Rigor and replication in time-frequency analyses of cognitive electrophysiology data

Michael X Cohen
University Medical Center, Science Faculty, and Donders Center for Neuroscience, Radboud University
mikexcohen@gmail.com

Keywords: EEG, time-frequency, oscillations, replication, methods, electrophysiology

Running head: Replications in cognitive electrophysiology

Acknowledgments: MXC is funded by an ERC-StG 638589

Abstract

Cognitive electrophysiology is a subfield of neuroscience that focused on linking M/EEG data to aspects of cognition and the neurophysiological processes that produce them. This field is growing in terms of the novelty and sophistication of findings, data, and data analysis methods. Simultaneously, many areas of modern sciences are experiencing a "replication crisis," prompting discussions of best practices to produce robust and replicable research. The purpose of this paper is to contribute to this discussion with a particular focus on cognitive electrophysiology. More issues are raised than are answered. Several recommendations are made, including (1) incorporate replications into new experiments, (2) write clear Methods and Results sections, and (3) publish null results.

Why replication is important

The concept of “Truth” needs a different definition in biology compared to physics. The mathematical laws of the universe do not vary according to the seasons or the number of hours since a good meal. Maxwell's equations for electromagnetic wave propagation are not influenced by cultural background or by genetic disorders.

Biology is messy, because Nature promotes diversity. Biological diversity and variability is generally a good thing. It is unlikely that intelligent life would have evolved in a perfectly stationary and ordered environment. On the other hand, diversity and variability are sources of frustration in science, and require scientists to awkwardly straddle ecological validity (poorly controlled and poorly measured diversity) and experimental control (overly constrained environments that might not reflect natural behavior). The brain is perhaps the best example of these difficulties, because it is not only diverse and variable across individuals and species, but it is also highly complex and dynamic within an individual.

Because of this, “Truth” in biology—and certainly in psychology—is difficult to ascertain, and may depend on a variety of factors. To make matters worse, we might not recognize the Truth even if we happen to stumble upon it. Therefore, the best we can strive for is “Consistency.” Results should be regarded in a positive light when they are observed repeatedly in several different situations, from different research groups, and when using different data collection or analysis techniques. In other words, in lieu of an unobtainable absolute Truth, we need replications.

It is very easy to pay lip-service to the importance of replications in science, but more difficult to achieve it in practice. And as data and data analyses become increasingly complicated, replications become increasingly difficult. Even determining whether a finding has been replicated can be difficult to quantify.

The purpose of this paper is to discuss some issues related to replications in the field of *cognitive electrophysiology*, which generally refers to using the brain's electromagnetic fields in order to understand aspects of cognition and how cognitive processes are implemented by neural circuits in the brain. This paper is not the definitive word on how to perform or evaluate replications in cognitive electrophysiology; instead, it is part of a nascent and important discourse about how we can develop and add to a corpus of knowledge that can be written into textbooks and will still be observed in comparable experiments in the future.

Some of the points raised here are general and could be applied to all branches of psychological and neural sciences, while other points are more specific and apply mainly to time-frequency decomposition of neural time series data.

Pressures for and against replication

Needless to say, we all want to do replicable research. No one actually wants to publish findings that cannot be replicated. This is the primary motivation for collecting data from N>1 subjects. Good reputations in science are also built on findings or methods that are replicated and used by other research groups. And reputations can be tarnished by repeated failed attempts to replicate findings. In

other words, there is both individual and career pressure to perform replicable research.

On the other hand, there are also several factors that perhaps unintentionally apply pressure against replications. Top-tier journals (and several mid-tier journals) will reject manuscripts on the basis of insufficient novelty. Major science funding agencies generally do not give grants that only fund replications. And few university departments will be interested in hiring faculty who spend most of their time replicating existing findings. In other words, at multiple levels of the business of science, there is pressure to focus on novel and exciting research rather than on replications.

But the scientist is not just a victim here: Many (or perhaps most) scientists want to do novel and exciting research, because—let's be honest—replicating previous findings gets boring. Scientists have been known to switch fields because they get bored with replicating their own findings.

In some branches of science including cognitive electrophysiology, the issue is exacerbated by the amount of time and expertise required to perform sophisticated analyses of the data. Unlike questionnaire-based or simple computer-based tasks, an EEG study focusing on time-frequency-based analyses might take 1-2 years to complete, and it might require several years of training before being able to analyze the dataset appropriately. A PhD student or postdoc who has limited time and who is under pressure to be competitive for a faculty position or a grant cannot be blamed for wanting to focus their energy on novel experiments rather than on confirmatory replications.

These competing pressures are understandable. There is no science without progress, and progress means looking forward and pushing the envelop of knowledge by making new discoveries. The brain has been such an uncharted territory for the past millenia that it is understandable that the focus has been on new discoveries instead of confirmation and replication. Certainly the brain remains an elusive mystery, but arguably, we've now come far along enough that it is time to shift priorities towards a balance between novelty and replication.

Time-frequency analyses in cognitive electrophysiology

Before discussing issues related to replications, I will first briefly introduce the motivations for and mechanisms of time-frequency analyses. In most cases, researchers use time-frequency-based analyses because they want to make inferences about neural oscillations. The study of oscillations in the brain is a field with growing interest and importance (readers interested in general reviews about neural oscillations may start with Buzsáki and Draguhn 2004; Wang 2010).

Neural oscillations are rhythmic fluctuations in the activity of populations of brain cells. Neural oscillations are present across nearly all of the vast spatiotemporal scales of brain function, from synapses and neurons to circuits, columns, and networks, to patches of brain tissue that are measurable with noninvasive imaging such as EEG or functional MRI. Neural oscillations are perhaps the best candidate feature for understanding how multiple spatial-temporal scales are inter-connected (Le Van Quyen 2011; Palva and Palva 2012; Cohen and Gulbinaite 2013). Furthermore, despite the huge differences in the sizes of the brain over different species, the speeds of neural oscillations have remained remarkably constant (Buzsáki et al. 2013). This suggests that oscillations have a fundamental role in brain function that is conserved across species. Time-frequency-based analyses are the best approach to allow inferences regarding neural oscillations.

