Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Biological Psychology 72 (2006) 333–343

BIOLOGICAL PSYCHOLOGY

www.elsevier.com/locate/biopsycho

Methodology

Interpreting event-related brain potential (ERP) distributions: Implications of baseline potentials and variability with application to amplitude normalization by vector scaling

Thomas P. Urbach [a,*], Marta Kutas [a,b]

[a] *Department of Cognitive Science, University of California, San Diego, United States*
[b] *Department of Neurosciences, University of California, San Diego, United States*

Received 9 November 2004; accepted 21 November 2005
Available online 30 January 2006

Abstract

Recent proposals regarding the purpose and validity of amplitude normalization by vector scaling including mitigation of baseline and noise problems in between-condition difference analyses are critically evaluated. In so doing, we elaborate on some of the points raised in Urbach and Kutas (2002) regarding baselines and noise, especially as these impact amplitude normalization by vector scaling and discuss the motivation for measuring event-related brain potential (ERP) amplitudes relative to a pre-stimulus baseline and the implications of this for certain (but not all) inferences. Throughout, our focus is on the logic of interpreting ERP measurements with an emphasis on the importance of specific assumptions and consideration of what conclusions are and are not supported.

Keywords: Event-related potentials; Voltage maps; Statistical analysis; Scaling methods

Statistically reliable between-condition differences in the distribution of scalp potentials show incontrovertibly that corresponding neural generators do differ *somehow* with respect to their location, polarity, or intensity. In Urbach and Kutas (2002), we critically evaluated an apparently widely held view that amplitude normalization, for example, by vector scaling, can sharpen this conclusion to allow the inference that the spatial configurations of the generators differ. We showed that although this proposition is indeed true for ideal distributions of generators and surface potentials under a suitable definition of ‘‘spatial configuration’’, the vector scaling procedure does little to narrow down the many possible ways that spatial configurations might differ. In particular, we demonstrated that different topographic shapes as reflected in a significant condition by electrode interaction after vector scaling do not support any inferences about changes in either the number or the location of the generators across conditions. Furthermore, computer simulations revealed that even the limited inferences that might be drawn in principle are not secure in experimental ERP practice because amplitude measurements relative to a baseline and noise can both lead to the misleading conclusion that post-stimulus spatial configurations of generators differ when they are actually the same.

Three main points emerged: (1) The distinction between distributions of scalp potentials and distributions of neural generators is not just terminological. Moreover, this distinction is critical to the purpose of vector scaling. Amplitude normalization aims to improve upon the general conclusion that the distributions of neural generators somehow differ by replacing it with a more specific conclusion about *how* they differ. Even if the normalization procedure were otherwise sound, it was not intended to be nor should it be treated as a post hoc test to ensure the reliability of distributional differences in surface potentials. It is not clear how widely this point is appreciated (see Dien and Santuzzi (2005) for an explicit endorsement of this misapplication of vector scaling). (2) Even for ideal distributions of generators and surface potentials, the extent to which vector scaling refines conclusions about generator distributions is limited to one special case. Prior to amplitude normalization, differences in scalp distributions show that neural generators differ in some combination of

DOI of related article: 10.1016/j.biopsycho.2005.12.002.

* Corresponding author at: Department of Cognitive Science, Mail Code 0515, University of California, San Diego, La Jolla, CA 92093-0515, United States.

E-mail address: turbach@cogsci.ucsd.edu (T.P. Urbach).

0301-0511/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.biopsycho.2005.11.012

location, polarity, and relative or overall strength. After amplitude normalization, residual differences merely attest to the fact that neural generators differ in some combination of location, polarity, or relative strength, that is, that they differ in spatial configuration. Of all the possible combinations of differences in generator locations, polarities, and strengths that could account for the different scalp distributions, amplitude normalization at best only rules out one special case: namely, where the generators in the two conditions all have the same locations and polarities and differ in strength by the same multiplicative factor. (3) Moreover, quite apart from these in principle limitations, simulations brought to light fundamental problems in applying amplitude normalization procedures to scalp potentials in standard ERP practice. Non-zero baseline potential distributions and noise are unavoidable and both pose problems for the interpretation of differences between amplitude-normalized distributions. First, distributions of scalp potentials recorded during the post-stimulus interval of interest are typically measured by subtracting some baseline potential distribution. Even if these baselines do not differ between conditions, nothing ensures that they are numerically zero at any recording site. When subtracted from the post-stimulus distributions of interest, non-zero baseline potentials can result in differences in the (apparent) topographic shape of the *measured* distributions even when the spatial configurations of the post-stimulus generators are the same. Because topographic shape alone cannot distinguish genuine differences in the spatial configuration of post-stimulus generators from the contribution of the baseline potential, the amplitude normalization procedure does not allow valid inference to different spatial configurations of post-stimulus generators. Indeed, non-zero baseline potentials pose a problem for identifying generator configurations for any procedure that operates on the algebraic difference of post-stimulus and baseline distributions. A second issue for amplitude normalization is noise. ERP measurements are never noise free. Setting aside technical artifacts, electrical interference, and non-cortical potentials, which might at least be mitigated, variability resulting from differences between individual subjects is unavoidable. Noise is a problem for vector scaling because it contributes to the amplitude of a distribution and tends to increase vector length. Noise-induced over-correction can result in residual differences in topographic shape after scaling, even when the spatial configurations of the generators are identical and the levels of noise are the same.

For these reasons, we recommended that the use of the vector scaling procedure for this purpose be discontinued. We have since received a number of queries about the consequences of our conclusions for ERP analyses. One in particular that has been asked by several people namely – whether the baseline problem can be avoided by scaling mean amplitudes of difference ERPs – is addressed by Wilding (2006). More generally, our conclusions have been questioned, challenged, and modified by Dien and Santuzzi (2005) and Wilding (2006), who offer alternative recommendations, albeit without any additional simulations or mathematical analyses.

The following discussion will elaborate some of the points raised in Urbach and Kutas (2002) regarding baseline distributions and noise, emphasizing aspects relevant to amplitude normalization by vector scaling. While there is presumably broad agreement that baseline potentials are an issue, there seems to be less consensus on what sort of an issue this is and how best to deal with it. Dien and Santuzzi (2005) suggest that baseline potentials are an unavoidable problem for all ERP research to which researchers must simply be resigned, whereas, Wilding (2006) proposes to circumvent the baseline problem for purposes of vector scaling by computing between-condition differences. In addressing these issues with regard to baseline, we examine the motivation for measuring amplitude relative to a baseline and the implications of this data transformation for subsequent inferences and interpretations. There is likewise broad agreement that noise is a problem in ERP research, although, again, there are markedly different approaches to dealing with it in the context of vector scaling. Dien and Santuzzi suggest the problems posed by variability-related vector length misestimation can be solved by visual inspection, although if topographic shapes could be determined by inspection alone, it is unclear why an analytic procedure like vector scaling is needed at all. As with the baseline problem, Wilding asserts that effects of noise can be mitigated by computing between-condition differences. We will show that this is not always so and that to know for sure requires rigorous mathematical analyses as well as computer simulations. Finally, we revisit the purpose of vector scaling. We will emphasize, contra Dien and Santuzzi, that vector scaling should not be used to evaluate the reliability of condition × electrode interactions in ANOVA and we will consider Wilding's proposal that differences in the spatial configuration of neural generators demonstrate qualitative differences in cognitive function.

