META-RESEARCH

Lessons from a catalogue of 6674 brain recordings

Abstract It is now possible for scientists to publicly catalogue all the data they have ever collected on one phenomenon. For a decade, we have been measuring a brain response to visual symmetry called the sustained posterior negativity (SPN). Here we report how we have made a total of 6674 individual SPNs from 2215 participants publicly available, along with data extraction and visualization tools (https://osf.io/2sncj/). We also report how re-analysis of the SPN catalogue has shed light on aspects of the scientific process, such as statistical power and publication bias, and revealed new scientific insights.

ALEXIS DJ MAKIN*, JOHN TYSON-CARR, GIULIA RAMPONE, YIOVANNA DERPSCH, DAMIEN WRIGHT AND MARCO BERTAMINI

Introduction

Many natural and man-made objects are symmetrical, and humans can detect visual symmetry very efficiently (***Bertamini et al., 2018***; ***Treder, 2010***; ***Tyler, 1995***; ***Wagemans, 1995***). Visual symmetry has been a topic of research within experimental psychology for more than a century. In recent decades, two techniques – functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) – have been used to investigate the impact of visual symmetry on a region of the brain called the extrastriate cortex. Since 2011, we have been using EEG in experiments at the University of Liverpool to measure a brain response to visual symmetry called the sustained posterior negativity (SPN; see ***Box 1*** and ***Figure 1***). By October 2020 we had completed 40 SPN projects: 17 of these had been published, and the remaining 23 were either unpublished or under review.

The COVID pandemic stopped all our EEG testing in March 2020, and we used this crisis/opportunity to organize and catalogue our existing data. The data from all 40 of our SPN projects are now available in a public repository called "The complete Liverpool SPN catalogue" (available at https://osf.io/2sncj/; see ***Box 2*** and ***Figure 2***). The catalogue allows us to draw conclusions that could not be gleaned from a single experiment. It can also support meta-scientific evaluation of our data and practices, as reported in the current article.

Meta-scientific lessons from the complete Liverpool SPN catalogue

There is growing anxiety about the trustworthiness of published science (***Munafò et al., 2017***; ***Open Science Collaboration, 2015***; ***Errington et al., 2021***). Many have argued that we should build cumulative research programs, where effects are measured reliably, and the truth becomes clearer over time. However, common practice often falls far short of this ideal. And although there has been a positive response to the replication crisis in psychology (***Nelson et al., 2018***), there is still – according to the cognitive neuroscientist Dorothy Bishop – room for improvement: "many researchers persist in working in a way almost guaranteed not to deliver meaningful results. They ride what I refer to as the four horsemen of the irreproducibility apocalypse" (***Bishop, 2019***). The four horsemen are: (i) publication bias; (ii) low statistical power; (iii) p value hacking; (iv) HARKing (hypothesizing after results known).

The "manifesto for reproducible science" includes these four items and two more: poor quality control in data collection and analysis, and the failure to control for bias (***Munafò et al., 2017***). Such critiques challenge all scientists to answer a simple question: are you practicing cumulative science, or is your research is undermined by the four horsemen of the irreproducibility apocalypse? Indeed, before compiling the

***For correspondence:** alexis.makin@liverpool.ac.uk

Competing interest: The authors declare that no competing interests exist.

Funding: See page 18

Reviewing Editor: Peter Rodgers, eLife, United Kingdom

Box 1. Symmetry and the sustained posterior negativity (SPN).