The roles of neural oscillations in brain function and neural computation have been discussed and debated for over a century. There are several dominant theories, and many models, simulations, and empirical findings, that support the role of neural oscillations in brain function and cognition. Discussing this literature is beyond the scope of the present paper, but taking a "bird's eye view" of the literature reveals two important ideas about neural oscillations that permeate much of the theorizing and empirical findings.

First, oscillations facilitate the dynamic routing of information across anatomically distinct neural networks (Fries 2005; Jensen and Mazaheri 2010). In part, this occurs because strong oscillations can constrain action potential timing (Vinck et al. 2011; Lisman and Jensen 2013; Reimann et al. 2013), and convergent and synchronized input from many afferent neurons provides a nonlinear boost in post-synaptic input (Kepecs et al. 2002; Eyherabide et al. 2009). Oscillations as a mechanism for controlling the flow of information in the brain is faster and less permanent than synaptic plasticity or other structural changes associated with long-term learning. This allows oscillations to regulate neural information flow over the course of tens to hundreds of ms, i.e., the time-frame of many cognitive processes.

Second, neural oscillations are thought to provide an internal clocking mechanism for coordinating neural computations (Buzsáki and Moser 2013). Neural information processing is highly temporally precise (Cohen 2011), and oscillations provide a temporally precise framework in which a sequence of information processing can be preserved, and in which the upcoming state of a neural network or circuit configuration can be predicted.

Interest in understanding the roles of neural oscillations in the brain has triggered its own rapidly growing subfield of data analysis methods to detect and characterize neural oscillations (Cohen 2014). There are many methods to quantify oscillations in time series data; most of the dominant analysis methods involve using "template matching" procedures in which sine waves or parts of sine waves are compared against the EEG signal, and an algorithm determines the extent to which the time series contains patterns that are similar to the sine wave templates. The Fourier transform and wavelet convolution are two examples of template-matching methods.

The repertoire of time-frequency analysis methods is too expansive to detail here. A quick graphical overview of one of the dominant time-frequency analysis methods is presented in Figure 1.

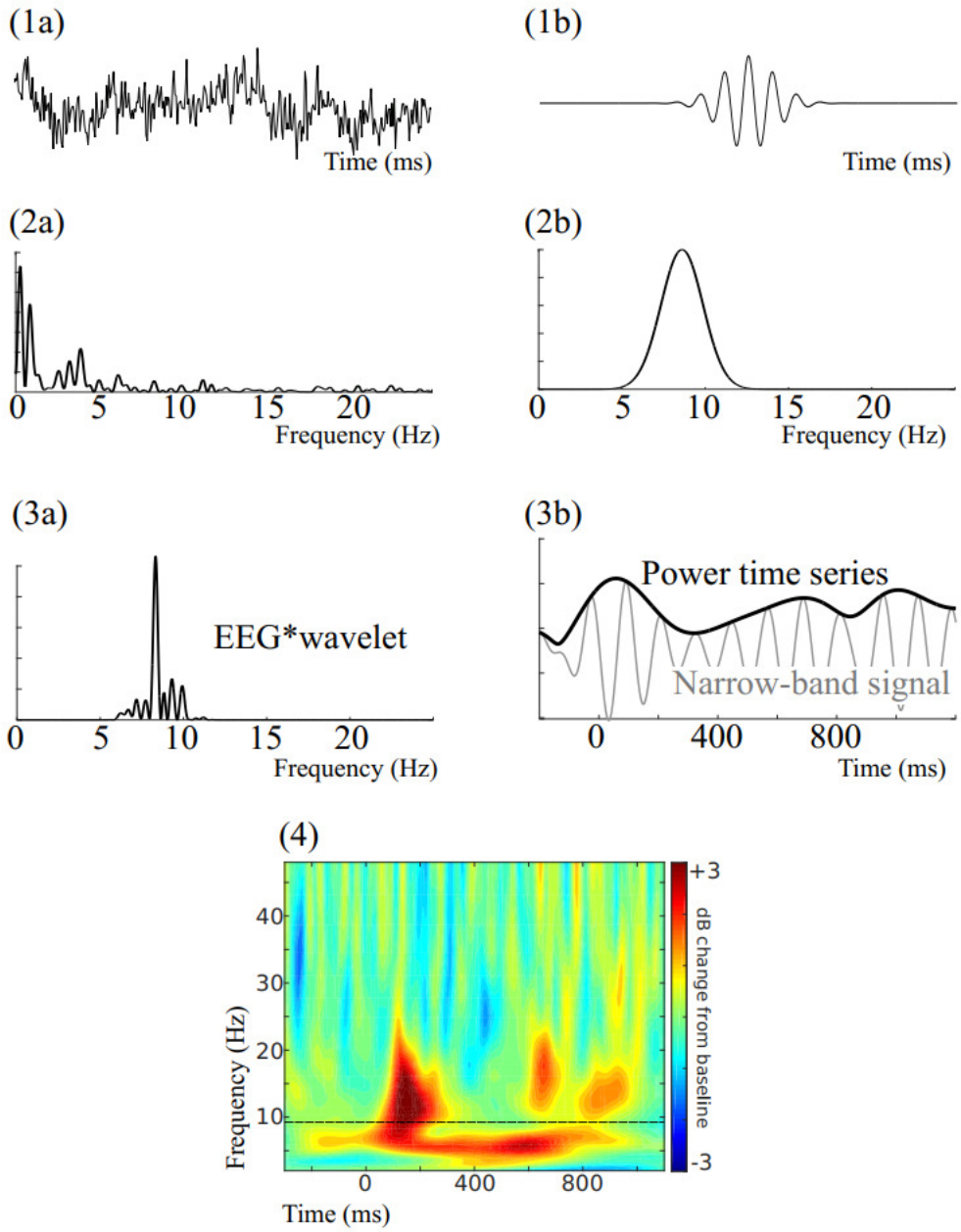

Figure 1. *A graphical overview of using Morlet wavelet convolution for time-frequency analysis. The single-trial data (1a) are convolved with a Gaussian-tapered sine wave (a.k.a. a Morlet wavelet) (1b). In practice, the convolution is done in the frequency domain, in which the frequency spectrum of the EEG data (2a) is multiplied by the frequency spectrum of the wavelet (2b). From the resulting filtered frequency spectrum (3a), the time-course of the band-pass filtered signal and power can be extracted for each single trial (panel 3b shows an example of one trial). When steps 1-3 are repeated over many frequencies, the results can be inspected in a time-frequency power plot (4; the horizontal dashed line is for 8 Hz, the peak frequency of the wavelet shown in steps 1 and 2). It is common to apply a pre-stimulus baseline normalization (here: decibel, dB) to facilitate interpretation and statistical comparisons.*