1. Why measure ERP amplitude relative to a baseline?

Measuring ERP amplitudes as post-stimulus mean or peak amplitudes relative to a pre-stimulus baseline is ubiquitous in cognitive ERP research. The computation is straightforward—at each channel, the mean amplitude recorded in a pre-stimulus interval is subtracted from the recorded post-stimulus amplitude of interest. This measure is described as a matter of course in discussions of ERP methodology along with caveats, e.g., about the untoward consequences of residual noise in the baseline interval (Handy, 2005; Picton et al., 2000). The motivation for this transformation, however, receives less attention. Why measure post-stimulus amplitude against a pre-stimulus baseline at all? What question does this transformation answer that the analysis of the recorded data alone does not? Various answers might be imagined. For example, Dien and Santuzzi (2005, p. 71) write, “Because the baseline period is normally used to estimate the value of true zero in the EEG, activity in this period can result in misestimates of zero.” This assertion is puzzling. Subtracting the mean potential recorded in a baseline interval from each time point in the time series mathematically centers the data in the baseline interval around zero but this post hoc shift of the measurement scale is the same as placing the probes of a voltmeter across the terminals of a battery and turning the adjustment screw until the needle points to zero. Recentering the needle does nothing to determine the true potential of the battery and recentering the baseline interval does not determine the zero of the EEG. The motivation for measuring ERP amplitudes relative to a baseline potential must lie elsewhere and the alternative rehearsed next, without claim to novelty, is that measurements relative to a baseline are a key step in the interpretation of post-stimulus scalp potentials as causal consequences of an event. This is a general point about experimental design that does not depend on anything about brain activity or scalp potentials and can be illustrated with simple garden-variety experiments.

Suppose fertilizer is applied to the 10-week-old pepper plants in the backyard and 4 weeks later there are an average of 8.0 peppers per plant (p/p). The causal consequences of this fertilizing event, i.e., whether it increased, decreased, or had no effect on pepper yields can only be determined by comparison with a suitable control. If an unfertilized but otherwise identically treated plot yields an average of 4.0 p/p, it is natural to take the two together as evidence that the fertilizer did increase pepper yields (with the probability of error in rejecting the null hypothesis determined by statistical analysis of the variability about the mean yields, as usual). However, the pepper yields at 14 weeks may reflect the contribution of many factors and clearly the fertilizer cannot be responsible for yields that occurred before its application. Attributing all or even some of the 4.0 p/p difference between yields at 14 weeks to the fertilizer requires an additional assumption about what the pepper levels were prior to fertilizer application, i.e., the inference tacitly presupposes measurement of baseline pepper yields. The numerical values of such baseline measurements crucially constrain the inferences that can be drawn about the causal consequences of the fertilization event. To take an extreme case, suppose that at week 10, immediately prior to fertilization, there were already 8.0 and 4.0 p/p in the to-be-fertilized plot and control plot, respectively. In this case, the causal inferences drawn from yield measurements of 8.0 and 4.0 p/p at week 14 would be quite different: the assumption of identical treatments in the two plots would be highly suspect; whatever unknown factor was responsible for the different yields, it could not possibly be the not-yet-applied fertilizer; there would be no evidence whatsoever that fertilizer has any effect on the yields.

The baseline measurements of pepper yields at 10 weeks are not, *pace* Dien and Santuzzi (2005), used to determine true zero yields. Rather, baseline measurements provide a reference point against which to measure changes over time following an event of interest. In an appropriate experimental design with a suitable control condition, these changes, in turn, can be used to draw inferences about the causal consequences of the (eliciting) event. The motivation for selecting the time immediately prior to the event as the baseline is to separate out the causal consequences of that event of interest from other, previously occurring, factors. Changes over other time periods may be of interest as well. If the time course of the fertilizer's action is at issue, the pepper yields at 2 and 4 weeks after fertilization in comparison with a control condition might also be of interest.

2. Interpreting difference scores

A key fact about amplitude measurements relative to baseline is that these are difference scores in which two values, the post-stimulus and baseline amplitudes, are collapsed into a single value; in this sense they are no different than peak-to-peak measurements. It may be instructive to illustrate the consequences without reference to scalp potentials or brain activity. Satellite photos of a Caribbean hurricane represent the spatial distribution of water vapor (clouds) and using the image at 11:15 h as the baseline spatial distribution, the change in clouds over time can be measured by subtracting the grayscale value of the baseline image pixel by pixel from the later photos H1 and H2 (Fig. 1A). Note that in the difference image, color now indicates the change in moisture level between the two times rather than moisture level at a given time. Lighter colors represent a greater relative increase in moisture, darker colors represent a greater relative decrease, and medium gray represents no change, i.e., cloudy stays cloudy or clear stays clear. These satellite photo difference maps clearly show the algebraic blending that results when spatial distributions at two different times are collapsed into one by measurement relative to a baseline (or for that matter any other interval).

The same points hold for ERPs and are illustrated here with waveforms from one condition in a simple reaction time experiment before and after subtracting the mean pre-stimulus baseline for each channel (Fig. 1B). The contour maps represent three distributions of recorded potentials across the scalp: mean amplitude in the 200 ms pre-stimulus baseline interval, peak negative amplitude between 50 and 120 ms (N1), and mean amplitude 120–200 ms post-stimulus (P2) (Fig. 1B, bottom row, left). The distribution of the corresponding N1 baseline-to-peak and P2 baseline-to-mean amplitude measurements are also represented with contour maps (Fig. 1B, bottom row, right). These contour maps show that when the baseline and post-stimulus distributions are similar as they are for N1 peak amplitude, the difference distribution tends to be more equipotential (Fig. 1B, compare N1 and N1-baseline). If the baseline and post-stimulus distributions are dissimilar as they are for P2 mean amplitudes, the range of values in the resulting difference distribution can be greater, resulting in more pronounced peaks and valleys in the contour map (Fig. 1B, compare P2 and P2-baseline).