Visual symmetry plays an important role in perceptual organization (**Koffka, 1935**; **Wagemans et al., 2012**) and mate choice (**Grammer et al., 2003**). This suggests sensitivity to visual symmetry is innate: however, symmetrical prototypes could also be learned from many asymmetrical exemplars (**Enquist and Johnstone, 1997**). Psychophysical experiments have taught us a great deal about symmetry perception (**Barlow and Reeves, 1979**; **Treder, 2010**; **Wagemans, 1995**), and the neural response to symmetry has been studied more recently (for reviews see **Bertamini and Makin, 2014**; **Bertamini et al., 2018**; **Cattaneo, 2017**). Functional MRI has reliably found symmetry activations in the extrastriate visual cortex (**Chen et al., 2007**; **Keefe et al., 2018**; **Kohler et al., 2016**; **Sasaki et al., 2005**; **Tyler et al., 2005**; **Van Meel et al., 2019**). The extrastriate symmetry response can also be measured with EEG. Visual symmetry generates an event related potential (ERP) called the sustained posterior negativity (SPN). The SPN is a difference wave – amplitude is more negative at posterior electrodes when participants view symmetrical displays compared to asymmetrical displays (**Jacobsen and Höfel, 2003**; **Makin et al., 2012**; **Makin et al., 2016**; **Norcia et al., 2002**). As shown in ***Figure 1***, SPN amplitude scales parametrically with the proportion of symmetry in the image (**Makin et al., 2020c**).

SPN catalogue, we were unable to answer this question for our own research program.

One problem with replication attempts is their potentially adversarial nature. Claiming that other people's published effects are unreliable insinuates bad practice, while solutions such as "adversarial collaboration" are still rare (***Cowan et al., 2020***). In this context and heeding the call to make research in psychology auditable (***Nelson et al., 2018***), we decided to take an exhaustive and critical look at the complete SPN catalogue in terms of the four horsemen of irreproducibility.

Horseman one: publication bias

Most scientists are familiar with the phrase "publish or perish" and know it is easier to publish a statistically significant effect (*P*<.05). Null results accumulate in the proverbial file drawer, while false positives enter the literature. This publication bias leads to systematic overestimation of effect sizes in meta-analysis (***Brysbaert, 2019***; ***Button et al., 2013***).

The cumulative distribution of the 249 SPN amplitudes is shown in ***Figure 3A*** (top panel), along with the cumulative distributions for those in the literature and those in the file drawer (middle panel). The unpublished SPNs were weaker than the published SPNs (mean difference = 0.354 microvolts [95% CI=0.162–0.546], t (218.003)=3.640, *P*<.001, equal variance not assumed). Furthermore, the published SPNs came from experiments with smaller sample sizes (mean sample sizes = 23.40 vs 29.49, *P*<.001, Mann-Whitney U test).

To further explore these effects, we ran three meta-analyses (using the metamean function from the dmetar library in R). Full results are described in supplementary materials (https://osf.io/q4jfw/). The weighted mean amplitude of the published SPNs was –1.138 microvolts [95% CI = –1.290; –0.986]. This was reduced to –0.801 microvolts [–0.914; –0.689] for the unpublished SPNs, and to –0.954 microvolts [–1.049; –0.860] for all SPNs. The funnel plot for the published SPNs (***Figure 3A***; bottom panel) is not symmetrically tapered (less accurate measures near the base of the funnel are skewed leftwards). This is a textbook fingerprint of publication bias. However, the funnel asymmetry was still significant for the unpublished SPNs and for all SPNs, so publication bias cannot be the explanation (see ***Zwetsloot et al., 2017*** for detailed analysis of funnel asymmetry).

The amplitudes of the P1 peak and the N1 trough from the same trials provide an instructive comparison. Our SPN papers do not need large P1 and N1 components, so these are unlikely to have a systematic effect on publication. P1 peak was essentially identical in published and unpublished work (4.672 vs 4.686, t (195.11) = –0.067, *P*=.946; ***Figure 3B***). This is potentially an interesting counterpoint to the SPN. However, the N1 trough was larger in published work (–8.736 vs. –7.155. t (183.61) = –5.636, *P*<.001; ***Figure 3C***).