Replication issues in time-frequency analyses

Time-frequency-based analyses of cognitive electrophysiology data add several dimensions of complexity to replication attempts. Not only are the data transformed into a multi-dimensional space (typically, a time X frequency X electrode X condition space), but time-frequency analyses provide a new framework for additional and physiologically inspired analyses, including functional connectivity, cross-frequency coupling, and spatial multivariate analyses (often called MVPA in the fMRI literature). This increased complexity with many dependent variables makes replications a bigger challenge compared to, e.g., ERP-based studies, in which there is only a small number of dependent variables (typically just one).

This is not a fatal limitation, and we should embrace rather than shy away from the complexity of the brain. But this complexity also means that replications of time-frequency results require additional considerations.

The following sections contain brief discussions of important issues that limit or promote replications. They are listed roughly in chronological order of doing experiments, from experiment design to data collection and processing to analyses and reporting results.

Experiment design for time-frequency analyses

Proper experiment design can improve scientific quality, reduce or prevent headaches during data analysis and publishing, and promote replications. For general discussions about experiment design in EEG, see Luck (2014) or Cohen (2014). Three specific points will be highlighted here.

1) First and most important: Design your experiments with replications in mind. This statement has two interpretations. First, when designing your experiment, keep in mind the possibility that other researchers will want to replicate your experiment and results. Furthermore, keep in mind that other researchers may ask for your experiment and analysis code. That should provide motivation to double-check code for errors and write clean and commented experiment and analysis code.

Second, whenever possible, try to incorporate replications of existing findings into new experiments. The reasoning behind this will be explained in a later section. Cognitive electrophysiology (and many other branches of science) needs to find a balance between replicating existing findings and producing new findings. This balance can be built naturally into research by ensuring that new experiments allow for replications and novel findings.

2) Think carefully about an appropriate time period for baseline normalization. For ERPs, baselining simply involves a linear subtraction of the pre-trial (or sometimes pre-response) average voltage value. For time-frequency power, however, a modification of this procedure is required for two reasons.

First, power (the amount of energy in a signal at each frequency) generally follows a “1/f” pattern, meaning that power decreases with increasing frequencies, typically in a supralogarithmic manner. Comparing power across frequencies therefore requires a nonlinear normalization—typically, decibel or percent change.

Second, because of temporal smoothing inherent in time-frequency analysis methods, a baseline period

that ends at time=0 is suboptimal because early post-stimulus activity may "leak" into the estimate of the baseline activity. In most situations, a pre-trial baseline period of -500 to -200 ms is sufficient. The choice of baseline should be considered during experiment design.

3) Have "enough" trials per condition and "enough" subjects. Unfortunately, it is difficult to give precise numbers that would be appropriate for all experiments, because it depends on the effect size and the quality of the data. As a general guideline, I recommend a minimum of 50 trials per condition per subject (ideally >100), and at least 20 subjects.

When possible, trial counts should be matched across experiment conditions. Trial count differences can lead to condition differences in signal-to-noise characteristics, and for some analyses including phase-based analyses, unbalanced trial counts can bias the results towards the conditions with fewer trials.

The appropriate trial count also depends on the frequency of interest and on the analysis parameters that are used. Low frequencies, and analysis parameters that involve more smoothing, generally produce higher signal-to-noise results and therefore may require fewer trials. Previous studies with similar experiment design and data analyses can be a guide, but piloting is nearly always the best way to estimate effect sizes.

In the pilot experiment, a few motivated subjects can perform the experiment using as many trials as they can tolerate. In the analyses, the reliability of the effects can be assessed using random subsets of trials. The minimum number of trials that reliably reproduces the effects observed with all of the trials can be taken as the minimum number of trials to produce a stable effect.

When data are acquired from living animals, be aware that there is a balance between a high trial count and engagement in the task. After too many trials, subjects may become disengaged or tired, and this may result in noisy data that only decreases statistical robustness and replicability. This is a greater danger in some populations such as children or patients.

Data preprocessing

"Preprocessing" refers to all of the steps taken after the raw data are collected and before single-subject analyses are performed. Preprocessing steps generally include high-pass filtering (typically around .1 or .5 Hz), epoching around the events of interest, rejecting artifacts (discussed more below), and interpolating bad channels. Most preprocessing steps are fairly standard and arouse little concern about biases or procedures that could affect replication.

The exception is artifact rejection. Artifact rejection is necessary because the data contain noise that can adversely affect the results, and the noise must be attenuated or removed. Artifact rejection generally involves two steps: trial rejection and ICA (independent components analysis) subtraction. Despite some well-intentioned efforts to automate artifact rejection, I will argue below that the best artifact rejection procedure is manual, i.e., human-based and visually guided, and therefore unfortunately open to subjectivity.

The argument for automatic artifact rejection is understandable: An ideal data-cleaning method should be fast and easy to use, and should produce the same results regardless of whether a novice student or

seasoned researcher is handling the data, and regardless of how many times the data are processed.

The problem is that I and many other M/EEG researchers with whom I have spoken agree that automatic algorithms produce both Type-I and Type-II errors. That is, algorithms reject data that a human would include, and fail to reject data that a human would reject. Part of the problem is that the criteria for what is “signal” and what is “noise” may depend on the experiment, the data, and the planned analyses.