The difference maps of the satellite photos illustrate the challenges inherent in interpreting this sort of algebraically blended spatial distribution. In the H2-baseline image (Fig. 1, top right) there are various roughly circular focal bright spots. In each case, these are incontrovertibly localized areas in which there are more clouds in the H2 image than in the baseline image. However, the explanation of why each is focal is a different matter. One of the bright spots arises because a focal dark spot in the baseline image (the eye of the hurricane, ○) coincides with part of a more extensive light (cloudy) area in H2. Another focal bright spot arises in the difference because a

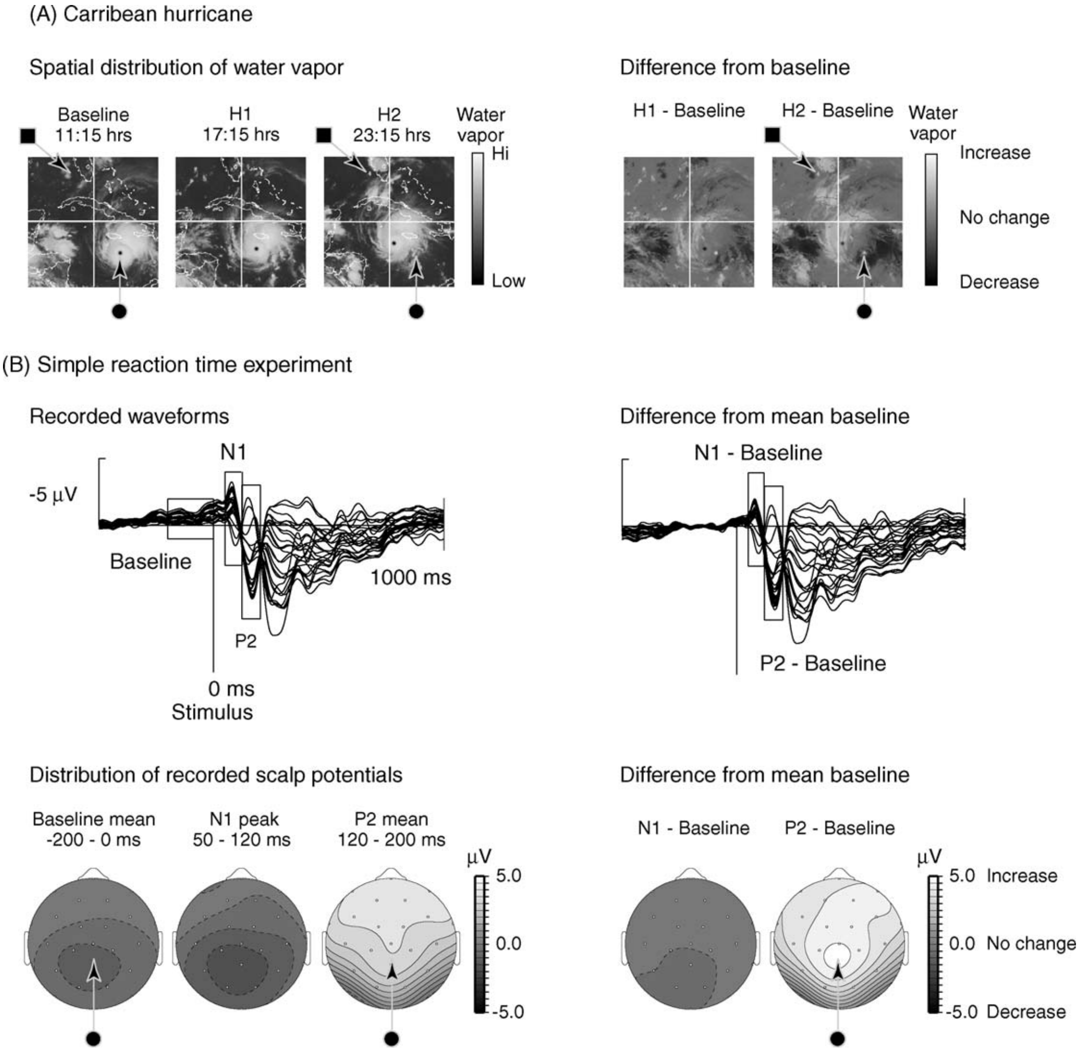

Fig. 1. (A) A sequence of photos taken from a geosynchronous satellite at 6 h intervals. The color at each pixel is given by one of 255 grayscale values as a red, green, blue (RGB) triples of 16-bit values ranging from black = [0, 0, 0] to white [65535, 65535, 65535]. The [r, g, b] difference values ranged from [−65535, −65535, −65535] for a black pixel to [65535, 65535, 65535] for a white pixel minus a black pixel and were remapped back to the original grayscale range as the integer portion of ([r, g, b] + 65535)/2. For purposes of illustration, the eye of the hurricane in each image was digitally enlarged prior to the computation of the difference images. The original Geostationary Operational Environmental Satellite (GOES) images are in the public domain and provided by the National Environmental Satellite, Data, and Information Service (NESDIS) of the National Oceanic and Atmospheric Administration (NOAA). (B) Grand average ERPs to visual targets in a simple reaction time experiment before and after subtracting mean amplitude in a 200 ms pre-stimulus baseline interval. The contour maps represent the scalp distribution of the potentials in the corresponding intervals: mean potential 200 ms (baseline), peak negative potential 50–120 ms post-stimulus (N1), and mean potential 120–200 ms post-stimulus (P2). See Fig. 2 caption for further description of the experimental paradigm.

focal bright spot (concentration of thunderheads, □) occurs in H2 in locations that are dark (clear) in the baseline. Inspection of the difference map alone cannot disentangle these two equally likely and plausible explanations. Since the waveforms and scalp maps of amplitudes measured relative to a baseline are this same sort of blend of activity from two time slices, they are liable to the same sort of interpretive ambiguity. For instance, the centro-parietal positivity in the P2-baseline contour map is uncontroversially a fairly focal area where the potential is relatively greater – more positive or less negative – during the P2 interval than in the baseline interval. However, it cannot be determined from the difference map alone whether this increase in potential from one time to the other is because of a focal centro-parietal negativity in the baseline interval (c.f., the hurricane eye) or a focal positivity in the P2 interval (c.f., the thunderheads). A further inference from focal potentials in a difference map to focal brain activity is more tenuous still and at a minimum requires additional assumptions.

For satellite images and ERPs alike, magnitudes derived by subtracting pre-stimulus baseline magnitudes are difference scores wherein information has been irretrievably lost. The subtraction transformation maps many pairs of arguments to one value, e.g., 7 − 5 = 2 and 6 − 4 = 2, and there is no inverse function from the difference back to the pair of arguments whence it came. As a result, analysis of differences relative to baseline is not sufficient to test some sorts of hypotheses. For instance, the H2-baseline difference photo is not informative with regard to hypotheses about the quantity of clouds at a particular location at 17:15 h (H2). If the difference image is medium gray at this location, all that follows is there was no change between the two times. There may have been clouds at both times or no clouds at both times. Or, if the difference image is dark gray at this location, it might be because dense clouds had decreased by H2, but clouds still remained or because moderate clouds were entirely gone by H2.