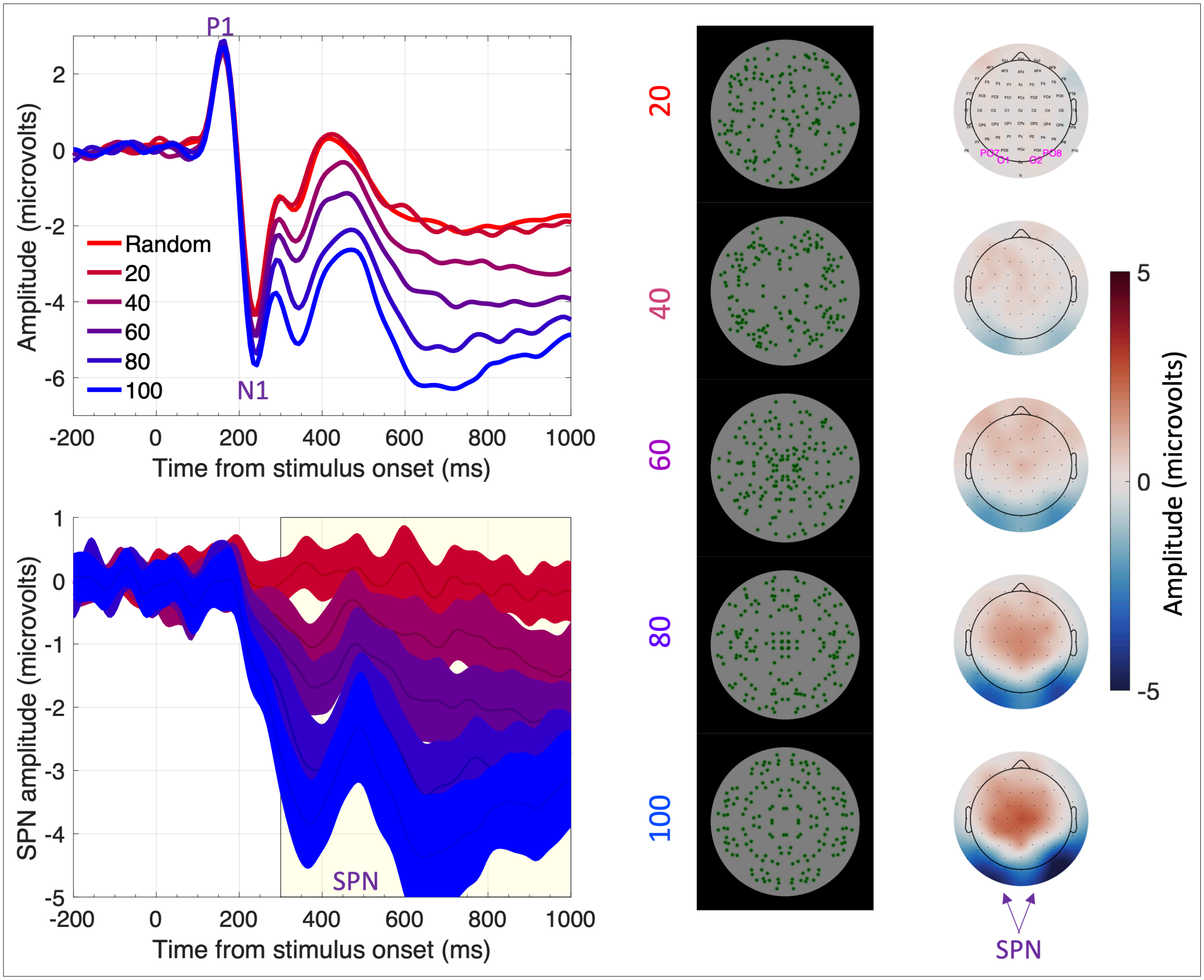


Figure 1. The sustained posterior negativity (SPN). The grand-average ERPs are shown in the upper left panel and difference waves (reflection-random) are shown in the lower left panel. A large SPN is a difference wave that falls a long way below zero. Topographic difference maps are shown on the right, aligned with the representative stimuli (black background). The difference maps depict a head from above, and the SPN appears as blue at the back. Purple labels indicate electrodes used for ERP waves [PO7, O1, O2 and PO8]. Note that SPN amplitude increases (that is, becomes more negative) with the proportion of symmetry in the image. In this experiment, the SPN increased from ~0 to –3.5 microvolts as symmetry increased from 20% to 100%. Adapted from Figures 1, 3 and 4 in ***Makin et al., 2020c***.