On the other hand, manual rejection is a source of subjectivity. Although some artifacts are obvious enough that anyone—or any computer—would remove them, other artifacts are not obvious, or are mixed in with signal, and two different individuals may remove somewhat non-overlapping parts of the data. It is also possible that the same individual would reject different parts of the same dataset when re-processing the data, e.g., in the morning vs. in the evening.

In my opinion, manual data cleaning is the lesser of two evils. A decent trade-off is to use automated procedures to flag trials or independent components for possible rejection, and a human can approve or veto each flagged epoch or component, as well as identifying unmarked data for rejection. Going through all of the raw data takes time, but if it is done carefully and meticulously, it needs to be done only once. Carefully inspecting data is the only way to ensure high data quality. I do not trust my own results until I have seen the raw data, and I hope other researchers feel the same.

Deciding which independent components to subtract from the data is another potential source of subjectivity. Components containing oculomotor artifacts (e.g., from blinks) are generally easy to find and safe to remove. Other components can be more difficult to make decisions about, because noisy components often contain some signal. Unless the noise is extreme, I recommend leaning on the side of not removing components. Thinking that EEG data can or should be noise-free is dangerous and incorrect, and it's better to leave some noise in than take signal out (extreme cases notwithstanding). When in doubt, ask a colleague to look at the component.

Given the inherent subjectivity in manual data cleaning, there should be a clear set of criteria that is applied to all datasets. Data cleaning is time-consuming and important, and yet can become boring. I encourage people not to clean more than 2-3 datasets per day. Exhaustion and frustration can change the criteria. One way to assess the subjectivity is to have two independent raters perform manual trial/ICA rejection on the same datasets, and measure inter-rater-reliability. This would promote identifications, discussions, and resolutions of differences in opinion of data rejection criteria. A systematic characterization of this reliability would be a welcomed addition to the literature.

Does this subjectivity translate to biases that make results more difficult to replicate? I think there is little cause for alarm here. It is important that the person doing the cleaning is blind to experiment condition and group assignment. One should not think “this trial is from condition 'A' and so therefore I should/shouldn't reject it.” Cleaning the data prior to separating the epochs into each condition is a good way to ensure condition-blindness. If the experiment involves different groups (e.g., patients and controls), the person cleaning the data should see only arbitrary dataset identifiers and not group membership. Ideally, only one person should clean the entire dataset.

A common theme in discussions about data cleaning/analysis is that when the potential for bias is unavoidable, biases should be applied to all parts of the dataset to minimize the possibility of

systematically affecting condition or group comparisons.

Data analyses with toolboxes vs. custom-written code

There are a few advantages of toolboxes: they are standardized and generally easy to use; they free the need to learn sufficient math and programming to implement the analyses on one's own; and they facilitate replications because the same functions can be used by different researchers (although functions and results may vary across toolbox or software versions).

On the other hand, only performing analyses that are available in toolboxes can be limiting. Most analysis toolboxes are developed by research groups, and their toolboxes are typically oriented to the kinds of data analyses preferred by those groups.

Thus, in practice, many researchers use a combination of custom-written code and toolbox functions, or sometimes entirely custom-written code.

Either way, there is no reason not to share code. There are now many websites that can be used to share code (in 2015, examples include github, google-code, figshare, and personal or lab websites). And code files are small and easy to transport digitally. This makes sharing code much easier than sharing empirical datasets, which can be hundreds of gigabytes.

Finally, sharing code has a hidden advantage beyond letting others replicate analysis methods. When someone knows that their analysis code may be seen and used by others, there is greater motivation for writing clean and commented code. And this in turn reduces the possibility of honest but careless programming errors. Mistakes are unavoidable in human endeavors, and strategies that help minimize mistakes should be embraced. Making code available also means that someone else will check your code and potentially find any mistakes. Of course, being notified that a programming mistake may have affected a published result is embarrassing and something we all want to avoid, but the progress of science is better served by catching and fixing mistakes.

Within-subject vs. group-level analyses

Statistical analyses can be done at two different “levels,” where “level” refers to what is considered the unit of data for statistical analysis. Within- subjects statistics (also called level-1) consider the trial to be the unit for analysis, while group-level statistics (level-2) consider trial-averaged data per subject to be the unit for analysis. Sometimes the term “single-trial analysis” is used in place of “within-subject analysis,” but this is typically a misnomer because many trials are included in the analysis.

It is important to realize that within-subjects and group-level analyses have different goals and different interpretations. Within-subjects analyses provide information regarding the cross-trial variability of an effect relative to the magnitude of the effect; they provide no information regarding the generalizability of the effect to other subjects. Group-level analyses, on the other hand, provide information regarding the consistency of the direction of the effect across the group of subjects, and provide little information regarding the within-subject variability across trials. Figure 2 shows how within-subject and group-level analyses can be dissociable.

Within-subject analyses can be informative and should not be discouraged. They provide insight into

the robustness of the effects (for example, by performing split-half replication, which may be impossible at the group level if there are too few subjects to create two groups), can be used to link EEG dynamics to behavior, and can be used to determine whether effects change in magnitude over time (as a function of practice or fatigue). In this sense, within-subjects analyses are complementary to group-level analyses.

In terms of reproducibility and replication attempts, group-level analyses are more relevant because they provide deeper insights into how likely the findings will generalize to individuals other than those from whom data were collected.