The same considerations hold for ERPs, and a moderate increase in positivity relative to baseline might be the result of a potential becoming less negative relative to a negative baseline or more positive relative to a positive baseline. The hypothesis that the potential in the post-stimulus interval was positive cannot be inferred from a positive change measured relative to baseline since it is equally consistent with a different alternative. These observations have implications for drawing inferences about post-stimulus brain activity. For instance, if a model of neural generators is computed for post-stimulus amplitudes measured relative to a pre-stimulus baseline, it is a model of the algebraic blend of pre- and post-stimulus generators (see Urbach and Kutas (2002), Fig. 4C). If such generator blends from two different times are of experimental interest as measures of change from baseline (c.f., the representation of moisture change from baseline in the satellite difference images), well and good. If the question at hand concerns which brain regions are or are not active as stimuli are being processed, conclusions cannot be drawn directly from the usual difference distributions alone. Between-condition difference waves (subtractions) have been much discussed in ERP research and there are well known constraints on their interpretation (e.g., Picton et al., 2000, p. 140): "When using difference waveforms, authors should bear in mind various factors that might affect the subtraction by differentially affecting the two recordings from which the difference is calculated. Cognitive factors include changes in the state of the subject and changes in the manner of processing information between the two recordings. More physiological factors include changes in latency of one or more components in the unsubtracted ERPs." These points are well taken and apply with equal force to ERP amplitude measured relative to a baseline in a single experimental condition since this too is a difference score.

Although difference maps may be ambiguous in some respects, they are still systematically related to the phenomena of interest and informative in important ways. For instance, a hypothesis in the hurricane example might be that clouds are building up at a particular location (a monotonic increase in clouds over time). This hypothesis entails that there is more cloud activity at H1 (17:15 h) than at the baseline (11:15 h) and yet more at H2 (23:15 h) than at H1. Two testable predictions are that in difference maps (H1-baseline) and (H2-baseline) the shade of gray at that location will be lighter than the medium gray corresponding to no change in clouds. In addition, the hypothesis predicts that the shade of gray at the relevant location will be lighter in the H2-baseline image than in H1-baseline image. There may be better ways to test hypotheses about hurricanes. The point for present purposes is that even though the difference maps are intrinsically ambiguous algebraic blends of activity from two different time points, this does not mean they are uninterpretable, as long as the questions asked are the sort that difference data can answer. In addition, considerations that go beyond the difference maps can be brought to bear to constrain the interpretation. For the satellite photos, auxiliary assumptions about the plausible sizes and speeds of hurricanes, their typical trajectories, and the number of eyes could make one focal bright spot a better candidate for an "inverted eye" interpretation than another. In the maps of scalp potentials, auxiliary assumptions can also constrain interpretations and support some explanations while militating against others. For instance, the bilateral negativity over occipital areas in the P2-baseline map could, mathematically, be the result of two asymmetric distributions that combine to give a symmetric bilateral effect in their difference. However, given what is known about the pre-target interval, central presentation of the stimuli, the architecture of the nervous system and the time course of visually evoked neural activity, attributing this effect to bilateral processing is a better hypothesis.

For both the photos and the scalp potentials, the most straightforward way to disentangle the blended difference is to look at the two distributions separately. Doing so may clarify which features – hurricane eyes versus thunder heads or symmetric versus inverted asymmetric distributions of potentials – are contributing to the difference. Subtracting the baseline interval from each point in a time series mathematically centers the waveform at each channel. Mathematically squashing the baseline distribution flat in this manner does not make it go away, it merely folds it in with the potentials recorded during the post-stimulus interval of interest (Fig. 1B, difference waveforms). The actual distribution of baseline potentials might be problematic or entirely benign for the relevant analyses and interpretations, but there is no way to know unless they are examined and presented separately. In this connection it is of some interest to note that ERP publication guidelines (Picton et al., 2000, H(vi)) dictate that if subtractions (difference waves) are presented, the original ERPs must be presented as well. This guideline was intended to apply to between-condition subtractions but it would confer the same benefits to the most common type of subtraction: post-stimulus amplitude measurements relative to a pre-stimulus baseline (Fig. 1B).

3. Implications of baseline distributions

Our comments on baselines thus far notwithstanding, we cannot overemphasize that we are not in total agreement with Dien and Santuzzi (2005, p. 72) when they write: "baseline effects are an issue for all ERP analyses. Taken to the extreme, one would have to abandon all ERP analyses by this logic." Rather, our position is that although baseline effects are indeed an issue for some ERP analyses, they are not for others. The hypothetical event-related pepper experiments share important features with event-related potential experiments and thus can illustrate why non-zero baseline values do not necessitate abandoning agricultural science or ERP research.

In a process that evolves over time, the dependent variable immediately prior to the event of experimental interest may or may not be numerically zero and may or may not differ between conditions. Even in two identically treated plots, it would be surprising to find identical pre-fertilizer baseline yields; small unsystematic between-condition differences in the baseline quantities are far more likely. Large systematic differences in baseline levels might also arise from the influence of an uncontrolled factor in a poor experimental design or from the influence of an experimentally manipulated independent variable in a good experimental design. For instance, to determine how fertilizer interacts with the amount of sun, it would be reasonable to split a sunny plot and a shady plot and treat half of each with fertilizer. Since peppers do well in full sun, baseline pepper yields at 10 weeks would be expected to be different (and higher) in the sunny plot than in the shaded one. To test hypotheses about the interaction between these factors, there is no logical requirement that the baseline levels be zero or even the same in the two conditions. The point of measuring relative to the baseline is to track event-related changes over time that occur above and beyond baseline levels that may or may not be the same.

Electrical brain activity is also a process that evolves over time, and the potentials prior to the event of experimental interest may or may not be numerically zero and may or may not differ between-conditions. The baseline subtraction transformation is illustrated for ERP waveforms and contour plots in three experimental paradigms: recognition memory, rapid serial visual presentation (RSVP) sentence comprehension, and reaction time (Fig. 2B). In all three paradigms there are systematic distributions of stimulus-locked potentials in the pre-stimulus baseline interval itself (Fig. 2, middle column). Urbach and Kutas (2002) argued that the topographic shape of measured distributions as determined by vector scaling varies with different non-zero baseline distributions and these typical ERP data suggest that the problem is not just hypothetical handwringing. The non-zero pre-stimulus baseline potentials are systematically distributed across the scalp with magnitudes of a few microvolts, comparable to many experimental ERP effects.

