Study One highlighted that minimizing the standard deviation of the condition effects can help improve the effect sizes. As discussed previously, one way of doing so is to average together multiple channels into a regional measure. This procedure has the benefit of improving the signal-to-noise ratio by effectively increasing the number of waveforms going into the average. It also accommodates individual differences in the peak channel location (see Figure Four), although at the possible risk of diluting the effects in the peak channel with weaker non-peak channels. Such an approach has been recommended by ERP methodologists (Dien & Santuzzi, 2005; Luck, 2014). Two studies of test-retest reliability reported that it provided some benefit (Baldwin, Larson, & Clayson, 2015; Huffmeijer, Bakermans-Kranenburg, Alink, & van Ijzendoorn, 2014), which means it should also improve statistical power; the present treatment will build on these prior reports by examining this issue in more depth and by evaluating the results in terms of the more widely used p-value statistic.

It is therefore hypothesized that some modest benefit will be seen by using a regional channel measure (the mean of a cluster of electrodes) rather than a single channel measure. By this logic, it may also be beneficial to do so as well for reference sites, both mean mastoid (as seen in Figure Three) and bipolar. Indeed, it may be that part of the reason the average reference performed better than expected is that it essentially serves as a regional reference channel (with all 129 channels). It is also predicted that a regional channel measure centered on a peak channel (which should normally be chosen *a priori*) will be more effective than omnibus ANOVAs using either a generic array of regional channels, twelve regions (Dien & Santuzzi, 2005) comprised of 106 of the 129 channels, or a single factor with individual channels as levels. In the latter case the factor was limited to a set of 15 major 10-20 electrode sites to retain some interpretability and to keep it generally comparable to the omnibus regional channel analysis.

3.2 Study Two - Methods

Methods were as for Study One with some additions. The omnibus ANOVAs were conducted using SPSS 23. The power calculations were again conducted using G*Power, based on SPSS's partial eta squared output (and using G*Power's SPSS effect size type option). The power calculation function for G*Power for repeated measure ANOVAs has not yet been documented and the author's attempt to use it resulted in nonsensical results. All statistical tests were dependent t-tests except for the omnibus tests, which were ANOVAs using the G-G epsilon correction (Geisser & Greenhouse, 1958) for factors with more than two levels. The G-G epsilon correction was chosen since it is more conservative (indeed too conservative) than the alternative H-F epsilon correction (which is too liberal).

For the omnibus channels ANOVA, the factors were condition (two-levels as per the t-tests) and electrode (fifteen levels). For the omnibus regions ANOVA, the factors were condition (two-levels as per the t-tests), y-axis (anterior vs. posterior), x-axis (left vs. right), and z-axis (ventral, middle,

dorsal). The labeling for the ANOVA factors is inspired by the Talairach coordinate system (Talairach & Tournoux, 1988) used in MRI studies.

Regional channels were computed based on a central channel and the surrounding electrodes: P3a (5 6 7 11 12 13 107 113), early P3b (32 38 53 54 55 61 62), late P3b (32 54 55 62 80 81 129), Positive Slow Wave (62 68 78 79 80 86 87). For the mastoid regional reference the channels were: 50 51 56 57 58 63 97 98 100 101 102 108. For the bipolar references, the electrodes were: P3a (64 65 69 70 71 74 75), early P3b (8 9 14 15 17 126), late P3b (22 23 26 27 33 127 128), Positive Slow Wave (14 17 22 126 127). The omnibus channels test relied on standard 10-20 locations: 11 (~Fz) 25 (~F3) 34 (F7) 37 (C3) 46 (T3) 58 (T5) 60 (P3) 62 (Pz) 86 (P4) 97 (T6) 105 (C4) 109 (T4) 122 (F8) 124 (~F4) 129 (Cz).

For the regional omnibus, the electrodes were 106 of the 129 channels: left anterior ventral (22 26 33 39 44 45 127 128), left anterior middle (18 19 23 24 27 28 34 35 40), left anterior dorsal (7 12 13 20 21 25 29 30 36), right anterior ventral (1 8 14 115 120 121 125 126), right anterior middle (2 3 9 10 15 116 117 122 123), right anterior dorsal (4 5 107 111 112 113 118 119 124), left posterior ventral (56 57 63 64 69 70 74 75), left posterior middle (47 50 51 58 59 65 66 71 72), left posterior dorsal (32 38 43 48 52 53 54 60 61 67), right posterior ventral (83 89 90 95 96 100 101 108), right posterior middle (77 84 85 91 92 97 98 102 103), and right posterior dorsal (78 79 80 81 86 87 88 93 94 99).

3.3 Study Two - Results

The results are presented in Table Three.

Reference	P3a	Early P3b	Late P3b	Positive Slow Wave
AR- Regional	.000063 (9)	.0015 (16)	.000031 (8)	.000015 (8)

Peak				
MM-Regional Peak	.000014 (7)	.0013 (15)	.00081 (13)	.000037 (8)
MMR-Regional Peak	.000073 (9)	.00080 (13)	.000067 (9)	.0000099 (7)
Regional Bipolar	.00078 (14)	.067 (54)	.000082 (10)	.0018 (16)
AR-Omnibus Channels	C: .0034 CxE: .027	CxE: .0048	C: .0016 CxE: .00000026	CxE: .000000097
AR-Omnibus Regions	C: .0080 CxZ: .0031 CxYxZ: .0050	C: .027 CxZ: .037	C: .00043 CxY: .0013 CxZ: .00012	C: .0041 CxY: .000080 CxZ: .0070
MM-Omnibus Channels	C: .0015 CxE: .027	C: .029 CxE: .0048	CxE: .00000026	CxE: .000000097
MM-Omnibus Regions	CxZ: .0031 CxYxZ: .0050	C: .027 CxZ: .037	CxY: .0013 CxZ: .00012	CxY: .000080 CxZ: .0070

Table Three. T-test and ANOVA Results for Different Regional Electrode Schemes. The p-value and the sample size required to achieve a power of .95 (in parentheses) are listed for each of the electrode schemes. Sample size computations were not available for ANOVAs. AR=average reference. MM=mean mastoid reference. MMR=mean mastoid region reference. C=cell. E=electrode. X=left vs. right electrodes. Y=anterior vs. posterior electrodes. Z=dorsal vs. middle vs. ventral electrodes. For ANOVAs, only significant effects involving the condition factor are listed.

3.4 Study Two - Discussion

In general, the regional channels did not consistently improve the results compared to the peak channel results. Furthermore, the pattern of differences between the average reference and mean mastoids reference became muddled. The regional peak channels did provide better p-values than the omnibus measures as expected. The average reference continued to perform comparably to the mean mastoid regional reference.

The regional channel approach did not have the expected consistent beneficial effects. Close examination of the data suggested that although the error variance (the denominator of the effect size) did generally decrease, the effect on the numerator was mixed. Presumably the reason is that the additional channels sometimes had an overall detrimental effect, diluting the desired signal. This was especially evident for the regional mean mastoid reference for the P3a, compared to the regular mean mastoid reference.

The regional peak approach did improve results over the single peak approach somewhat for the average reference, notably for the early P3b measure, but was largely a wash for the mean mastoids. A meaningful improvement was seen for the mean mastoid regional peak approach for all but the P3a measure. Using regional channels at both ends of the bipolar approach also yielded modest improvements for all but the P3a measure. The mean mastoid regional reference yielded some improvements for the early P3b measure over the conventional mean mastoid reference and did better than the average reference for two measures and worse for the other two. So overall it did appear that regional channels yielded some modest net improvement compared to single-electrode measures. Of some concern is that the expected differences between the reference schemes (mean mastoid reference better with P3a and average reference better with Positive Slow Wave) became more muddled, suggestive that the use of regional channels complicated matters by introducing additional factors (e.g., degree of presence of the ERP component in the non-peak channels).