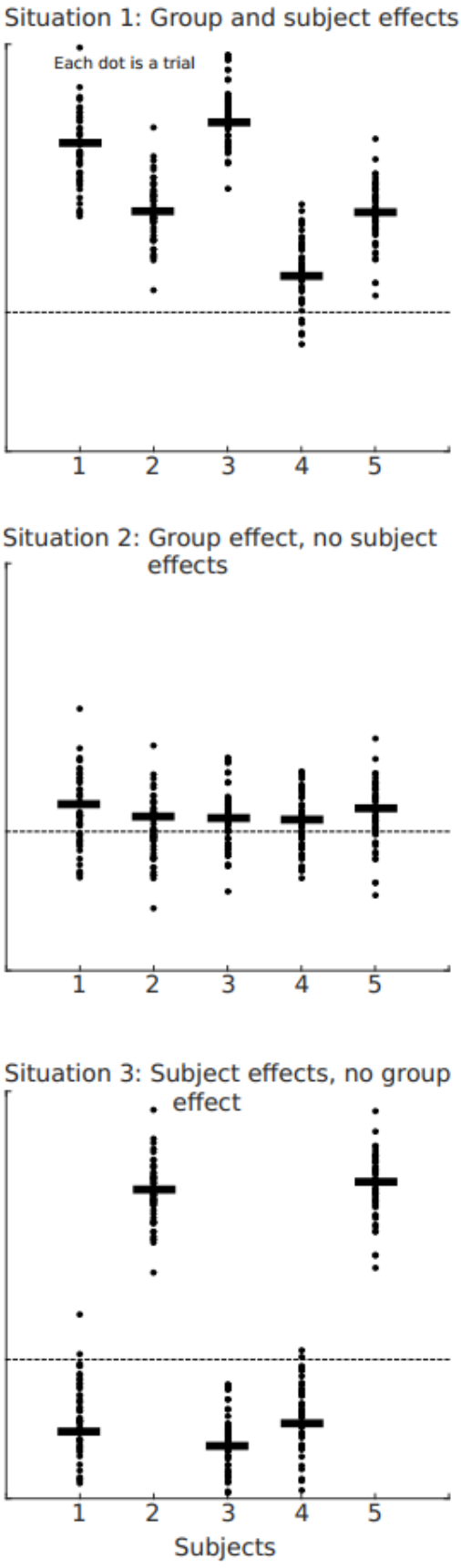

Figure 2. *Within-subjects and group-level analyses can be dissociated and reveal different patterns in the data. In these simulated data, each x-axis tick corresponds to a subject, each dot corresponds to a trial, and the short horizontal bars correspond to the average of trials within each subject. The dashed horizontal line corresponds to zero. An "effect" here refers to the trial-average being significantly different from zero. In situation 1, the effect is significant within each subject and at the group level. In situation 2, no individual subject exhibits a statistically significant effect due to high variability relative to the effect size (distance of the average across dots from the zero line), but because the trial-average effect is consistently above zero, the group-level effect would be significant. In situation 3, each subject individually exhibits a significant effect but there is no significant group effect. There could be a group effect if the cause of the variability is known, for example if subjects 2 and 5 received a medication and subjects 1, 3, and 4 received placebo.*

Selecting data from subject-specific vs. subject-average regions-of-interest

Group-level analyses generally involve pooling data over subjects and testing for significance within defined time-frequency-electrode windows. There is a strong—and typically unstated—assumption in group-level analyses, which is that there is no meaningful variability in temporal-spatial-spectral localization across subjects. That is, group-level analyses assume that, for example, activity from 8-10 Hz and 200-500 ms at electrode Pz reflects the same process for all subjects.

This is a questionable assumption. For example, there are considerable individual differences in peak frequency (Haegens et al. 2014), which are related to a combination of genetics (Posthuma et al. 2001), neurochemical factors (Muthukumaraswamy et al. 2009, but see Cousijn et al. 2014 for a non-replication), age (Polich 1997) and so on. This issue is relevant for replications, because to the extent that meaningful individual differences are ignored, statistical power is reduced and replication attempts are less likely to be successful.

On the other hand, that the approach of taking the same time-frequency-electrode window for all subjects is nearly ubiquitous in the cognitive electrophysiology literature demonstrates that the assumption is, to some extent, viable. The main advantage of selecting a common window is that it helps account for individual variability in temporal-spatial-spectral characteristics. This is a similar justification for why some temporal-spectral smoothing is beneficial.

There are two alternative approaches that may increase statistical sensitivity by incorporating individual variability in brain functional and structural anatomy. One approach is to select electrodes and/or time-frequency windows for each subject. As long as the data selection method is independent of the intended statistical contrast, circular inferences (a.k.a. double-dipping) can be avoided (Kriegeskorte et al. 2009; Cohen 2014). A second approach is to use weighted averaging techniques to combine data across all electrodes. This approach includes methods such as ICA, joint decorrelation, generalized eigenvalue decomposition, principle components analysis, and machine-learning techniques (Onton et al. 2006; Garcia et al. 2013; de Cheveigné and Parra 2014; Haufe et al. 2014). In these analyses, an electrode weighting matrix is defined based on statistical characteristics of each subject's data. Group analyses based on these weighted averages might be more sensitive than using the same electrode for all subjects.

New analysis methods

Data analysis methods development is an important subdiscipline of cognitive electrophysiology. Methods development is necessary because EEG data are multidimensional and we lack a complete understanding of how EEG dynamics are generated by and related to neural dynamics. The burden of benefit, however, should be greater for those developing new methods relative to those using established methods. New data analysis methods should be justified by their relationship (even if speculative) to neurophysiology, and there should be clear and demonstrated advantages over existing methods. Furthermore, analysis code should always be made available so non-experts can evaluate and use the new methods (this seems like too obvious a point to mention, but many methods papers report only equations and do not provide implementations).

Show data

Needless to say, results can be evaluated only if they are shown. There are two reasons to show data in figures. First, it will help determine whether the effect is specific to a certain time-frequency-electrode window. The interpretation of a finding depends in part on whether the effect is localized in time, frequency, and space. Second, showing data will allow readers to visually inspect a richer portion of the results. This may lead to additional task-related dynamics that were not specified by the hypotheses, and will facilitate subsequent replications. This is particularly important for new investigations when the time-frequency characteristics are not well established.

It is the authors' responsibility to create figures that illustrate a sufficient amount of data. And reviewers should not hesitate to ask authors during revisions to include additional time-frequency or topographical plots. On the other hand, we must recognize that it is impossible to show all results: Showing time-frequency plots from electrodes that have no task response will only lead to confusion and overwhelm the reader with plots. An example of how to show time-frequency results is presented in Figure 3. More generally, authors and reviewers can use the following list of questions when producing and evaluating figures (adapted from page 514 in Cohen, 2014).

