

Physiological plausibility can increase reproducibility in cognitive neuroscience

Freek van Ede¹ & Eric Maris²

¹ *Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford, United Kingdom*

² *Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands*

*Correspondence: e.maris@donders.ru.nl (E. Maris)

Key words: Reproducibility, Cognitive Neuroscience, Statistics, Dissemination, Electrophysiology

Physiological measurements offer the unique opportunity to assess plausibility along multiple data dimensions, in addition to the significance of statistical tests. Increased emphasis hereon should help increase reproducibility of research findings, by turning what is often considered a threat to reproducibility (a large search space) into part of the solution.

The last years have seen a surge of reports questioning the reproducibility of research findings in, amongst others, psychology and neuroscience (e.g. [1-3]). Part of this low reproducibility is due to the probabilistic nature of data-based inferences: observed effects are inferred to be genuine when their probability of being due to chance is smaller than 5% (i.e. $p < 0.05$). Based on this widely applied statistical norm, it follows that 1 out of every 20 inferences of a genuine effect will be, in reality, due to random error. Although 5% may be considered an acceptable risk, taking into consideration that insignificant results are often left unpublished, the actual percentage of statistical flukes is likely to be substantially higher when only considering published work. Thus, even leaving aside psychosociological factors that promote undesirable and biased research practice (as discussed in e.g. [1,4-6]), it is perhaps not surprising that many published research findings do not reproduce well.

For many cognitive psychology experiments involving a limited number of outcome measures per trial or per participant (such as accuracy and reaction time, resp., their across-trial averages), statistical measures are often the only available means to distinguish genuine from random differences between experimental conditions. Physiological measurements (such as electroencephalographic [EEG] recordings) on the contrary, often involve measurements that span across the levels of multiple dimensions, such as space, time, and frequency. As we will argue, physiological measures therefore provide the unique opportunity to assess plausibility in addition to p-values. Such physiological plausibility is ideally suited to help further distinguish genuine from random differences and an increased emphasis hereon should help increase the reproducibility of genuine research findings.

Key to our argument is that genuine physiological events (and hence their experimental modulations) are often not stand-alone events that are restricted to a single recording site, time point, or frequency bin. Rather, such events are highly structured, involving a large degree of correlation between neighboring recording sites, time points and frequency bins. As such, the degree to which an experimental modulation is structured along these dimensions provides strong complementary information to statistical p-values alone. In fact, whereas a given (possibly hypothesized) point in the multi-dimensional space may be characterized by the same statistical value (see highlighted data points in Fig. 1), the structure in the data will provide critical complementary information by deeming the effect in the left scenario in Figure 1 far less plausible than the “same” effect in the right scenario.

In essence, our proposal thus capitalizes on the fact that experimental manipulation of physiological processes show up as spatially, temporally, and/or spectrally correlated effects. This immediately raises the question how one has to distinguish patterns of correlated effects from patterns of correlated noise. In fact, for biophysical and signal processing reasons (resp., volume conduction, and artificial signal smoothing), also the noise in the physiological signals is likely to exhibit correlation. Fortunately, recent advances in both parametric [7] and nonparametric statistics [8] have provided the tools to control false discoveries in the context of correlated noise.

Although our argument is straightforward, we believe that its acknowledgement, when taken seriously, bears important implications for current research practice. In fact, in current practice, inferences are often solely based on p-values, while the very subject of these inferences (the spatio-spectro-temporal pattern in the data) is often only sparsely described. For example, although a typical EEG/MEG analysis unfolds over three dimensions (time, space, frequency), one often comes across published research articles in which the physiological effects of interest are described and depicted solely with regard to a single dimension (such as a time course averaged across several recordings sites and frequency bins), or even reduced to a single bar graph. Moreover, when considering analyses that involve between-site and between-frequency interactions (e.g. analyses of distributed phase-amplitude coupling), the degree of data reduction is often much larger, because these analyses involve *two* spatial and *two* spectral dimensions over which one can average.

There are at least two limitations of such data reduction at the dissemination stage. First, it obscures the ability to evaluate an effect's physiological plausibility, and thereby refrains an audience from relevant complementary information. Second, and more seriously, data may sometimes be depicted in reduced format because an effect appears more convincing in one dimension than in another – and authors may be inclined to depict their data in its “best” light. Such undisclosed flexibility in data analysis and reporting is a highly potent contributor to low reproducibility [1]. This is especially problematic for the same types of data that we consider to offer the unique opportunities for evaluating plausibility. This is because such data provide a large search space in which effects can be found at many places. The key point is that when patterns of interest would always be evaluated and depicted along *all* relevant dimensions of the data, then this would not only better reveal their plausibility but also make the potential contribution of data selection much more transparent. As such, the very feature of the data that is often considered a threat to reproducibility (a too large search space) can become part of the solution (see also Box 1 for a list of further recommendations).

Noteworthy, the issue of selective reporting is also relevant when employing statistical tests that evaluate the full dataspace and that deal with the multiple comparisons problem. This is because popular methods for multiple comparison correction [7, 8] produce so-called “clusters of significant activity” without the false alarm rate being controlled at the level of the elements of these clusters (e.g., sensor-time pairs, sensor-time-frequency triplets). Instead, only the false discovery rate is controlled [7] or the false alarm rate is only controlled at the level of the full dataspace [8]. This issue has been raised previously [9,10], and it is relevant here because these multi-dimensional clusters of “significant” activity can be depicted selectively. For example, the time-frequency profiles from two different sites that are both part of the same three dimensional cluster may still differ substantially (in Box 1 we make a specific recommendation for dealing with this).