Turning to the omnibus measures, they yielded lower p-values as expected. Targeted regional channels generally provided stronger results than the untargeted omnibus regional channel approach. The reference choice was essentially irrelevant for the omnibus regional channel approach. The reference did not affect interactions including channel factors due to the way the statistic is computed, which essentially rereferences the data to average reference (based on the channels included in the ANOVA). With regard to the condition main effect, the average reference results are somewhat arbitrary insofar as it is dependent on what channels were included in the omnibus test; if every channel was included then the condition main effect would always zero out (the channels would sum to zero in each condition). Conversely, the condition main effect for the mean mastoid omnibus regional channel approach reflects the extent to which the mean of all the channels differ in each condition, which from the standpoint of biophysics is probably mostly reflecting the condition effects in the mastoid region (which has been rereferenced to all the other channels). For omnibus regional ANOVAs, it seems best to just ignore the condition main effect (at least for high-density montages, where an average reference is justifiable) in favor of interaction effects, which have the advantages of being reference-independent, more statistically powerful (at least for this dataset), and more informative about scalp topography.

Thus far there seems to be a small advantage for the regional peak average reference approach as both the involvement of multiple channels in both the regional peak and the reference computation has generally reduced noise levels. This observation was consistent with prior reports (Baldwin et al., 2015; Huffmeijer et al., 2014) that used largely the same hardware and average reference (PARE-corrected average reference in the former case) to examine test-retest reliability effects, although the present findings introduce caveats regarding the effects of reference scheme and the effects of scalp topography and were overall less clearly positive.

4.0 Study Three – Robust ANOVA

4.1 Study Three - Introduction

Conventional ANOVAs (and t-tests, which are a special case of ANOVAs) are susceptible to three issues: deviations from normality, sensitivity to outliers, and unequal variances. One way of addressing these issues is to use a robust ANOVA procedure. One such implementation (Keselman, Wilcox, & Lix, 2003; Lix & Keselman, 1995; Wilcox & Keselman, 2003) is computed in the same manner as a conventional ANOVA but has modifications to address the three issues. This implementation is available through the EP Toolkit, based on SAS/IML code provided by Lisa Lix. In a prior informal comparison (Dien, Franklin, & May, 2006), it provided results comparable to that of conventional ANOVAs.

The first ANOVA issue is that they assume normality, relying on the Central Limit Theorem (Fischer, 2010) to assure that sample means from the population can be assumed to be normally distributed even if the population itself is not; however, the general consensus is that the theorem applies only when the sample size is at least thirty (e.g., Gravetter & Wallnau, 1992), which is often not the case in ERP studies, as in the present dataset. As seen in Figure Five-a, the windowed measurements in the present oddball dataset are not normally distributed, indicative that the population distribution is likely not normally distributed either; thus, for this data one is indeed relying on the Central Limit Theorem to ensure that the sample means are normally distributed. Not only does the present data only have fifteen observations, the Central Limit Theorem does not actually guarantee normality even for samples of larger than thirty (Bradley, 1980; Westfall & Young, 1993); in the present case, the Central Limit Theorem performs reasonably well but does lose some power due to a failure to fully achieve the assumed normal distribution, as can be seen in Figure Five-b. If the sample mean distribution fully achieved normality, the histogram would conform to the brown normal line and 5% of the distribution would fall into the tails beyond the alpha threshold. The number provided below each histogram shows the actual percentage falling into the tails and hence the

degree to which the lack of normality has affected the test results. Numbers less than .05 indicate a loss of power by the conventional ANOVA due to a mismatch between the expected normal distribution and the actual distribution (as estimated by the bootstrapping procedure).

The robust ANOVA procedure addresses such non-normal distributions by using bootstrapping (Efron, 1979; Wilcox, 2010), a method to estimate the shape of the sample mean distribution by repeatedly drawing subsets from the sample and compiling the results(Efron, 1979; Davidson & MacKinnon, 2000). In effect, the members of the sample are considered to be representative of the population distribution (with unlimited copies of each observation) and then a series of random simulated samples are generated by drawing from the sample with replacement (allowing for multiple draws of the copies of an observation) to empirically generate an estimate of the actual sample mean distribution. Thus the bootstrapping procedure will have some degree of random variability that will be reflected in the resulting p-value. The more simulation runs in the bootstrapping, the more stable the resulting estimate but also the more time-consuming the procedure. This is an issue inherent in any procedure with a stochastic element, including independent components analysis (Dien, Khoe, & Mangun, 2007).

An interrelated concern is that of random number generation. Computers are not capable of generating truly random numbers. Instead they use pseudo-random number generation in which a starting number, a seed, is subjected to a set of mathematical operations that result in numbers that are not predictable by humans (see Deng & Lin, 2000; Park & Miller, 1988). This seed is usually obtained either from the millisecond clock time (for varying results) or from the user (to allow for replications). The fact that statistical significance may depend on a wholly arbitrary choice of seed should be of clear concern to any empirical researcher and gives new meaning to the concept of "massaging the data" if one can simply try out different seeds (or keep rerunning the analysis if the seed is provided by the clock time) until one obtains significance. The question then is whether this unavoidable randomness is large enough to affect results and, if so, whether it can be managed to an acceptable degree. This is a question that has not thus far been addressed by the developers of the

robust ANOVA procedure as they have been focused on developing the deeper mathematical aspects of the technique and yet this should be of critical concern to any applied researchers who might be considering adopting it in their work.

One question of interest for the present report is how many bootstrapping simulations would be appropriate for ERP data. While bootstrapping has been extensively applied to ERP data for the purpose of analyzing within-subject effects (Di Nocera & Ferlazzo, 2000; Rosenfeld et al., 2008; Wasserman & Bockenholt, 1989), the number of simulations required has not been systematically evaluated for the purpose of replacing conventional between-subject ANOVAs. The number 599 has been recommended for bootstrapping simulation runs (Wilcox, 2010). The reasoning was in part because to be an exact test it has to yield an integer number when entered into the equation: a * (ß+1) where a is the alpha level and ß is the number of simulation runs (Hall, 1986; Racine & Mackinnon, 2007). The effect of increasing numbers of simulation runs will be examined in order to identify an optimal setting. This question can also be posed as an evaluation of how robust the results are to changes in the starting seed.

The second ANOVA issue is being overly sensitive to outliers. As can be seen in Figure Six, while the oddball ERP components replicate quite reliably at the group level, there is substantial variability in the individual subject averages. An overly good participant, as well as an overly bad one, can result in the Type I error of missing a true effect, because it will increase the error variance even more than the condition effect (Jolliffe & Lukudu, 1993; Wilcox, 2010). The present robust ANOVA procedure addresses outliers by the use of trimmed means and winsorized variances/covariances. For trimmed means, some percentage of the smallest and the largest observations in each cell is dropped, with a 20% rate (40% total) being recommended (Keselman, Algina, Lix, Wilcox, & Deering, 2008; Wilcox, 2010; Wilcox, 2012). This procedure drops the anomalous values, resulting in a more robust estimate of the central tendency of the cell. To accomplish the same for variances and covariances, the n extreme observations can be replaced with the value of the n-1 observation in a process termed winsorizing (Dixon & Yuen, 1974).

The potential drawback is that this process also reduces the degrees of freedom, in effect reducing the size of the sample. The cost-benefit ratio depends on the amount of trimming and the extent to which it is eliminating non-outliers versus outliers. This is another concern that will be examined next in the context of a typical ERP dataset.

The third ANOVA issue is that they assume homogeneity of variance and covariance. For within-subject factors, the standard practice for conventional univariate ANOVAs is to apply an epsilon correction factor to adjust the degrees of freedom (Box, 1954; Geisser & Greenhouse, 1958; Huynh, 1978), but the common ones have the known drawback (Maxwell & Arvey, 1982) that they are either overly conservative (G-G) or are overly liberal (H-F). Furthermore, for between-subject tests, the conventional test statistic is not suitable when group sizes are unequal (Bradley, 1980; Ramsey, 1980). An approach for addressing this problem is to use the Welch-James statistic (Johansen, 1980; Welch, 1938; Welch, 1947; Welch, 1951), which does not make the assumption of equal variances. This technique has been extended to within-group factors (Keselman, Carriere, & Lix, 1993; Lix & Keselman, 1995). Many statisticians (Best & Rayner, 1987; Fagerland & Sandvik, 2009; Ruxton, 2006) recommend routinely using the Welch-James statistic over the Student's t-test (and its ANOVA extensions presumably) as it performs comparably in the case of equal variances while being more robust to cases of unequal variances. There does not appear to be any notable drawback to using this statistic, but it will be of interest to see how results compare to that of the standard ANOVA procedure. Note that whereas the standard practice for conventional within-subject ANOVAs is to present the epsilon correction factor and the uncorrected degrees of freedom, the standard practice for robust ANOVAs is to present the corrected degrees of freedom.