Systematic pre-stimulus neuroelectric activity need not mean that the experiments were poorly designed or that the potentials were recorded inappropriately. In the recognition memory and reaction time experiments, the fixation stimulus is a warning cue that a target is impending and even though the interstimulus interval was varied, it is not surprising to find systematic preparatory potentials. In the sentence comprehension experiment, the pre-stimulus baseline interval occurs immediately after the main verb of a sentence and again, it is not unreasonable to find a systematic pattern of potentials at the scalp associated with the cognitive processing at this point in a sentence. As with the peppers, some between-condition baseline differences might be expected as the result of the experimental manipulation even in a sound design. While not all ERP experiments necessarily result in non-zero baseline distributions, some do (Fig. 2) and others probably do as well. In typical ERP experiments conscious alert human subjects are maintaining instructions and response mappings in memory, processing previous stimuli, directing attention in preparation for processing upcoming stimuli, and in some cases, preparing to make a decision and map it to an overt response. The experimental design may lead one to expect one thing or another but at the end of the day whether or not the baseline distributions are or are not flat, non-zero, or the same between conditions is an empirical question.

For both event-related pepper yields and event-related potentials, difference scores derived by measuring post-event quantities against pre-event baseline quantities in controlled experiments permit some kinds of conclusions to be drawn but not others. Post-event changes relative to a pre-event baseline can be compared with changes from baseline in other conditions – control or experimental – and used to draw inferences about the causal consequences of the stimulus event. For some causal inferences the numerical value of the pre-event baseline potential can be ignored without prejudicing the conclusions because the relevant quantity is change-from-baseline regardless of what the baseline value is. A pre-to-post event change of 4 p/p (or μV) in an experimental condition relative to a change of 2 p/p (or μV) in a control condition suffices to establish both a causal effect of the manipulation (fertilizer affects yield) and a direction of the effect (fertilizer increases yield). For these inferences, the conclusions follow regardless of what the baseline values are. Other inferences however are not invariant with respect to the baseline values. For instance, this ratio of fertilizer event-related changes from baseline, i.e., 4:2, does not necessarily show that the fertilizer doubles the final yield. This conclusion would be reasonable if the pre-fertilizer baseline yields were zero in both plots, but not if they were already 10. That is, the event-related increase from 10 to 14 in the fertilized plot is still twice the increase from 10 to 12 in the unfertilized plot, but the yield does not double. The key point here is that some valid inferences can be drawn solely from amplitude changes from baseline, whether or not these baseline distributions are zero or the same across the scalp or the same in both conditions, and some inferences cannot.

Wilding (2006) proposes a procedure for vector scaling between-condition differences that makes a useful case study for illustrating how different inferences require different assumptions about baseline potentials. Wilding purports to have identified a class of experiments for which vector scaling is useful, namely those in which comparisons among pair-wise between-condition differences are of interest. His recommendation is based on the observation that distributions of between-condition differences (experimental effects) are invariant when measured against different baseline distributions provided the baseline distributions in the two conditions are the same. This observation is correct, and Wilding's justification based on the inspection of a figure can be supplemented by mathematical considerations. Suppose that P_1 and P_2 are vectors of post-stimulus scalp potentials (i.e., scalp distributions) recorded in two experimental conditions and that B_1 and B_2 are the corresponding distributions of pre-stimulus baseline potentials. The measured baseline-to-mean amplitudes in the two conditions are, $M_1 = P_1 - B_1$