- What is the main message of the figure?
- Does the figure convey one idea, several related ideas, or several different ideas?
- Can you easily discern which lines/bars/plots correspond to which conditions/groups/ analyses?
- Can you easily distinguish the significant results from the nonsignificant results?
- Is the color necessary, helpful, or annoying?
- Will the figure be interpretable when printed in grayscale?
- Can the figure be interpreted on its own, or do you need to read the text to understand it?
- Is the figure legend clear and understandable without reading the text?
- For time-frequency plots, are there important results that are cut off by the limits of the plot? If so, the plot should be recreated to show where the key findings start and stop in time and in frequency (and, if relevant, in space).

It is difficult to understand why authors choose to show little data. Too often, the figures show a highly processed abstraction of the results without showing the necessary steps before getting to that point. It is likely that you have had the experience of thinking that, for example, the paper's figure 2 was step 6 of an analysis pipeline, and steps 1-5 are not shown. This is really to the paper's detriment: Sparsity of results make papers more easily criticized and more easily forgotten, and therefore less likely to be cited.

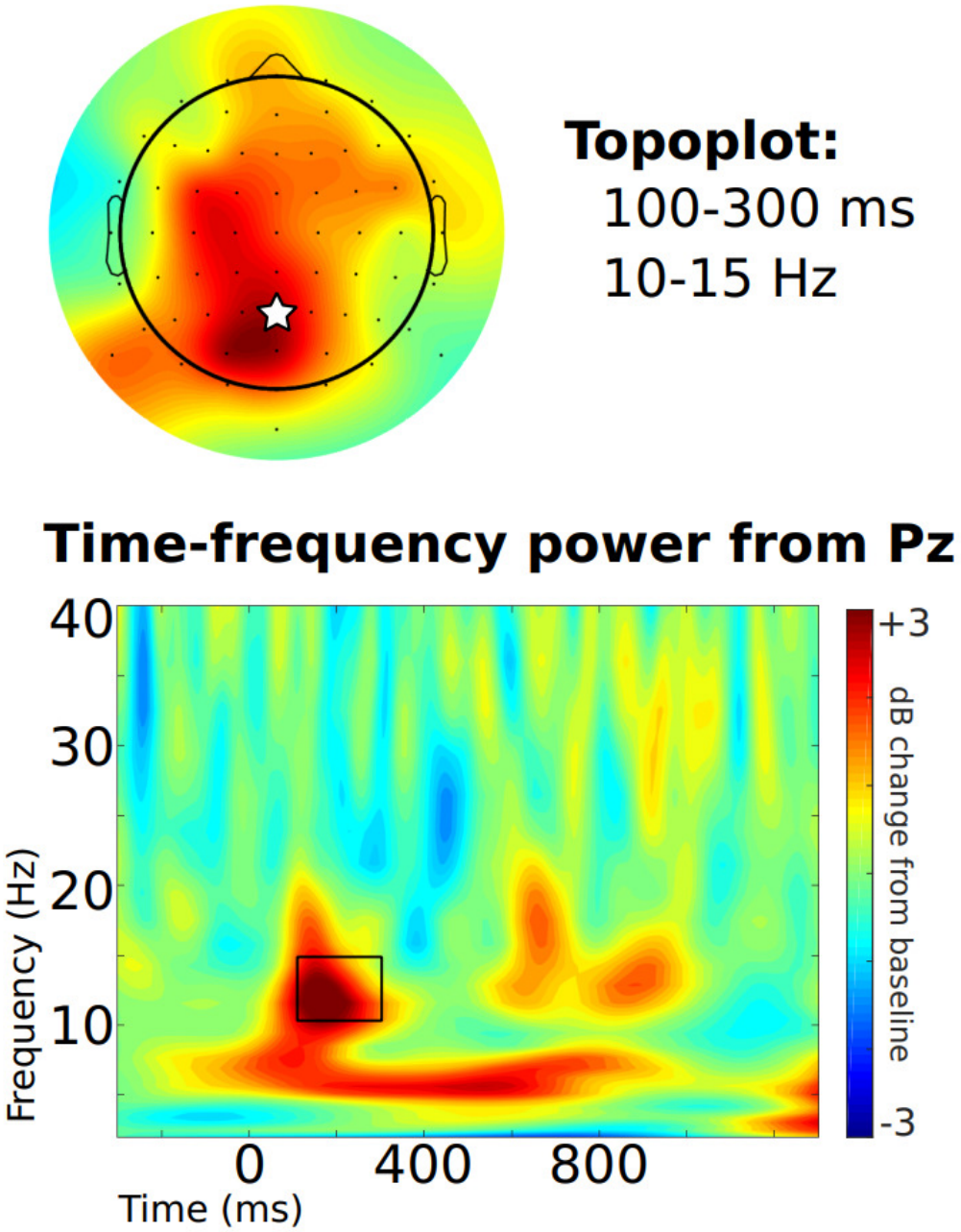

Figure 3. *An example of how to show time-frequency results in a figure. These two images (topographical map and time-frequency plot) show power over time, frequency, and space. The time-frequency plot is take from the electrode with a star (Pz), and the topographical map is taken from the power data averaged in the time-frequency window indicated by the black box. This is not the only appropriate way to show such data, but rather is one illustration of how to show the rich dynamics revealed by time-frequency-based analyses.*

Inconsistent terminology

One issue that impedes replication and causes confusion is inconsistent terminology. Inconsistent terminology means that different terms can refer to the same analysis, and the same term can refer to different analyses. One example is time-frequency power. The power of a frequency-band-specific signal is the squared amplitude of the analytic signal, and can be conceptualized as the trough-to-peak distance for sinusoidal processes. The term “power” is widely accepted in engineering and in physics. In cognitive electrophysiology, on the other hand, various terms are used, including synchronization, representation, spectral perturbation, activation, and ERD and ERS (event-related synchronization and desynchronization). I don't know what motivated this terminological proliferation, but it impedes direct comparisons across findings and research groups. “Synchronization” is a particularly poor term, because it can refer either to the power within a single electrode, or to a measure of functional connectivity across different electrodes.