Thus, this study will examine the effects of some robust ANOVA parameters (trimming levels, number of bootstrapping runs, effect of random seeds) to determine optimal settings, compare the performance of robust ANOVAs to conventional inferential tests with this ERP dataset, and examine effects of outlier ERP observations in general. It is hypothesized that robust ANOVAs will yield sufficiently substantial differences from conventional ANOVAs (presumably beneficial) that it makes

sense to consider their usage despite the difficulties posed by their absence from commercial statistics packages.

4.2 Study Three - Methods

In addition to the methods used in Study Two, robust ANOVA procedures were applied. They were carried out via the EP Toolkit's Matlab implementation of the SAS/IML code posted by Lisa Lix (http://homepage.usask.ca/~lml321/SAS_Programs.html). Power calculations were not available since the effect sizes presented by the SAS/IML code are only intended for between-group contrasts and should be ignored for within-group contrasts (personal communication, Lisa Lix, December 2015).

For the purpose of investigating the effects of the robust ANOVA parameters, the average reference regional channel measure was used for simplicity's sake. Variability was measured in terms of two times the standard deviation. As discussed in Section 4.4, twice the standard deviation provides a 95.45% confidence interval that can be used to judge whether the observed range of p-values still meets the alpha criterion level of .05 . According to this proposal, if the full range of the confidence interval (e.g., .01 +/- .002) meets the .05 alpha threshold, then the significance criterion is considered to be fulfilled. In effect, the alpha threshold is reduced by twice the standard deviation of the p-values (e.g., .05 - .002 = .0498).

To examine the effects of number of bootstrapping runs, t-tests were run on the four measures with 99 to 899 bootstrapping simulations runs in increments of 100, 999 to 8999 in increments of 1000, and 9,999 to 99,999 in increments of 10,000. Eleven repetitions were made at each level, with seeds for the pseudo-random number generator running from 100 to 1,100 in increments of 100.

To examine more deeply the range of random variability, 1,000 seeds from 1 to 1,000 were used for each of the t-tests on the four measures using 4,999 bootstrapping simulation runs.

4.3 Study Three – Results

For the four regional peak measures, the improvement in p-value variability seemed to start leveling out at about 4,999 simulation runs (Figure Seven). At this level, two standard deviations of the p-value variability were: P3a (0.0013), early P3b (0.0015), late P3b (0.0008), and Positive Slow Wave (0.0004). The elapsed time for these eleven repetitions on a Mac Pro with a 6-core 3.33 GHz Intel Xeon CPU was about 1.5 seconds.

In order to have a better sense of the potential variability due to seed choice, 1,000 seed values were utilized at the 4,999 simulation runs level. Variability of the p-values (two standard deviations) was comparable across the measures: P3a (0.0012), early P3b (0.0017), late P3b (0.0007), and Positive Slow Wave (0.0005). Or posed in terms of range: P3a (0.0002-0.0038), early P3b (0.0012-0.0070), late P3b (0-0.0022), and Positive Slow Wave (0-0.0012).

The effects of trimming levels on the four average reference regional channel measures is reported, using 4,999 simulation runs and the median p-value of eleven repetitions, in Table Four. Examination of the numerator and denominator of the effect size revealed no trend for the denominator but a clear trend in the numerator. The dependent measures t-test was computed with both difference scores and with separate paired scores as the trimming procedure interacts with these two cases differently.

%Trim	0 (15)		.07 (13)		.13 (11)		.20 (9)		.27 (7)	
	d	p	d	p	d	p	d	p	d	p
P3a	0.0016	0.0016	0.0024	0.0078	0.0078	0.03	0.0054	0.073	0.0194	0.043
	3.9/.70		3.7/.69		3.6/.69		3.5/.61		3.4/.65	
Early P3b	0.0036	0.0036	0.0060	0.0058	0.022	0.025	0.070	0.041	0.12	0.045
	2.8/.71		2.7/.60		2.6/.66		2.5/.72		2.3/.84	
Late P3b	0.0006	0.0006	0.0034	0.0026	0.0032	0.0014	0.0058	0.0020	0.031	0.011

	5.8/.97		5.7/1.1		5.4/.99		5.2/.96		5.2/1.1	
Positive	0.00 02	0.00 02	0.00 12	0.00 08	0.00 30	0.00 32	0.00 96	0.00 60	0.00 72	0.014
Slow Wave	6.0/.93		5.9/1.1		5.8/1.2		5.6/1.2		5.4/.90	

Table Four. Effects of Trimming Levels on the Four Regional Channel Measures. The top row is the %age trimming and the number of resulting observations in parentheses. The median p-value of eleven repetitions is reported for a test of difference scores (d) and then for a test of paired scores (p). In parentheses is the effect size numerator and denominator for the difference scores.

For a more fine-grained examination of the effects of each observation on the final p-value, leave-one-out ANOVAs were computed for each observation of the four measures (Figure Eight). It can be seen how in general dropping the middlemost observations tended to have the most detrimental effect whereas dropping the extremes resulted in lesser effects. Dropping the strongest P3a effect (4.2 microvolts) actually improved the p-value (.0002) over retaining it (.0016).

Based on these findings, the four measures were analyzed with robust t-test and ANOVAs.

Reference	P3a	Early P3b	Late P3b	Positive Slow Wave
AR-Regional Peak	.00160	.0036	.00060	.00020
MM-Regional Peak	.000000	.0048	.00040	.0016
MMR-Regional Peak	.000000	.0080	.00020	.0016
Regional Bipolar	.00080	.068	.00080	.0022
AR-Omnibus Channels	none	none	none	none
AR-Omnibus Regions	C: .012 CxYxZ: .057	CxZ: .050*	C: .013 CxY: .0030	C: .0058 CxY: .000000

			CxZ: .0050	CxZ: .018 CxYxZ: .046*
MM-Omnibus Channels	none	none	none	none
MM-Omnibus Regions	CxYxZ: .057	C: .023 CxZ: .050*	CxY: .0030 CxZ: .0050	CxY: .000000 CxZ: .018 CxYxZ: .046*

Table Five. Robust T-test and ANOVA Results for Different Regional Electrode Schemes. The p-value is listed for each of the electrode schemes. Sample size computations were not available for robust ANOVAs. AR=average reference. MM=mean mastoid reference. MMR=mean mastoid region reference. C=cell. E=electrode. X=left vs. right electrodes. Y=anterior vs. posterior electrodes. Z=dorsal vs. middle vs. ventral electrodes. Only significant effects including the condition factor are listed. *=2 x standard deviation of replications failed to confirm. "none" means that no results are available because the computation failed due to near-singularity.

4.4 Study Three - Discussion

First, an approach for mitigating p-value variability was determined. Then, robust ANOVAs were applied to the dataset and compared with the results using conventional ANOVAs. Overall, results suggested that while the p-values were weaker than with the conventional ANOVAs, the pattern of results were also more in line with expectations, suggesting that these p-values were truer to the data.

Systematic examination of the robust ANOVA output confirmed that variability of p-values from the bootstrapping procedure is not so great as to preclude its use but is sufficiently large that it needs to be managed. The results suggest that 4,999 bootstrapping simulation runs might be a better choice than 699 (Wilcox, 2010), providing an acceptable level of p-value variability for the present typical ERP dataset while being sufficiently rapid. Even so, there was still some amount of

variability. Two standard deviations of the p-value variability for the four measures was at worst 0.0015, meaning that for most tests it should be well below the alpha threshold.

Although the presence of such p-value variability is somewhat unsettling, at least it is explicit and can be managed; the alternative is to accept, in principle, a deterministic conventional ANOVA procedure with unknown error. Based on these observations, the approach taken in this approach was to run the analysis eleven times using a standard set of seeds, reporting the median value, and confirming that the p-value still meets the alpha threshold at double the standard deviation. Failure to confirm has been treated as a borderline significant effect. Other more complex approaches (Andrews & Buchinsky, 1998; Davidson & MacKinnon, 2000) have been proposed and can be evaluated in the future.