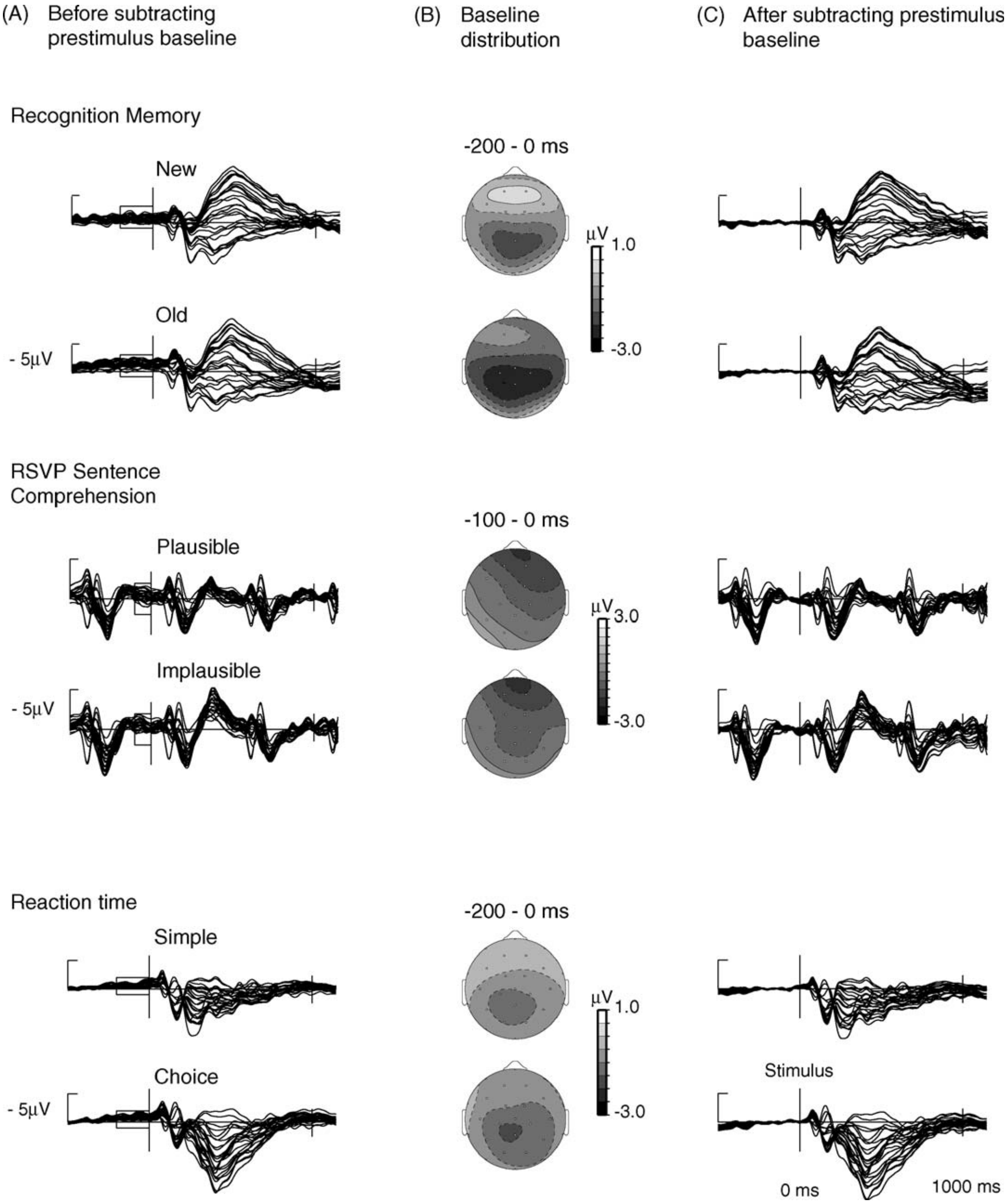

Fig. 2. Grand average potentials across subjects from three ERP experiments in which visual stimuli were presented on a computer monitor and 26 channels of EEG data were recorded from electrodes evenly distributed across the scalp and rereferenced offline to the mathematical average of the left (A1) and right (A2) mastoids. Two conditions in each paradigm are illustrated with the grand average waveforms across subjects from all 26 channels superimposed in each plot of potential (μV) over time (ms). (A) The recorded waveforms before subtracting mean amplitude in the pre-stimulus interval. (B) Spline interpolated contour maps of the mean potentials recorded in the pre-stimulus baseline interval (0.5 μV per contour on all scales). (C) Waveforms that would be plotted separately for each channel in a typical ERP research report after subtracting the mean baseline potential. In the recognition memory experiment, target stimuli are fragments of photographic images presented for 2000 ms. Targets were preceded by a fixation frame with the frame-to-target intertrial interval varying between 1500 and 2500 ms. The task was to verbally identify what the entire image depicts based on what can be discerned from the fragment. The waveforms are from the test phase in which half of the image fragments were presented during a study phase (old) and the other half were not (new). The images were counter balanced during study so that each test phase image appeared as old for half the subjects and as New for half the subjects. In the sentence comprehension experiment, sentences of the form subject verb object were presented word by word at an SOA of 500 ms. The target word is a plausible or implausible object noun and the same subject verb contexts appeared in both Object noun conditions. In the reaction time experiment with targets were letter strings consisting of five L's or five R's. The simple task was a speeded button press in blocked presentation of the targets, left button for L targets and right button for R targets. The choice task was the same except that the R and L targets were presented randomly intermixed. Target stimuli were preceded by a centrally presented fixation cross and the intertrial interval varied between 850 and 2050 ms.

and $M_2 = P_2 - B_2$ and the experimental effect is given by $M_2 - M_1 = (P_2 - B_2) - (P_1 - B_1) = (P_2 - P_1) - (B_2 - B_1)$. Under Wilding's assumption that the baseline distributions are the same, i.e., that $B_2 = B_1$, the two baseline terms cancel and the experimental effect is simply $M_2 - M_1 = (P_2 = P_1)$. The assumption of identical baseline distributions is not only sufficient for the invariance of the post-stimulus effect as this analysis shows, it is also necessary. If the baseline distributions differ, the term $(B_2 - B_1)$ does not drop out and the experimental effect $M_2 - M_1$ is no longer invariant with respect to the baseline distributions. In this case, the difference between the baselines is folded into the difference between the post-stimulus distributions and the result is a between-condition blend of pre- and post-stimulus distributions, with all that this entails.

The crucial assumption that the baseline distributions are identical is quite restrictive and whether or not it is met is an empirical question that must be decided in each case before any additional analysis (see Fig. 2B). Although some sorts of experiments may be designed to minimize between-condition differences in the pre-stimulus interval, there is no class of designs where the design alone guarantees that the pre-stimulus baseline distributions are in fact the same. Moreover, making such an empirical determination is sticky because of the familiar problem that failing to find a difference does not suffice to show that a difference does not exist. In practice, no two distributions are likely to be identical and the question will become how close is close enough, i.e., how many and what sizes and sorts of differences can be tolerated. Working this out systematically, e.g., for larger numbers of small differences and smaller numbers of large differences in a way that generalizes from 3 to 256 electrodes, is no small task. Although these open questions need to be addressed before Wildling's procedure can be used in practice, the proposal serves a useful purpose by illustrating some important points. First, the inferences that would be secured by Wilding's procedure are entirely dependent on the values of the pre-stimulus baseline potentials, and this provides an instructive contrast with the causal inferences illustrated above that go through regardless of whether the baseline values were zero, non-zero, or the same or different in the two conditions. Second, the proposal highlights the difference between specifying conditions that must be satisfied in principle and providing an explicit method for determining whether or not the conditions are in fact satisfied in any given experiment. The latter is what distinguishes a proposal for an analytic procedure from an analytic procedure.

4. Variability and vector length

Variability in potentials recorded within and across individual subjects is another fact of life in ERP research that poses challenges of its own. The computer simulations of multichannel ERP data sets in Urbach and Kutas (2002) showed that uncorrelated variability in the measurements at the simulated electrode channels led to systematic overestimations in the lengths of the vector representations. As variability increased relative to a given distribution, the estimates became progressively worse, disproportionately affecting distributions with the smaller overall amplitudes.

Various approaches to the problem of variability-related vector length misestimation in application to ERPs might be imagined. One is proposed by Dien and Santuzzi (2005, p.71) who acknowledge the problem but continue to endorse vector scaling and recommend that condition × electrode interactions that remain after vector scaling be graphed and inspected: "If the two graphs have the same relative distributions across the electrodes and differ only in the relative amplitude, then one can conclude that the significant interaction is instead due to insufficient correction by the vector scaling." The idea behind this suggestion is not so much wrong as incomplete. If two distributions of scalp potentials have the same topographic shape, i.e., are related by a scalar factor, the distributions obtained by scaling each separately by any factor whatsoever will also have the same topographic shape even if differences in amplitude remain (c.f., two similar triangles of different sizes, each scaled by a different factor remain similar). However, Dien and Santuzzi appear to walk the ERP researcher in a very narrow circle right back to the question vector scaling set out to answer: are the differences observed in the (now vector scaled) distributions evidence of different topographic shapes or evidence of (merely) different overall amplitudes? If this question could be reliably decided by inspection of graphed cell means, one could simply graph the unscaled potentials and forego amplitude normalization entirely. Inspection of cell means is a subjective procedure in the first place and does nothing to take account of the size of remaining effects relative to the variability in the second. Conclusions drawn from the mere inspection of unscaled ERP cell means in lieu of inferential statistics are of heuristic value only and without further justification it is unclear why inspection of vector scaled magnitudes should be treated any differently.