Ideally, there should be a one-to-one mapping between the name of a method and the math behind that method. Consider, for example, the terms correlation, ANOVA, and factor analysis. There is no ambiguity about which sets of equations are employed when someone says that he performed an ANOVA. In cognitive electrophysiology, if someone says that there was an increase in alpha synchronization, without further interrogation it is unclear whether that person means an increase in power or phase-based connectivity. This is an important distinction because these two analyses have very different interpretations, putative neurophysiological origins, theoretical implications, and methodological concerns.

Unfortunately, the situation is unlikely to improve any time soon, in part because many researchers continue to use their own terminology for the sake of internal consistency, and in part because some researchers seem unconcerned about terminological inconsistencies. Realistically then, it is more important to write clearly what analyses were used and to describe how the analyses were done. Ambiguous terms like “synchronization” and “time-frequency response,” if used, should be clearly and carefully defined in the Methods section.

Preregistered reports

Preregistration is a model for scientific publishing in which authors essentially have the Methods section peer-reviewed prior to starting the study, and once approved by reviewers and editors, the manuscript is in-principle accepted provided the experiment is conducted along the lines of the approved manuscript.

It is difficult to argue against this being a good model of scientific communication. Scientists should be rewarded for their ability to think and reason critically and carefully, for their ideas and hypotheses, and for being able to anticipate the analyses that best test their initial hypotheses. Preregistered reports also allow for unplanned, post-hoc, and exploratory analyses, so long as those analyses are clearly indicated as being post-hoc in the final paper.

Preregistered reports increases pressure on scientists to think carefully and critically about their hypotheses and analyses, while reducing pressure to spin a story around getting and selling “sexy” results. At present, only a few peer-reviewed journals offer preregistered reports (in 2015, the journal

Cortex is the most frequently mentioned in this context).

Report null results

If you see an exciting finding in one paper but have not seen it replicated in other papers, it is difficult to know why. It could be that other people have tried and failed to replicate the finding but did not publish it (the "file drawer phenomenon"), or it could simply be that no one has tried to replicate it.

Not publishing null results when there were no serious flaws in the experiment or analyses should be considered a minor violation of scientific ethics, for two reasons. First, scientific progress comes not only from knowing which hypotheses and theories are confirmed; progress also requires knowing which hypotheses and theories need to be revised or abandoned. Publishing only positive results and hiding null results supports only the first stream of scientific progress.

Second, not publishing null results can waste the time of other researchers. Who knows how many times the same experiment has been attempted by different groups, independently of each other and unbeknownst to each other because most people refuse to publish (and sometimes even to talk about) valiant scientific efforts that produce null effects. I'll share a personal experience with this situation: A student and I conducted a study on visual flicker and response conflict processing. The study was well-designed, well-thought-through, and the quality of the behavioral and EEG data was high. But we found no support for our hypotheses, and the study went unpublished. A few years later a colleague from a different university mentioned to me their repeated but unsuccessful attempts at a similar experiment that were never published. Our independent but collective failures to publish good science with null results translated into wasted time testing ideas that other researchers had demonstrated are unlikely to succeed.

Null findings are difficult to publish in peer-reviewed journals, particularly high-impact and "luxury" journals that often use media-friendly titles and author lists as considerations for acceptance. Fortunately, the growing use of online repositories such as arXiv.org and bioRxiv.org, as well as blogs and other non-peer-reviewed online sources, allow more findings to be available to the scientific community. The aforementioned study from my student and me is now posted on bioRxiv (Cohen and Steel 2014). Unfortunately, these outlets are generally not listed in pubmed.com or other scientific search engines, but can be found using internet searches. But null findings will be more likely to be published in peer-reviewed journals when bundled with additional analyses, as discussed in the next section.

Increasing replications in cognitive electrophysiology

As described earlier, there are multiple sources of pressure in science against replications and towards novel findings. But replications are important. In my opinion, attempts to focus entire research plans or careers on replications is not a viable or sustainable long-term strategy, because few researchers are willing to dedicate their time entirely to replications. Instead, a more realistic and sustainable strategy is to combine replications and novel results into the same experiments and the same publications.

Electrophysiology is advantageous in this approach. Because EEG provides rich and multivariate datasets, replicating an EEG study need not be as dry as replicating experiments with a small number of dependent variables. That is, the replication itself could be only one part of the Results section.

Additional analyses can then be performed to help contextualize the original findings, to link the findings to other areas of the literature, and to perform novel analyses not previously reported. If the original findings are not replicated, additional analyses can be performed to investigate why the original results were not replicated, for example by selecting a subsample of subjects who exhibit some particular behavioral effect.

Attempts to replicate a result need not be driven by a lack of trust in the original authors. Furthermore, failures to replicate do not imply that the original authors were lying or that they manipulated their data. Of course, we all hope that our findings will generalize and replicate in other studies, but it is inevitable that some findings in the literature are statistical false alarms, or are limited to a narrow population or specific experiment design. It is important for progress in cognitive electrophysiology (and science more generally) to determine which findings are overinterpreted and which are more stable, and we must all take some responsibility to work towards this goal.

Conclusion and outlook

The branch of cognitive electrophysiology that is focused on linking M/EEG data to neurophysiology and brain computations underlying cognition has grown from inception to infancy, and continues to produce large amounts of rich, multivariate, and multidimensional data. The purpose of this paper is to contribute to nascent and ongoing discussions about how to make sure that our field is building a corpus of reproducible results.

There is a plurality of opinions about the current publishing system—ranging from leaving it as-is to completely over-hauling or even abandoning it. In my opinion, radical proposals to make qualitative changes to how science is conducted and published will be met with friction and are unlikely to succeed. Given how interwoven our current system is in career evaluations, grant proposals, etc., I believe that the best strategies to promote replicable research will involve positive adjustments to existing research practices, including publishing null results, making analysis scripts and data available, and balancing replications and novel findings in the same publication.

Looking back on this paper and recalling the related discussions I have had on this topic with colleagues, this area of science is too young and too exploratory to impose strict rules on what can and cannot be done in experiments or with data. Cognitive electrophysiology has a “wild-west” feel to it, and this should be embraced rather than discouraged.