It was determined that at least for a typical ERP dataset such as this, the 20% trimming recommended elsewhere (Wilcox, 2010) was definitely not beneficial. Even a modest 6% trimming (two participants out of the full fifteen) resulted in a loss of statistical power. Closer examination of the results suggests that the loss of power is due to a combination of loss of degrees of freedom and a reduction of the effect size numerator. While trimming would be expected to have a neutral effect on a symmetrical distribution, the present data had a negative skew (Figure Five) that resulted in a net loss to the mean condition difference. Furthermore, although loss of the largest observation can have an overall beneficial effect by reducing the denominator variance (Figure Eight), it appears that again the negative skew meant that the accompanying loss of the smallest observation outweighed the benefits. While more investigation is called for, it is suggested that no more than 5% trimming be used for typical ERP datasets, rounded down so that for samples smaller than twenty no trimming is performed at all. There is also no clear pattern on whether the trimming procedure is better applied to a t-test of two separate values or of a single difference score.

Although the overall robust ANOVA p-values were not as significant as those from the conventional ANOVAs, the pattern was closer to the expected pattern, suggesting that the robust ANOVA values were indeed more accurate or at least more readily interpretable. The mean mastoid

reference and the regional mean mastoid reference results showed a stronger sensitivity to the orientation of the ERP components, with more significance for the vertical P3a and less significance for the horizontal Positive Slow Wave, than was seen for the average reference. This predicted pattern of results was stronger for the robust ANOVA (Table Five) than that for the conventional ANOVA (Table Three). Also notable is that the mean mastoid regional reference did not display the same benefit for the robust ANOVA, perhaps showing that the robust ANOVA addressed the same error variance addressed by the regional reference. Although the apparent loss of statistical power is unfortunate, if indeed it was simply being more accurate, then in principle it could go either way and for other datasets one could find the opposite pattern.

6.0 General Discussion

Overall, this author recommends using robust ANOVAs as protection against spurious findings, even if there is some loss of statistical power. Furthermore, the results suggest that at least for a high-density montage, average reference provides an approach that is relatively unbiased with respect to ERP component orientation while providing reasonable statistical sensitivity. In part this is because it effectively provides a regional reference channel, which is less noisy since it averages together multiple channels. Where statistical sensitivity is paramount and the ERP component orientation is favorable (i.e., vertically oriented), a mean mastoid reference can be deployed, ideally using regional mastoid channels. Of course, there are other considerations beyond statistical power when making a reference choice (Dien, 1998b).

As for channel arrangements, the regional peak channel approach provided minimal benefit over that of a single peak channel, but was at least overall comparable and provides some insurance against modest inaccuracies in the *a priori* choice of a peak channel and bad channels. In general, when it is possible to determine *a priori* the location of the peak channels, then a regional peak channel seems better than an exploratory omnibus regional channel approach. When using an omnibus regional channel approach, the main effect should be ignored. Regardless, bipolar

arrangements should be avoided unless both peaks are known to be located in low noise electrode regions, in which case it would in principle be the best choice, or unless the hardware is limited to a handful of carefully situated electrodes.

An alternative approach not addressed herein is that of PCA. Instead of applying ANOVAs to a windowed measure one can use PCA to generate dependent measures (factor scores) and apply ANOVAs to them. Given the complexities of the PCA of ERPs (Dien, 2006; Dien, 2010a; Dien, 2012; Dien, 1998a; Dien et al., 2005; Dien et al., 2007), this author generally recommends using PCA as a complement to windowed measures rather than as a replacement.

A caveat regarding these recommendations is that they were limited to the available hardware (a high-density 128-channel EGI system). Other EEG systems may very well have characteristics that would necessitate different conclusions. For example, some systems are designed to ensure especially high quality recordings from the mastoid electrodes, which is not the case for the EGI system (which is instead optimized for rapid high-density montages that are well suited for the average reference). In general, the recommendations regarding regional channels arrangements and average reference are likely to be most appropriate for high-density montages. The present report provides a template that researchers can use to evaluate their own systems.

The present conclusions were also limited by the nature of the example dataset. It is quite possible that ERP components with other scalp topographies or statistical parameters might display differing characteristics. It would therefore be prudent for researchers to conduct their own comparisons, using the present report as a guide. One might expect that just as MMN researchers have found it most effective to standardize on a nose-reference, other ERP components might be best studied with reference sites customized for their particular topography. While there is some merit to using a standardized reference site to verify what ERP components one is observing (as illustrated by the "Frontal Negativity" in the present dataset), once that has been done (perhaps in a figure) there is no reason not to customize the reference choice to optimize statistical power, especially when studying one of the smaller ERP components. While the majority of ERP studies

have focused on large robust ERP components like the P300 and the N400, neuroimaging studies have graphically revealed the extent to which such a simple approach fails to recognize the richness of the neural systems. For example, it is well-known that the lateral surface of the temporal lobes have a central role in language comprehension and yet a focus on mean mastoids reference is less likely to detect signals emanating from this region (Dien, 2009) since the reference site is located close by (reference choice will reallocate effects at the reference sites to the other electrodes rather than eliminating them but will result in a diffuse topography that will be easier to overlook). Average reference may be best (for high-density montages) when a number of ERP components are of interest as a general compromise while customized references (taking into account both electrode montage characteristics and ERP scalp topography) may be better when a single ERP component is the focus of interest.

The present report's conclusions were also limited by the use of a real dataset rather than a simulated dataset. Simulated datasets have the advantage that the true answer is known, but are limited by the degree to which they are truly representative of real datasets (Beauducel & Debener, 2003). Thus, the finding that channel noise was an important contributor to the results might have escaped notice in an artificial dataset. On the other hand, since the true nature of the experimental effects is not exactly known, it cannot be known for certain that an improvement in the p-values is always indicative of a more accurate procedure. What can be said for certain is that the novelty oddball paradigm is well known to produce P3a, P3b, and Positive Slow Wave effects and so significance should be obtained. It can also be said that where a robust statistic resulted in a less significant p-value, it is at the least more conservative than a conventional ANOVA and by design such a difference should reflect greater statistical accuracy. The important observation with respect to the present robust ANOVA results is that even when the p-levels were worse, the increased rigor did not come at an undue cost to statistical power.

7.0 Conclusions

In conclusion, this report illustrates how decisions about how an ANOVA is applied to ERP data can have substantial effects on the statistical power and thus whether true effects are detected. It has been demonstrated how the choice of reference can affect the data and guidance has been provided under what conditions the mean mastoids or the average reference might be appropriate for obtaining optimal results. Evidence has been provided on the effects of different options for aggregating channels and how they might affect the results. Finally, the argument has been made for using robust ANOVA statistics, with the caution that doing so may reduce statistical power even as it protects against spurious results, and suggestions have been made on the optimal settings for doing so (i.e., number of bootstrap runs, determining the variability of the p-values and reducing the alpha threshold accordingly, degree of trimming). Regardless, it is recommended that researchers be consistent in their analysis procedure across studies, and to justify deviations from it, to avoid the appearance of massaging the data.