The funnel asymmetry was also significant for both P1 and N1.

Summary for horseman one: publication bias

This analysis suggests a tendency for large SPNs to make it into the literature, and small ones to linger in the file drawer. In contrast, the P1 was essentially identical in published and unpublished work. However, we do not think publication bias is a problem in our SPN research. The published and unpublished SPNs are from heterogenous studies, with different stimuli and tasks. We would not necessarily expect them to have the same mean amplitude. In other words, the SPN varies across studies not only due to neural noise and measurement noise, but also due to experimental manipulations affecting the neural signal (although W-load and Task were similar in published and unpublished work, see ***Box 2***). Furthermore, it is *not* the case that our published papers selectively report one side of a

Box 2. The complete Liverpool SPN catalogue.

The complete Liverpool SPN catalogue was compiled from 40 projects, each of which involved between 1 and 5 experiments (with an experiment being defined as a study producing a dataset composed only of within-subject conditions). Sample size ranged from 12 to 48 participants per experiment (mean = 26.37; mode = 24; median = 24). Each experiment provided 1–8 grand-average SPN waves. In total, we reanalysed 249 grand-average SPNs, 6,674 participant-level SPNs, and 850,312 single trials. SPNs from other labs are not yet included in the catalogue (***Höfel and Jacobsen, 2007a***; ***Höfel and Jacobsen, 2007b***; ***Jacobsen et al., 2018***; ***Kohler et al., 2018***; ***Martinovic et al., 2018***; ***Wright et al., 2018***). Steady-state visual evoked potential responses to symmetry are also unavailable (***Kohler et al., 2016***; ***Kohler and Clarke, 2021***; ***Norcia et al., 2002***; ***Oka et al., 2007***). However, the intention is to keep the catalogue open, and the design allows many contributions. In the future we hope to integrate data from other labs. This will increase the generalizability of our conclusions. Anyone wishing to use the catalogue can start with the beginner's guide, available on open science framework (https://osf.io/bq9ka/).

The catalogue also includes several supplementary files, including a file called "One SPN Gallery.pdf" (https://osf.io/eqhd5/) which has one page for each of the 249 SPNs, along with all technical information about the stimuli and analysis (***Figure 2*** shows the first SPN from the gallery). Browsing this gallery reveals that 39/40 projects used abstract stimuli, such as dot patterns or polygons (see, for example, Project 1: ***Makin et al., 2012***). The exception was Project 14, which used flowers and landscapes (***Makin et al., 2020b***). The SPN is generated automatically when symmetry is present in the image (e.g., Project 13: ***Makin et al., 2020c***). However, the brain can sometimes go beyond the image and recover symmetry in objects, irrespective of changes in view angle (Project 7: ***Makin et al., 2015***).

Almost half the SPNs (125/249) were recorded in experiments where participants were engaged in active regularity discrimination (e.g., press one key to report symmetry and another to report random). The other 124 SPNs were recorded in conditions where participants were performing a different task, such as discriminating the colour of the dots or holding information in visual working memory (***Derpsch et al., 2021***). In most projects the stimuli were presented for at least 1 second and the judgment was entered in a non-speeded fashion after stimulus offset. Key mapping was usually shown on the response screen to avoid lateralized preparatory motor responses during stimulus presentation.