References

Buzsáki G, Draguhn A. Neuronal oscillations in cortical networks. Science [Internet]. 2004 Jun 25 [cited 2014 Apr 28];304(5679):1926–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15218136

Buzsáki G, Logothetis N, Singer W. Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. Neuron [Internet]. 2013 Oct 30 [cited 2013 Nov 6];80(3):751–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24183025

Buzsáki G, Moser EI. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. Nat Neurosci [Internet]. 2013 Feb [cited 2014 Jul 11];16(2):130–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4079500&tool=pmcentrez&rendertype=abstract

de Cheveigné A, Parra LC. Joint decorrelation, a versatile tool for multichannel data analysis. Neuroimage [Internet]. 2014 Sep [cited 2014 Jul 15];98C:487–505. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24990357

Cohen MX. It's about Time. Front Hum Neurosci [Internet]. 2011;5(January):2. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3025647&tool=pmcentrez&rendertype=abstract

Cohen MX. Analyzing Neural Time Series Data: Theory and Practice. Cambridge: MIT Press; 2014.

Cohen MX, Gulbinaite R. Five methodological challenges in cognitive electrophysiology. Neuroimage [Internet]. 2013 Aug [cited 2013 Aug 20];85:702–10. Available from: http://dx.doi.org/10.1016/j.neuroimage.2013.08.010

Cohen MX, Steel FW. Pre-trial exogenous visual flicker does not affect behavioral or EEG signatures of conflict processing. bioRxiv [Internet]. 2014 May 5; Available from: http://biorxiv.org/content/early/2014/05/05/004747.abstract

Cousijn H, Haegens S, Wallis G, Near J, Stokes MG, Harrison PJ, et al. Resting GABA and

glutamate concentrations do not predict visual gamma frequency or amplitude. Proc Natl Acad Sci [Internet]. 2014 Jun 9 [cited 2015 Oct 22];111(25):9301–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4078853&tool=pmcentrez&rendertype=abstract

Eyherabide HG, Rokem A, Herz AVM, Samengo I. Bursts generate a non-reducible spike-pattern code. Front Neurosci [Internet]. 2009 May [cited 2013 Sep 5];3(1):8–14. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2695386&tool=pmcentrez&rendertype=abstract

Fries P. A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. Trends Cogn. Sci. 2005. p. 474–80.

Garcia JO, Srinivasan R, Serences JT. Near-real-time feature-selective modulations in human cortex. Curr Biol [Internet]. 2013 Mar 18 [cited 2015 Oct 23];23(6):515–22. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3608396&tool=pmcentrez&rendertype=abstract

Haegens S, Cousijn H, Wallis G, Harrison PJ, Nobre AC. Inter- and intra-individual variability in alpha peak frequency. Neuroimage [Internet]. Elsevier B.V.; 2014 Feb 6 [cited 2014 Feb 19];1–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24508648

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage [Internet]. 2014 Feb 15 [cited 2015 Jul 7];87:96–110. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24239590

Jensen O, Mazaheri A. Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Front Hum Neurosci [Internet]. 2010 Jan [cited 2014 Jul 11];4:186. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2990626&tool=pmcentrez&rendertype=abstract

Kepecs A, Wang X-J, Lisman J. Bursting neurons signal input slope. J Neurosci [Internet]. 2002 Oct 15;22(20):9053–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12388612

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci [Internet]. 2009 May [cited 2014 Jul

10];12(5):535–40. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2841687&tool=pmcentrez&rendertype=abstract

Lisman JE, Jensen O. The θ-γ neural code. Neuron [Internet]. 2013 Mar 20 [cited 2014 Feb 20];77(6):1002–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23522038

Luck SJ. An Introduction to the Event-Related Potential Technique. 2nd ed. Cambridge: MIT Press; 2014.

Muthukumaraswamy SD, Edden RAE, Jones DK, Swettenham JB, Singh KD. Resting GABA concentration predicts peak gamma frequency and fMRI amplitude in response to visual stimulation in humans. Proc Natl Acad Sci U S A [Internet]. 2009 May 19 [cited 2015 Oct 22];106(20):8356–61. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2688873&tool=pmcentrez&rendertype=abstract

Onton J, Westerfield M, Townsend J, Makeig S. Imaging human EEG dynamics using independent component analysis. Neurosci Biobehav Rev [Internet]. 2006 Jan [cited 2015 Sep 5];30(6):808–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16904745

Palva S, Palva JM. Discovering oscillatory interaction networks with M/EEG: challenges and breakthroughs. Trends Cogn Sci [Internet]. 2012 Apr [cited 2014 Dec 9];16(4):219–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22440830

Polich J. EEG and ERP assessment of normal aging. Electroencephalogr Clin Neurophysiol [Internet]. 1997 May [cited 2015 Oct 22];104(3):244–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9186239

Posthuma D, Neale MC, Boomsma DI, de Geus EJ. Are smarter brains running faster? Heritability of alpha peak frequency, IQ, and their interrelation. Behav Genet [Internet]. 2001 Nov [cited 2015 Oct 22];31(6):567–79. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11838534

Reimann MW, Anastassiou CA, Perin R, Hill SL, Markram H, Koch C. A biophysically detailed model of neocortical local field potentials predicts the critical role of active membrane currents. Neuron [Internet]. 2013 Jul 24 [cited 2014 Feb 24];79(2):375–90. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/23889937

Le Van Quyen M. The brainweb of cross-scale interactions. New Ideas Psychol. 2011;29:57–63.

Vinck M, Oostenveld R, van Wingerden M, Battaglia F, Pennartz CMA. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. Neuroimage [Internet]. 2011 Apr 15 [cited 2014 Mar 19];55(4):1548–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21276857

Wang X-J. Neurophysiological and computational principles of cortical rhythms in cognition. Physiol Rev [Internet]. 2010 Jul [cited 2013 Dec 25];90(3):1195–268. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923921&tool=pmcentrez&rendertype=abstract