Figure Legends

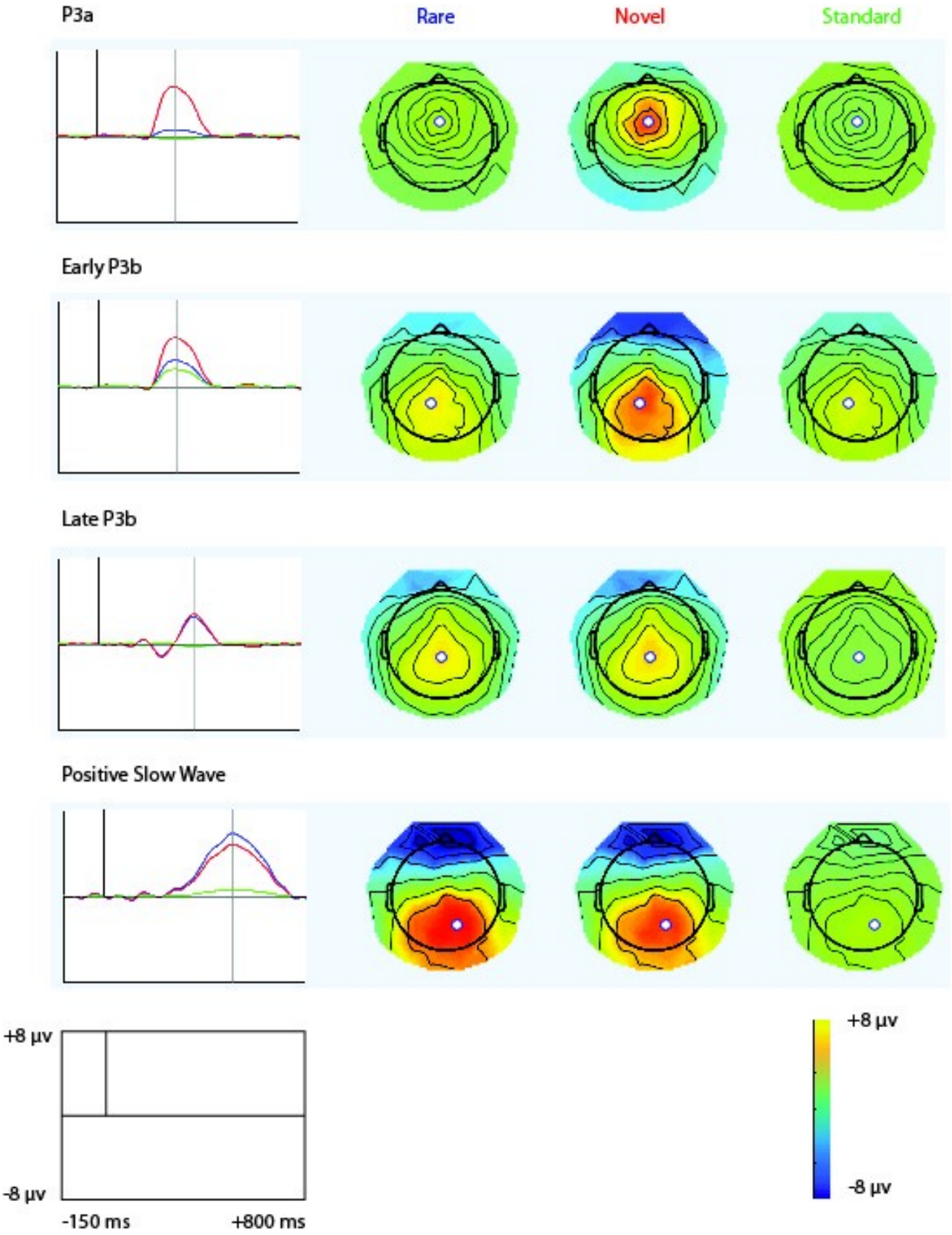
P3a
Rare
Novel
Standard
Early P3b
Late P3b
Positive Slow Wave
+8 µv
-8 µv
-150 ms
+800 ms
+8 µv
-8 µv

Figure One. ERP components as characterized by a two-step PCA. This temporo-spatial PCA of the average reference data displays the four factors of interest. These were then used as the basis for the four ERP measures used throughout this report. The white dot indicates the positive peak channel. The P3a and the early P3b were derived from different spatial factors of the same temporal factor.

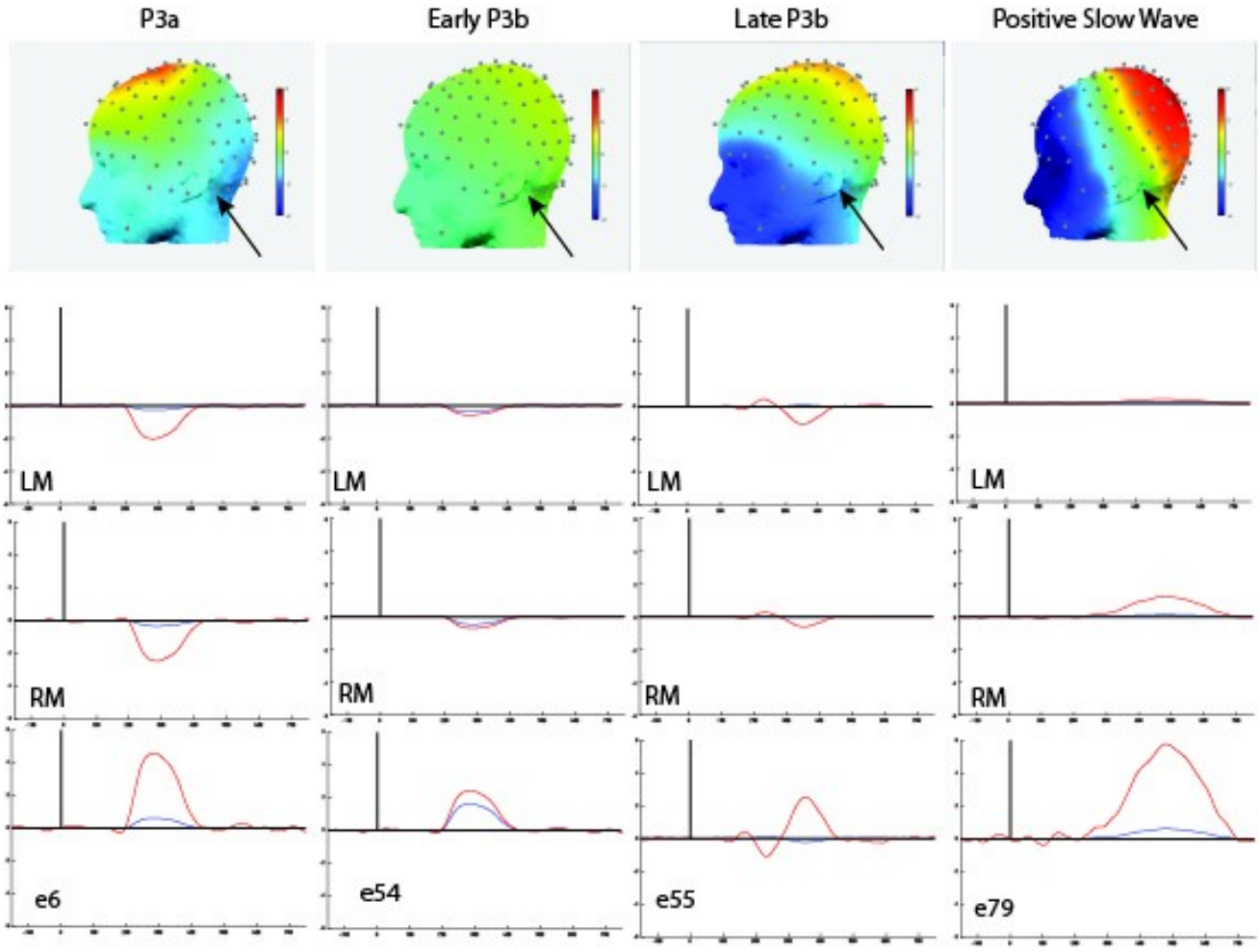

Figure Two. Waveforms at the mastoids and peak channels for the four measures. The Maps present the scalp topography of the four measures (corresponding to the Novel-Target and the Target-Standard difference waves) at the peak time point, as characterized by the two-step PCA of the average reference dataset. The black arrow indicates the location of the mastoid electrode.

LM=left mastoid. RM=right mastoid. The last row of waveforms are those of the peak channels. The scale of the figures are -6 to +6 μv. The zero isopotential line in the scalp maps is light blue.