The catalogue is designed to be FAIR (Findable, Accessible, Interoperable and Reusable). For each project we have included uniform data files from five subsequent stages of the pipeline: (i) raw BDF files; (ii) epoched data before ICA pruning; (iii) epoched data after ICA pruning; (iv) epoched data after ICA pruning and trial rejection; (v) pre-processed data averaged across trials for each participant and condition (stage v is the starting point for most ERP visualization and meta-analysis in this article). The catalogue also includes Brain Imaging Data Structure (BIDS) formatted files from stage iv (https://osf.io/e8r95/). BIDS files from earlier processing stages can be compiled from available codes or GUI (https://github.com/JohnTyCa/The-SPN-Catalogue) by users of MATLAB with EEGLAB and BIOSEMI toolbox.

Furthermore, we developed an app that allows users to: (a) view the data and summary statistics as they were originally published; (b) select data subsets, electrode clusters, and time windows; (c) visualize the patterns; (d) export data for further statistical analysis. This is available to Windows or Mac users with a Matlab license, and a standalone version can be used on Windows without a Matlab license. The app, executable and standalone scripts, and dependencies are available on Github (https://github.com/JohnTyCa/The-SPN-Catalogue, copy archived at swh:1:rev:75e729f867c275433b68807bc3f2228c57a3ccac, ***Tyson-Carr, 2022***). This repository and app will be maintained and expanded to accommodate data from future projects.

continued on next page

The folder called "SPN user guides and summary analysis" (https://osf.io/gjpr7/) also contains supplementary files that give all the technical details required for reproducible EEG research, as recommended by the Organization for Human Brain Mapping ***Pernet et al., 2020***. For instance, the file called "SPN effect size and power V8.xlsx" has one worksheet for each project (https://osf.io/c8jgy/). This file documents all extracted ERP data along with details about the electrodes, time windows, ICA components removed, and trials removed. With a few minor exceptions, anyone can now reproduce any figure or analysis in our SPN research. Users can also run alternative analyses that depart from the original pipeline at any given stage. Finally, the folder called "Analysis in eLife paper" contains all materials from this manuscript (https://osf.io/4cs2p/).

Although this paper focuses on meta-science, we can briefly summarize the scientific utility of the catalogue. Analysis of the whole data set shows that SPN amplitude scales with the salience of visual regularity. This can be estimated with the 'W-load' from theoretical models of perceptual goodness (***van der Helm and Leeuwenberg, 1996***). SPN amplitude also increases when regularity is task relevant. Linear regression with two predictors (W-load and Task, both coded on a 0–1 scale) explained 33% variance in grand-average SPN amplitude (SPN (microvolts) = –1.669 W – 0.416Task +0.071). The SPN is slightly stronger over the right hemisphere, but the laws of perceptual organization, that determine SPN amplitude, are similar on both sides of the brain. Source dipole analysis can also be applied to the whole data set (following findings of ***Tyson-Carr et al., 2021***). We envisage that most future papers will begin with meta-analysis of the SPN catalogue, before reporting a new purpose-built experiment. The SPN catalogue also allows meta-analysis of other ERPs, such as P1 or N1, which may be systematically influenced by stimulus properties (although apparently not W-load).

distribution with a mean close to zero. Our best estimate of mean SPN amplitude (–0.954 microvolts) is far below zero [95% CI = –1.049; –0.860]. The file drawer has no embarrassing preponderance of sustained posterior *positivity*.

Some theoretically important effects can appear robust in meta-analysis of published studies, but then disappear once the file drawer studies are incorporated. Fortunately, this does not apply to the SPN. We suggest that assessment of publication bias is a feasible first step for other researchers undertaking a catalogue-evaluate exercise.

Horseman two: low statistical power

According to ***Brysbaert, 2019***, many cognitive psychologists have overly sunny intuitions about power analysis and fail to understand it properly. The first misunderstanding is that an effect on the cusp of significance (*P*=.05) has a 95% chance of successful replication, when in fact the probability of successful replication is only 50% (power = 0.5). Researchers often work on the cusp of significance, where power is barely more than 0.5. Indeed, one influential analysis estimated that median statistical power in cognitive neuroscience is just 0.21 (***Button et al., 2013***). In stark contrast, the conventional threshold for adequate power is 0.8. Although things may be improving, many labs still conduct underpowered experiments without sufficient awareness of the problem.