Reduction of the data space in the stage of reporting must be distinguished from its reduction in the analysis stage. In the latter stage, one can sometimes make use of an independent localizer or an a priori hypothesis about the location (in space, frequency and/or time) of the effect of interest. Assuming validity of the independent localizer or the a priori hypothesis, this type of data reduction will increase the sensitivity of the statistical test, because no (of fewer) corrections for multiple comparisons will be required. Thus, data reduction in the analysis stage can be clever and perfectly legitimate. However, it does not thereby also justify selective dissemination of only this aspect of the data. This becomes evident by considering that in both scenarios in Figure 1 the data selection (as marked by the highlighted data points in magenta and cyan) may have been based on the same independent localizer or a priori hypothesis. Rather, if an effect is predicted to be

structured in a particular way in space, frequency and time, then confirmations as well as violations of this prediction should be reported. This will also increase transparency and, in this regard, our plea parallels a recent plea for completeness in the description of the analyses of multi-variable behavioral data [9].

Evidently, the evaluation of plausibility is not based on a quantitative (thresholdable) aspect of the data, as is the case for a p-value. However, this does not make the current argument invalid or useless. In our view, p-values should be considered one (but preferable not the only) source by which genuine and random effects can be distinguished, and researchers should actively seek to complement them where possible. Here, we have emphasized one such complementary force: the inherent structure of genuine physiological phenomena.

In sum, physiological data offer the great opportunity to evaluate plausibility in addition to statistical tests, and establishing stricter norms for reporting such data provides a powerful avenue toward increasing reproducibility of research findings in cognitive neuroscience. Such norms include always evaluating and disseminating effects of interest with regard to all dimensions in which they unfold, and placing more emphasis on the predicted structuring of effects (and their confirmations as well as violations) in scientific reports.

Acknowledgements

FvE was supported by a Newton International Fellowship from the British Academy and the Royal Society.

Figure

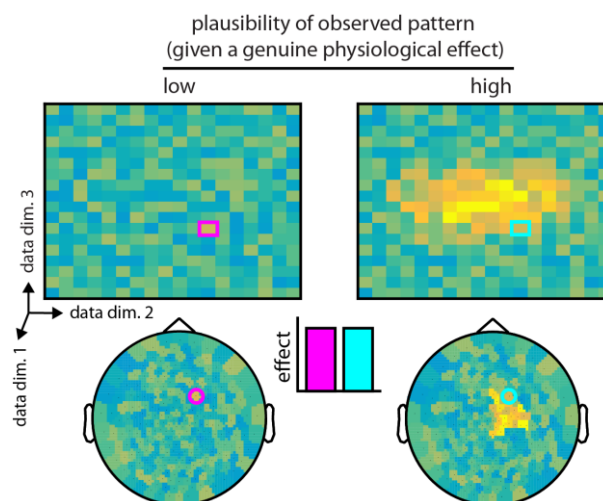


Figure 1. Two scenarios in which the same data point (highlighted in magenta and cyan) is embedded in a more or less plausible physiological pattern that unfolds over three dimensions (e.g. 1: space, 2: time, 3: frequency). When only provided with the bar graph, it is impossible to evaluate the plausibility of the pattern the lies “under the hood” and thereby to distinguish both scenarios.

Additional element

Box 1. Recommendations when dealing with high-dimensional data

- Always explore and disseminate the data with regard to all of its dimensions, even when statistical evaluation is (legitimately) restricted to specific aspects of the data.
- When there is no good justification for zooming in on a particular aspect of the data for statistical evaluation, employ statistical tests that capitalize on the structure of the data (as described in e.g. [8,10]).

- When dealing with multi-dimensional significant clusters, depict the relevant dimensions with regard to all samples of the cluster. For example, plot time-frequency profiles averaged over all sites of the cluster and, conversely, plot topographical maps averaged over all time-frequency samples of the cluster.
- Whenever possible, articulate the expected structuring of an effect and use this when evaluating and discussing the plausibility of the observed data. For example, when studying movement preparation, one expects the neural effects of interest to build up from the preparatory cue to the go signal.
- Employ dimensionality reduction techniques to extract components with profiles in all relevant dimensions of the data. As an example, see [12] where patterns of phase-amplitude coupling are decomposed into components with four profiles: one spatial and one spectral profile, both for the phase- and the amplitude-providing oscillations.

References

- [1] Simmons, J.P. et al. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psych. Sci.* 22, 1359-1366
- [2] Button, K.S. et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365-376
- [3] Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349, aac4716
- [4] Kriegeskorte, N. et al. (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535-540
- [5] Nosek, B.A. et al. (2012) Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psych. Sci.* 7, 615-631
- [6] Nuzzo T (2015) How scientists fool themselves – and how they can stop. *Nature* 526, 182-185
- [7] Genovese, C.R. et al. (2012) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* 15, 870-878
- [8] Maris, E. and Oostenveld, R. (2007) Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Meth.* 164, 177-190
- [9] Friston K.J. et al. (1996) Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage* 4, 223-235
- [10] Maris, E. (2012) Statistical testing in electrophysiological studies. *Psychophysiology* 49, 549–565
- [11] Wigboldus, D.H.J. and Dotsch, R. (2015) Encourage playing with data and discourage questionable reporting practices. *Psychometrika* 81, 1-6
- [12] van der Meij, R. et al. (2012) Phase-amplitude coupling in human electrocorticography is spatially distributed and phase diverse. *J. Neurosci.* 32, 111-123