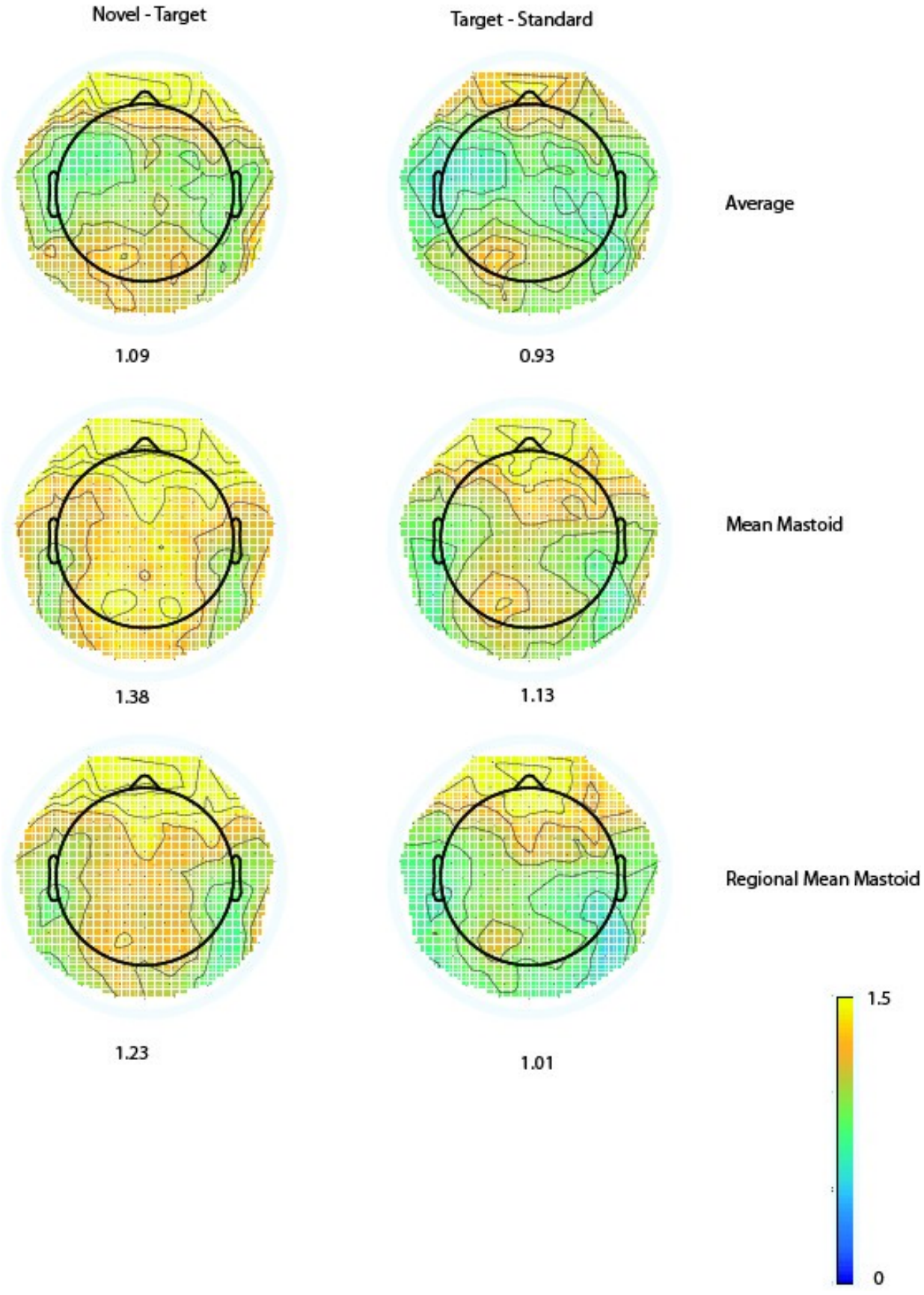
Novel - Target
Target - Standard
Average
1.09
0.93
Mean Mastoid
1.38
1.13
Regional Mean Mastoid
1.23
1.01
1.5
0

Figure Three. Topographical map of noise levels in the baseline period. The Maps present the mean of the standard deviation of the voltages across the 152 ms period of the individual subject averages (corresponding to the Novel-Target and the Target-Standard difference waves). The mean across the entire set of electrodes is also presented beneath each topographical map. The Target-Standard map is overall less noisy as more trials were included in the Standard averages.

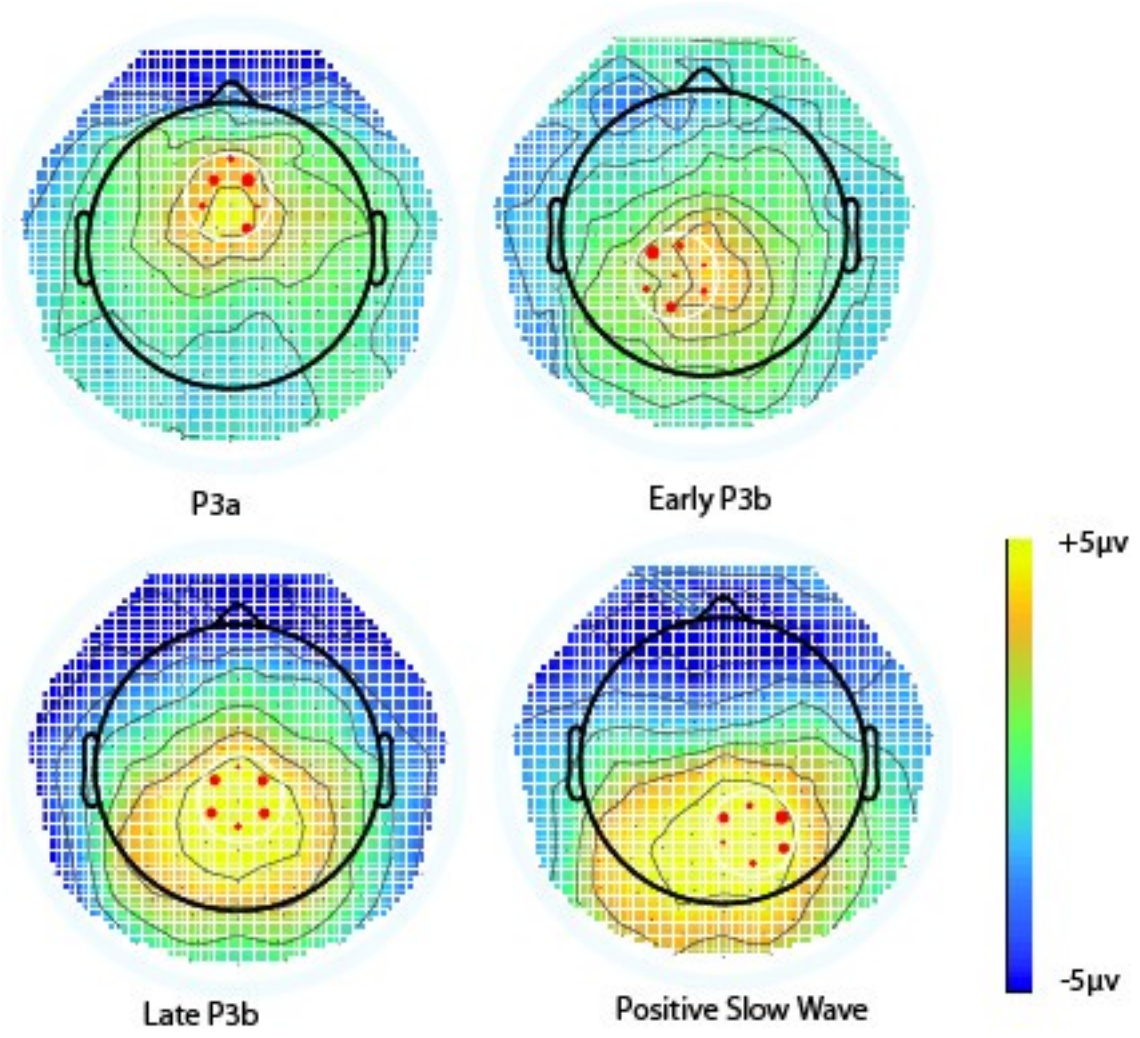

Figure Four. Variability of peak channels of the four measures. Figures show the scalp topography of the difference waves (novel-target for P3a, target-standard for the other three) collapsed across the window for each average reference grand average effect. The white circle shows the extent of the regional channels. The peak subject channels (out of those in the channel region) is indicated

with a red dot whose size is proportional to the number of peak channels at that electrode. Only channels within the regional channel region are considered for this figure; the overall peak channel is often not within the regional channel region.