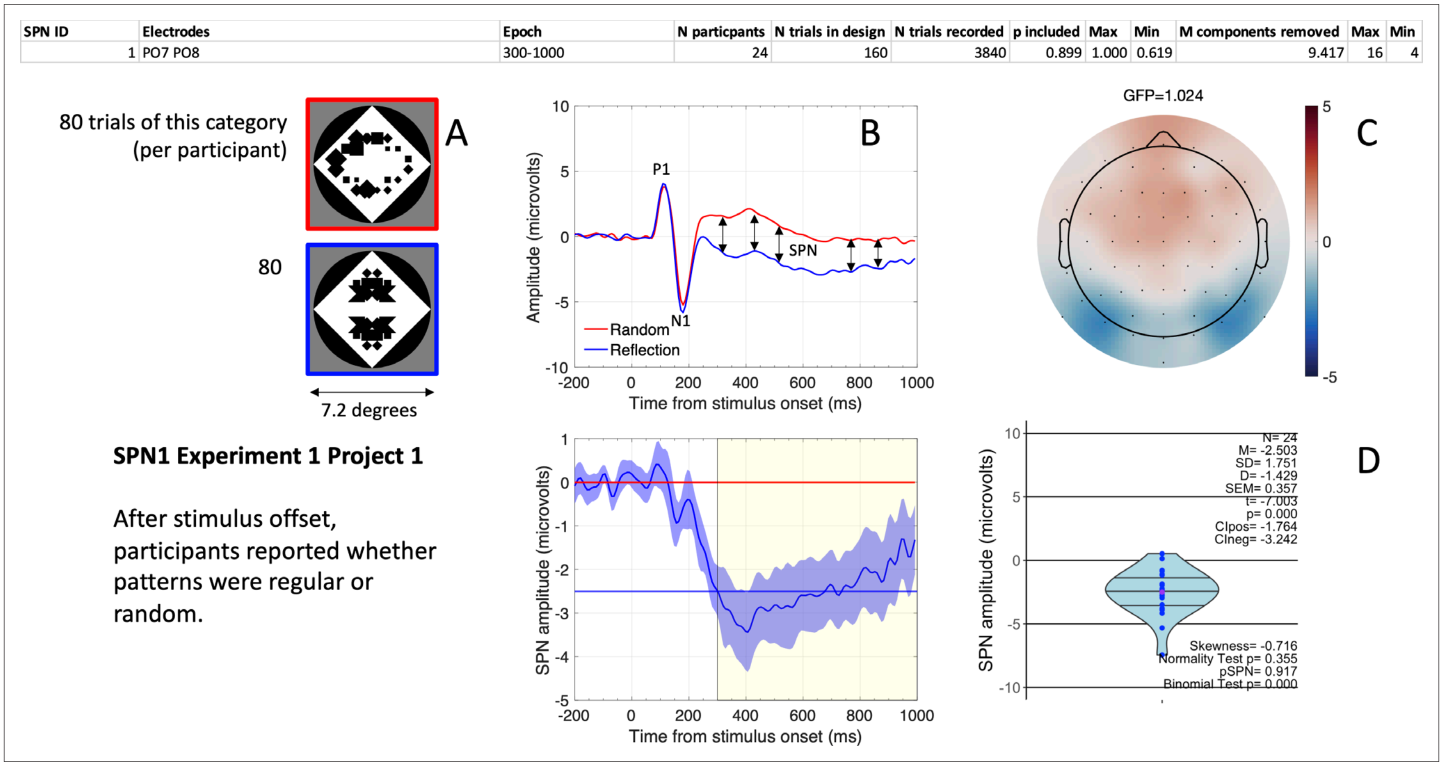
To estimate statistical power, one needs a reliable estimate of effect size, and this is rarely available a priori. In the words of ***Brysbaert, 2019***, "you need an estimate of effect size to get started, and it is very difficult to get a useful estimate". It is well known that effect size estimates from published work will be exaggerated because of the file drawer problem (as described previously). It is less well known that one pilot experiment does not provide a reliable estimate of effect size (especially when the pilot itself has a small sample; ***Albers and Lakens, 2018***). Fortunately, we can estimate SPN effect size from many experiments, published and unpublished, and this allows informative power analysis.