Another approach to variability related vector length misestimation may be found in Wilding (2006) where it is first asserted that variability related vector misestimation is less problematic for the recommended across-subject vector scaling procedure than within-subject scaling. Even if this is true, the "less problematic" across-subject vector scaling procedure still inflates Type I error rates (see Urbach and Kutas (2002), Fig. 6C), and Wilding offers no justification for the leap from "less problematic" to "unproblematic". Wilding's sanguine attitude toward variability in connection with his proposal to vector scale between-condition difference scores appears to be equally unfounded: "when the data submitted to analysis comprise difference scores, these are the conditions under which changes in noise levels are the least influential (see the 'zero baseline' condition in Fig. 10C of Urbach and Kutas (2002))." This reference is puzzling because there is no analysis of difference scores in Urbach and Kutas and it is unclear how Fig. 10C offers any guidance in this regard. Furthermore, considered on its own merits the assertion seems problematic. For any two sets of scores $X = \{X_i | 1 \leq i \leq j\}$ and $Y = \{Y_i | 1 \leq i \leq j\}$, the variance σ^2_{X-Y} of the set of pair-wise difference scores, $X - Y = \{X_i - Y_i | 1 \leq i \leq j\}$ is given by $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\sigma^2_{XY}$, i.e., the sum of the variances of X and Y minus twice the covariance. In the usual case, σ^2_X and σ^2_Y are greater than zero. If X and Y are uncorrelated, the covariance term is zero and the variance of the difference scores is the sum of the variances and greater than either variance considered separately. In a typical within-subjects ERP comparison between two conditions, scores may well covary, in which case the covariance term $2\sigma^2_{XY}$ will not be zero. If so, the variance of the difference scores will be less than the sum of the variances of X and Y, how much less depends on the degree of correlation. Contrary to Wilding's assertion, nothing about the mathematical relation between difference scores and variance systematically ensures that computing difference scores mitigates variability-related vector length misestimation. Computing difference scores may even exacerbate the problem in some cases. At best the matter is undecided and without a compelling argument it would be premature to casually dismiss the problem.

Variability related vector length misestimation is not a trivial problem, but it may well be tractable. By way of comparison, consider violations of the sphericity assumption in repeated-measures ANOVA. Although inhomogeneity of variance inflates the Type I error rate when critical *F* is based on the nominal degrees of freedom, the error rate can be controlled by adjusting the degrees of freedom as a function of the degree of inhomogeneity (Greenhouse and Geisser, 1959; Huynh and Feldt, 1976). These adjustments derive from mathematical analyses of the consequences of sphericity violations, and so too, a rigorous mathematical analysis of the relation between variability and vector length might likewise afford a systematic, objective, and generally applicable procedure for controlling error rates. Urbach and Kutas (2002) argued at length that even if amplitude normalization by vector scaling were valid and reliable, the questions it could actually answer about the spatial configurations of neural generators are of limited interest. Before going to the trouble of addressing the problem of variability related vector length misestimation – which would clearly involve lots of work – it might be worth reconsidering what would be gained by doing so.

5. Why normalize amplitude by vector scaling?

In the final analysis, the justification for any analytic procedure is pragmatic and depends on the extent to which the procedure reliably contributes to answering any scientifically important questions. At issue here is what scientifically important questions does a reliable procedure for comparing topographic shapes answer that cannot be answered without it. The widely, though not universally, received view critically evaluated in Urbach and Kutas (2002) was that the vector scaling transformation afforded comparison of the topographic shapes of distributions of scalp potentials which, in turn, afforded conclusions about the spatial configurations of the neural generators of the corresponding scalp potentials (McCarthy and Wood, 1985; Haig et al., 1997; Ruchkin et al., 1999; Picton et al., 2000). The first section of the critical discussion in Urbach and Kutas elaborated a point made in Alain et al. (1999) and noted also in Picton et al. (2000) that differences between the spatial configurations of generators could result from differences in neural generator locations or differences in the relative strengths of generators in fixed locations. This is an observation about what sort of specifics about the spatial configurations of neural generators can and cannot be inferred from the analysis of the topographic shape of the corresponding scalp potentials. Urbach and Kutas further argued that under suitable idealizations, if the neural generators in two experimental conditions are identical with respect to their number, location, and polarity and the only difference is that every generator in one condition differs in intensity (strength) by the same factor from the corresponding generator in the other, then the distributions of potentials at the scalp would have the same topographic shape. From this it follows that if some procedure can demonstrate that topographic shapes are not the same, then this and only this configuration of generators could be ruled out (see Urbach and Kutas (2002) for details). Neither vector scaling nor any other procedure that simply compares topographic shapes of amplitudes can answer the arguably more interesting questions about whether the number or locations of neural generators differ between experimental conditions. The next, and perhaps more controversial, issue is whether ruling out this particular relation between generator configurations is of much scientific interest. Urbach and Kutas concluded, in essence, that the questions that comparisons of topographic shape can answer are not interesting and those that are interesting it cannot answer.

A strategy for responding to this would be to demonstrate that there is something interesting about the questions vector scaling can answer after all. Dien and Santuzzi (2005, p. 71), for example, appear to take the point of vector scaling to be rather different than that outlined above: ‘‘vector scaling is intended only to provide a test of the reliability of the [condition × electrode] interaction’’. This seems to express the view that conducting an ANOVA on vector scaled distributions is a kind of post hoc test conducted as a follow-up to confirm a reliable interaction in the unscaled potentials. Although there are cases in which a data transformation is conducted to improve on a statistical analysis, e.g., the log transformation of reaction times to make a skewed distribution more nearly normal, this is not one of them.

Wilding (2006) makes a different explicit proposal about what interesting question vector scaling can answer by asserting that differences between the topographic shapes of between-condition differences demonstrate that functionally distinct processes are engaged: ‘‘to make the functional inference that qualitatively different cognitive processes are engaged, the fact that the spatial configurations – as defined by Urbach and Kutas – are different is sufficient’’. Although the project of relating brain activity and cognitive function would be much easier if this were so, this mapping principle seems rather dubious as formulated. No reasons are offered in support of this principle and the citation of Rugg and Coles (1995) in this connection is puzzling, since that chapter appears to make a reasoned argument to exactly the opposite conclusion, i.e., that the mapping between differences in surface potentials and differences in cognitive function is not at all simple. There is neither explicit argumentation nor textual support for this mapping principle and, furthermore, general considerations appear to militate against it.

If differences in the distribution of scaled potentials did license the conclusion that (A) spatial configurations of generators differed and this in turn licensed the conclusion that (B) qualitatively distinct cognitive processes are engaged, then there might indeed be an in-principle motivation for scaling. Even setting aside the practical difficulties in establishing (A), there do not appear to be any grounds for drawing the conclusion (B). Perhaps if differences in spatial configuration entailed that there were generators in different locations, i.e., that neural tissue was active in different places, then a presumption that qualitatively different cognitive processes were involved might be a more plausible prima facie hypothesis. However, one of the central points in Urbach and Kutas (2002) is that differences in vector scaled distributions do *not* entail that generators are in different locations. Between-condition differences in the intensity (strength) of some generators and not others without any other differences in location or polarity, yield differences in spatial configuration of the sort that can (ideally) be identified by the comparison of topographic shape after vector scaling (see Urbach and Kutas (2002), Fig. 1C). If Wilding's mapping principle were correct, experimentally ratcheting the intensity of a generator (or multiple generators) up or down while leaving all other generators unchanged would be enough to establish a qualitative difference in cognitive processing. A parsimonious alternative interpretation is that the change in intensity reflects a quantitative change in the same cognitive function(s). This alternative is at least as plausible as Wilding's proposal. Indeed, to the extent that there is consensus on the matter, it would seem to fit better with a prevailing view that amplitude changes alone (which this is) reflect quantitative and not qualitative differences in cognitive function. Since there is an equally plausible alternative cognitive interpretation of the difference in the spatial configuration of generators, such a finding alone does not even provide evidence, let alone suffice, for the conclusion that cognitive functions differ qualitatively.