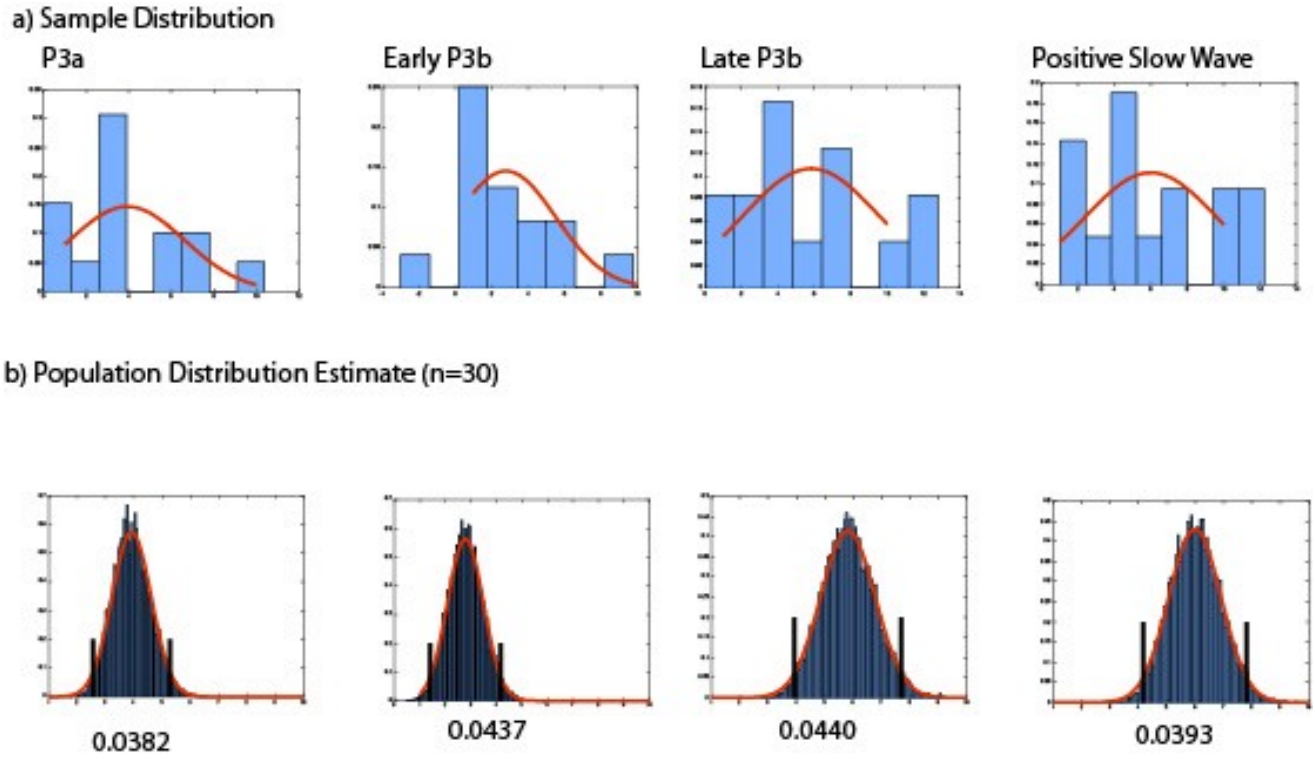

Figure Five. Histogram of sample distributions. The histograms present the difference scores (novel-target for P3a, target-standard for the other three) for the average reference regional peak channel. The first line (a) consists of the sample scores. The red line indicates the normal distribution expected by the conventional ANOVA. The second line (b) consists of the bootstrap estimate of the sample distribution based on resampling of the sample data with an n of 30 (which is the minimum number often cited as necessary for the Central Limit Theorem to apply). The black bars indicate the .05 alpha threshold based on the conventional ANOVA procedure. The number beneath these graphs indicates the proportion of the bootstrapped samples that meet the conventional alpha threshold of .05.

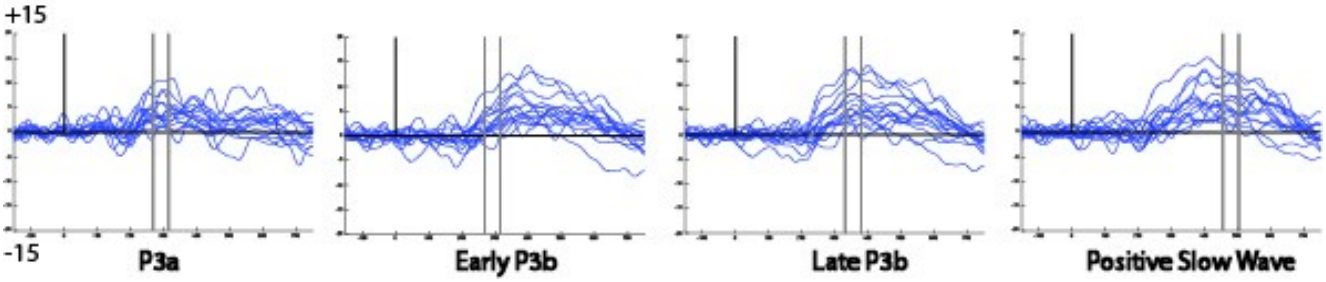

Figure Six. Difference waves of the four ERP measures. The difference waves (novel-target for P3a, target-standard for the other three) are the average of the channels comprising the average reference regional channels. The two lines indicate the boundaries of the measure's window.

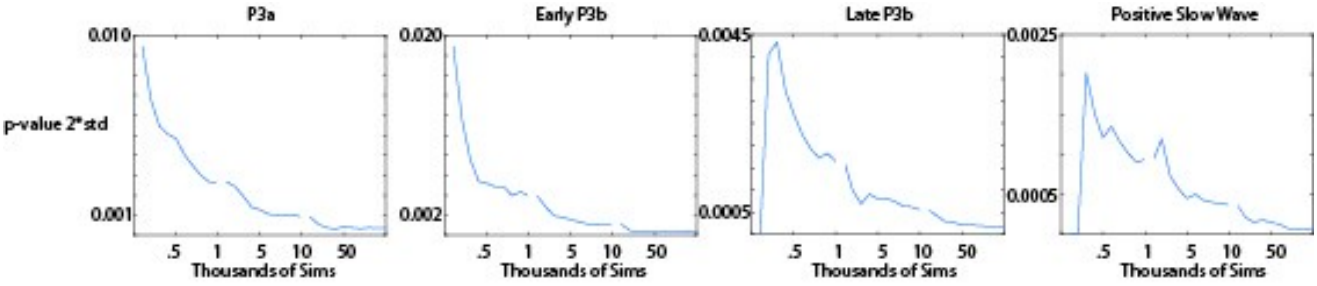

Figure Seven. Variability of p-values versus number of bootstrap simulation runs. For the four average reference regional channel measures, the variability of eleven repetitions of the robust ANOVA at different numbers of bootstrapping simulations runs, expressed as twice the standard deviation.

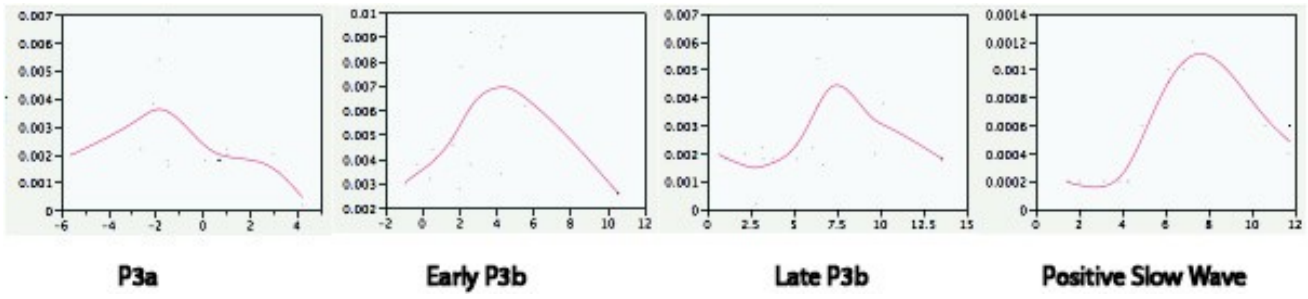

Figure Eight. Leverage plots for the observations of each measure. For the four average reference regional channel measures, each observation's condition effect (x-axis) is plotted against the p-value

(y-axis) obtained when it is dropped from the robust ANOVA (median p-value of eleven repetitions, 4,999 simulation runs, no trimming). The red line is the best fit to the observations (using a smoothing spline with a lambda of 1, a parameter controlling the degree to which the line is smoothed).

Acknowledgements

Supported in part by the Therapeutic Cognitive Neuroscience Fund, Johns Hopkins Medical Institutions, Barry Gordon, PI.

References

Andrews, D. W. K., & Buchinsky, M. (1998). *On the number of bootstrap repetitions for bootstrap standard errors, confidence intervals, confidence regions, and tests* (Cowles Foundation Discussion Paper No. 1141R, revised.). Yale University.

Baldwin, S. A., Larson, M. J., & Clayson, P. E. (2015). The dependability of electrophysiological measurements of performance monitoring in a clinical sample: A generalizability and decision analysis of the ERN and Pe. *Psychophysiology*, *52*(6), 790-800.