For a single SPN, the relevant effect size metric is Cohen's d_z (mean amplitude difference/SD of amplitude differences). ***Figure 4A*** shows the relationship between SPN amplitude and effect size d_z. The larger (more negative) the SPN in microvolts, the larger d_z. The curve tails off for strong SPNs, resulting in a nonlinear relationship. The second order polynomial trendline was a better fit than the first order linear trendline (see the supplementary polynomial regression analysis at https://osf.io/f2659/). The same relationship is found whether regularity is task relevant or not (***Figure 4B***) and in published and unpublished work (***Figure 4C***). Crucially, we can now estimate typical effect size for an SPN with a given amplitude using this polynomial regression equation (see the SPN effect size calculator at https://osf.io/gm734/).

This approach can be illustrated with 0.5 microvolt SPNs. Although these are at the low end of the distribution (***Figure 4A***), they can be interpreted and published (e.g., ***Makin et al., 2020a***). The average d_z for a 0.5 microvolt SPN is –0.469. Power analysis shows that to have an 80% chance of finding an effect of this size (*P*<.05, two tailed) we need a sample of 38 participants. In contrast, our median sample size is 24, which gives us an observed power of just 60%. In other words, if we were to choose a significant 0.5 microvolt SPN and rerun the exact same experiment, there is a 100%–60%=40% chance we would not find the significant SPN again. This is not a solid foundation for cumulative research.

Only a third of the 249 SPNs are 0.5 microvolts or less. However, many papers do not merely report the presence of a significant SPN. Instead, the headline effect is usually a within-subjects difference between experimental conditions. As a first approximation, we can assume the same power analysis applies to pairwise SPN modulations. We thus need 38 participants for an 80% chance of detecting an ~0.5 microvolt SPN difference between two regular conditions (and more participants for between-subject designs).

The table in ***Figure 4G*** gives required sample size (N) for 80% chance of obtaining SPNs of a particular amplitude (power = 0.8, alpha = 0.05, two-tailed). This suggests relatively large 1.5 microvolt SPNs could be obtained with just 9 participants. However, estimates of effect size are less precise at the high end (see the supplementary polynomial regression analysis at https://osf.io/f2659). A conservative yet feasible approach is to collect at least 20 participants even when confident of obtaining a large SPN or SPN modulation. Alternatively, researchers may require a sample that allows them to find the



SPN ID	Electrodes	Epoch	N particpants	N trials in design	N trials recorded	p included	Max	Min	M components removed	Max	Min
1	PO7 PO8	300-1000	24	160	3840	0.899	1.000	0.619	9.417	16	4

Figure 2. The first SPN from the SPN Gallery. (**A**) Examples of stimuli. (**B**). Grand-average ERP waves from electrodes PO7 and PO8 (upper panel), and the SPN as a reflection-random difference wave (with 95% CI; lower panel). The typical 300–1000ms SPN window is highlighted in yellow. Mean amplitude during this window was –2.503 microvolts (horizontal blue line). (**C**) SPN as a topographic difference map. (**D**) Violin plot showing SPN amplitude for each participant plus descriptive and inferential statistics. The file "One SPN Gallery.pdf" (https://osf.io/eqhd5/) contains a figure like this for all 249 SPNs. The analysis details shown at the top of the figure are also explained in this file.