For ease of exposition this point has been illustrated with a between-condition comparison but the same considerations apply to comparisons of pair-wise between-condition differences. Suppose the post-stimulus generators in condition 1 and condition 2 differ only in overall strength and that between-condition differences are computed with a common subtractor as Wilding recommends, i.e., Diff_{1-3} = (condition 1 − condition 3) and Diff_{2-3} = (condition 2 − condition 3). Setting aside concerns about the baseline presented above, suppose further that the baseline distributions in all three conditions are identical and mathematically cancel out (see Wilding (2006), Fig. 2, bottom row for an illustration of this best case situation). If the distribution of surface potentials in condition 3 is non-zero, the pair-wise condition differences Diff_{1-3} and Diff_{2-3} are once again algebraic blends. Since the generators in conditions 1 and 2 differ in strength (alone) and the generators in condition 3 are the same in both, by definition the spatial configuration of generators in Diff_{1-3} and Diff_{2-3} will be different. Urbach and Kutas (2002, Fig. 4) illustrated how subtracting identical (non-zero) baseline distributions from two post-stimulus distributions that differ only in strength results in different spatial configurations of generators. The situation is exactly the same when a (non-zero) third condition is subtracted from two post-stimulus distributions that differ only in strength. The upshot is that even if (identical) baseline potentials are cancelled out by computing pair-wise between-condition differences, the problem just rehearsed for simple between-condition comparisons remains. That is, conditions 1 and 3 differ only in the strength of their generators but the spatial configurations of generators are different in the pair-wise condition differences Diff_{1-3} and Diff_{2-3}. Wilding's mapping principle would entail that this difference in strength alone between conditions 1 and 2 suffices to establish qualitative differences in cognitive function. Again, this inference is a huge leap at face value and logically undermined by the existence of alternative interpretations of these differences in generator strength.

Although questions of qualitative differences in cognitive processes are surely of great interest, it is not clear that vector scaling can answer them as directly as Wilding supposes. If the foregoing is right, conclusions about the spatial distribution of generators that can be drawn from vector scaling – individual conditions and pair-wise condition differences equally – are not sufficient for drawing inferences to strong claims about the existence of qualitative differences in cognitive functions. This does not establish that comparison of topographic shapes by vector scaling cannot answer any scientifically interesting questions since the space of such questions is large. A compelling case for vector scaling has yet to be made, however.

6. Conclusion

The discussion here attempted to apply some commonsense methodological principles to particular issues in ERP research. Throughout, our focus is on inference and interpretation, in particular the importance of determining exactly what experimental question is being asked, what conclusions can be drawn from a particular analysis, and the extent to which the conclusions do or do not answer the question asked. Along the way, the shortcomings of relying on handwaving and unsupported assertions about the consequences of an analytic procedure instead of a rigorous analysis via mathematical argument sometimes supplemented with computer simulations were illustrated.

We argued that amplitude measurements relative to non-zero pre-stimulus baseline potentials were problematic for some sorts of inferences, but not for others. Inferences that rely on assumptions about baseline potential distributions, such as those that proceed from vector scaled distributions to conclusions about neural generators, face the practical problem of how to determine that the relevant assumptions are satisfied, e.g., that baseline distributions are zero or flat or the same between conditions. We recommended routine mapping of baseline potential distributions. We also evaluated approaches to the problem that noise poses for drawing inferences from vector scaled distributions and speculated that this might be tractable. The discussion returned to what questions vector scaling might hope to answer and critically evaluated two recent proposals in this regard. The importance of clearly articulating what scientifically interesting questions might be answered by a rehabilitated vector scaling procedure (or any other) was emphasized.

Having rehearsed interpretive limitations at some length, it is important to emphasize again that conclusions drawn from the analysis of ERP amplitudes measured relative to a pre-event baseline in controlled experiments are not intrinsically more problematic than conclusions drawn about water vapor from the analysis of satellite photos or about the effects of sun and fertilizer on pepper yields. Well-designed ERP experiments allow sound causal inferences to conclusions about the relation between experimentally manipulated independent variables and brain activity and these conclusions may in turn play a role in a much longer story about the relation between cognitive processes and brain function. If, in navigating the Scylla of unreliable or untested analytic procedures and the Charybdis of abandoning ERP research, we do not go as far or as fast as might be hoped, there is some consolation in keeping the boat off the rocks and thus moving forward safely.

Acknowledgements

This research was supported in part by grants HD22614 and AG088313 to Marta Kutas. M. Kutas is a Lady Davis Fellow at Hebrew University.

References

Alain, C., Achim, A., Woods, D.L., 1999. Separate memory-related processing for auditory frequencies and patterns. Psychophysiology 36, 737–744.

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

Greenhouse, S.W., Geisser, S., 1959. On methods in the analysis of profile data. Psychometrika 55, 431–433.

Haig, A.R., Gordon, E., Hook, S., 1997. To scale or not to scale: McCarthy and Wood revisited. Electroencephalography and Clinical Neurophysiology 103, 323–325.

Handy, T.C., 2005. Basic principles of ERP quantification. In: Handy, T.C. (Ed.), Event-Related Potentials: A Methods Handbook. MIT Press, Cambridge, MA, pp. 33–56.

Huynh, H., Feldt, L.S., 1976. Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational Statistics 1, 69–82.

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

Picton, T.W., Bentin, S., Berg, P., Donchin, E., Hillyard, S.A., Johnson Jr., R., Miller, G.A., Ritter, W., Ruchkin, D.S., Rugg, M.D., Taylor, M.J., 2000. Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. Psychophysiology 37, 127–152.

Rugg, M.D., Coles, M.G.H., 1995. The ERP and cognitive psychology: Conceptual issues. In: Rugg, M.D., Coles, M.G.H. (Eds.), Electrophysiology of Mind: Event-related brain potentials and cognition. Oxford University Press, Oxford, pp. 27–39.

Ruchkin, D.S., Johnson Jr., R., Friedman, D., 1999. Scaling is necessary when making comparisons between shapes of event-related potential topographies. Psychophysiology 36, 832–834.

Urbach, T., Kutas, M., 2002. The intractability of scaling scalp distributions to infer neuroelectric sources. Psychophysiology 39, 791–808.

Wilding, E.L., 2006. Technical note: on the practice of rescaling scalp-recorded event-related potentials. Biol. Psychol. 72, 325–332.