Beauducel, A., & Debener, S. (2003). Misallocation of variance in event-related potentials: Simulation studies on the effects of test power, topography, and baseline-to-peak versus principal component quantifications. *Journal of Neuroscience Methods*, *124*, 103-112.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129-1159.

Bertrand, O., Perrin, F., & Pernier, J. (1985). A theoretical justification of the average reference in topographic evoked potential studies. *Electroencephalography and Clinical Neurophysiology*, *62*, 462-464.

Best, D. J., & Rayner, J. C. W. (1987). Welch's approximate solution for the Behrens‐Fisher problem. *Technometrics*, *29*(2), 205-210.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The annals of mathematical statistics*, *25*(2), 290-302.

Bradley, J. V. (1980). Nonrobustness in Z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society*, *16*(5), 333-336.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Courchesne, E., Hillyard, S. A., & Galambos, R. (1975). Stimulus novelty, task relevance and the

visual evoked potential in man. *Electroencephalography and Clinical Neurophysiology*, *39*, 131-143.

Davidson, R., & MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps. *Econometric Reviews*, *19*(1), 55-68.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9-21.

Deng, L.-Y., & Lin, D. K. J. (2000). Random number generation for the new century. *The American Statistician*, *54*(2), 145-150.

Di Nocera, F., & Ferlazzo, F. (2000). Resampling approach to statistical inference: bootstrapping from event-related potentials data. *Behav Res Methods Instrum Comput*, *32*(1), 111-119.

Dien, J. (2006). Progressing towards a consensus on PCA of ERPs. *Clin Neurophysiol*, *117*(3), 699-702; author reply 703.

Dien, J. (2009). The Neurocognitive Basis of Reading Single Words As Seen Through Early Latency ERPs: A Model of Converging Pathways. *Biological Psychology*, *80*(1), 10-22.

Dien, J. (2010a). Evaluating Two-Step PCA Of ERP Data With Geomin, Infomax, Oblimin, Promax, And Varimax Rotations. *Psychophysiology*, *47*(1), 170-183.

Dien, J. (2010b). The ERP PCA Toolkit: An Open Source Program For Advanced Statistical Analysis of Event Related Potential Data. *Journal of Neuroscience Methods*, *187*(1), 138-145.

Dien, J. (2012). Applying Principal Components Analysis to Event Related Potentials: A Tutorial. *Developmental Neuropsychology*, *37*(6), 497-517.

Dien, J. (1998a). Addressing misallocation of variance in principal components analysis of event-related potentials. *Brain Topography*, *11*(1), 43-55.

Dien, J. (1998b). Issues in the application of the average reference: Review, critiques, and recommendations. *Behavior Research Methods, Instruments, and Computers*, *30*(1), 34-43.

Dien, J., Beal, D. J., & Berg, P. (2005). Optimizing principal components analysis of event-related

potential analysis: Matrix type, factor loading weighting, extraction, and rotations. *Clinical Neurophysiology*, *116*(8), 1808-1825.

Dien, J., Franklin, M., & May, C. (2006). Is “blank” a suitable neutral prime for event-related potential experiments? *Brain and Language*, *97*, 91-101.

Dien, J., Khoe, W., & Mangun, G. R. (2007). Evaluation of PCA and ICA of simulated ERPs: Promax versus Infomax rotations. *Human Brain Mapping*, *28*(8), 742-763.

Dien, J., & Santuzzi, A. M. (2005). Application of repeated measures ANOVA to high-density ERP datasets: A review and tutorial. In T. Handy (Ed.), *Event-Related Potentials: A Methods Handbook* (pp. 57-82). Cambridge, Mass: MIT Press.

Dixon, W. J., & Yuen, K. K. (1974). Trimming and winsorization: A review. *Statistische Hefte*, *15*(2-3), 157-170.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*, 357-374.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 1-26.

Fagerland, M. W., & Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemp Clin Trials*, *30*(5), 490-496.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*, *39*(2), 175-191.

Fischer, H. (2010). *A history of the central limit theorem: From classical to modern probability theory*. Springer Science & Business Media.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box’s results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics*, *29*, 885-891.

Gravetter, F. J., & Wallnau, L. B. (1992). *Statistics for the Behavioral Sciences* (3rd ed.). St. Paul, Minnesota: West Publishing.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval.

The Annals of Statistics, 1453-1462.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R., & van Ijzendoorn, M. H. (2014). Reliability of event-related potentials: the influence of number of trials and electrodes. *Physiol Behav*, *130*, 13-22.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika*, *43*, 161-175.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, *67*(1), 85-92.

Jolliffe, I. T., & Lukudu, S. G. (1993). The influence of a single observation on some standard test statistics. *Journal of Applied Statistics*, *20*(1), 143-151.

Junghöfer, M., Elbert, T., Tucker, D. M., & Braun, C. (1999). The polar average reference effect: A bias in estimating the head surface integral in EEG recording. *Clinical Neurophysiology*, *110*(6), 1149-1155.

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology*, *40*, 586-596.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*(2), 110.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. *Journal of Educational and Behavioral Statistics*, *18*(4), 305-319.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol*, *4*, 863.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on

testing for mean equality. *Psychological Bulletin*, *117*(3), 547-560.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). MIT press.

Maxwell, S. E., & Arvey, R. D. (1982). Small sample profile analysis with many variables. *Psychological Bulletin*, *92*(3), 778-785.

Park, S. K., & Miller, K. W. (1988). Random number generators: good ones are hard to find. *Communications of the ACM*, *31*(10), 1192-1201.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clin Neurophysiol*, *118*(10), 2128-2148.

Racine, J. S., & Mackinnon, J. G. (2007). Simulation-based tests that can use any number of simulations. *Communications in Statistics– - Simulation and Computation*, *36*(2), 357-365.

Ramsey, P. H. (1980). Exact type 1 error rates for robustness of Student's t test with unequal variances. *Journal of Educational and Behavioral Statistics*, *5*(4), 337-349.

Rosenfeld, J. P., Labkovsky, E., Winograd, M., Lui, M. A., Vandenboom, C., & Chedid, E. (2008). The Complex Trial Protocol (CTP): a new, countermeasure-resistant, accurate, P300-based method for detection of concealed information. *Psychophysiology*, *45*(6), 906-919.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (6). Sage. Retrieved from http://books.google.com/books?hl=en&lr=&id=-4CVIVsBSiIC&oi=fnd&pg=PA3&dq=rosenthal+1991&ots=EG6vq0EMDm&sig=XqKJP1rR0LENZ5UwOMU7180rEGk

Ruchkin, D. S., Munson, R., & Sutton, S. (1982). P300 and slow wave in a message consisting of two events. *Psychophysiology*, *19*(6), 629-642.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, *17*(4), 688-690.

Simons, R. F., Graham, F. K., Miles, M. A., & Chen, X. (2001). On the relationship of P3a and the Novelty-P3. *Biological Psychology*, *56*, 207-218.

Spencer, K. M., Dien, J., & Donchin, E. (1999a). A componential analysis of the ERP elicited by novel

events using a dense electrode array. *Psychophysiology*, *36*, 409-414.

Spencer, K. M., Dien, J., & Donchin, E. (2001). Spatiotemporal Analysis of the Late ERP Responses to Deviant Stimuli. *Psychophysiology*, *38*(2), 343-358.

Spencer, K. M., Goldstein, A., & Donchin, E. (1999b). What is novel about “novel” stimuli? The effects of event probability on P300 and Novelty P3. *Psychophysiology*, *36*, S111.

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology*, *38*, 387-401.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science*, *150*, 1187-1188.

Talairach, J., & Tournoux, P. (1988). *A co-planar stereotaxic atlas of a human brain*. Stuttgart: Thieme.

Tucker, D. M. (1993). Spatial sampling of head electrical fields: the geodesic sensor net. *Electroencephalography and Clinical Neurophysiology*, *87*(3), 154-163.

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: applications to psychophysiology. *Psychophysiology*, *26*(2), 208-221.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*(3/4), 350-362.

Welch, B. L. (1947). The generalization of ‘Student’s’ problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28-35.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*(3/4), 330-336.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (279). John Wiley & Sons.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Springer Science & Business Media.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=zZ0snCw9aYMC&oi=fnd&pg=PP1&dq=wilcox+2012+robust&ots=PNtAVPbSvG&sig=eeBwsk8-w-NqqQG0al8FYdhhbxs

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