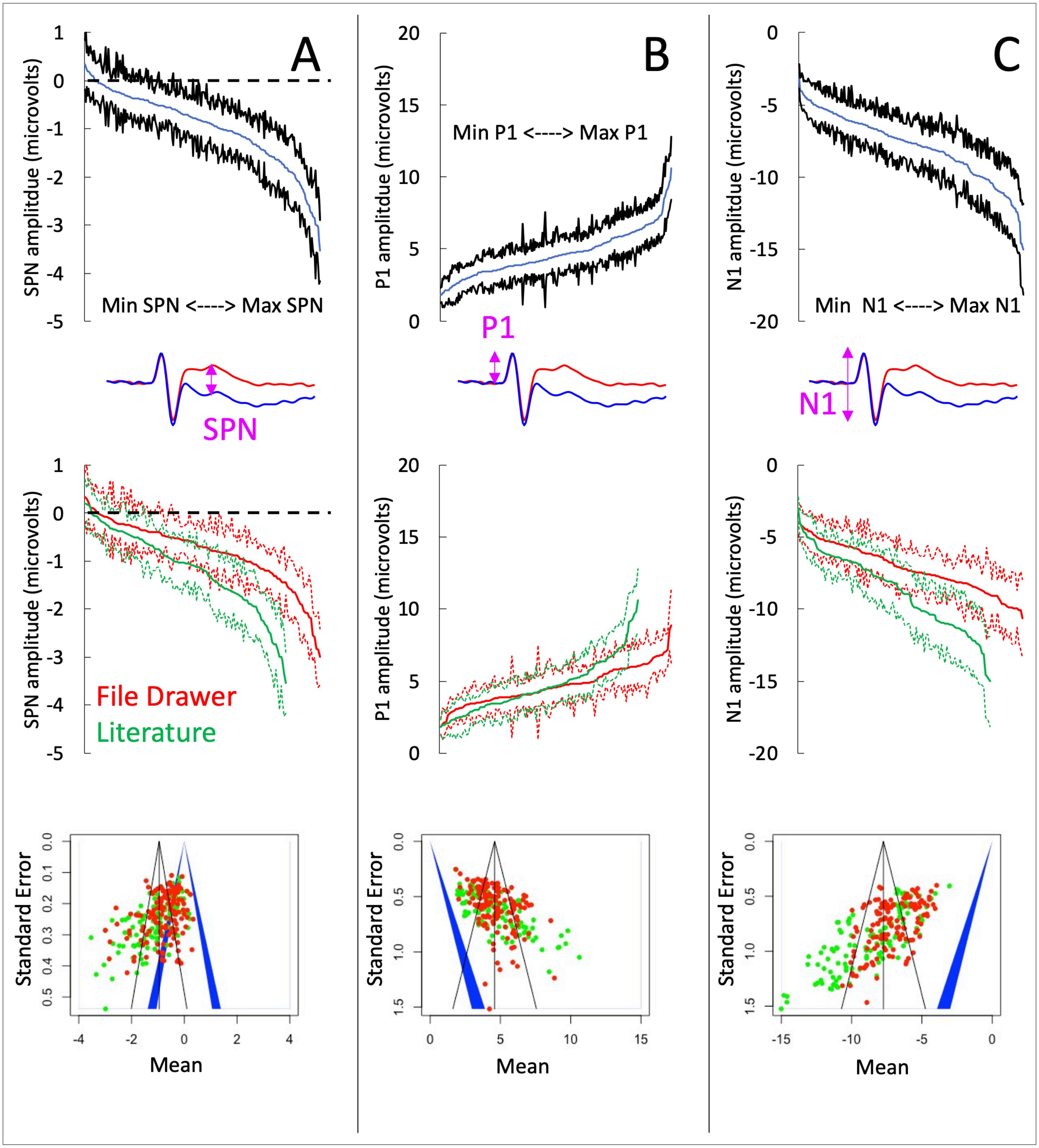


Figure 3. SPN amplitudes in published and unpublished work. (**A**) The top panel shows the cumulative distribution of all 249 grand-average SPNs. The smallest SPN is at the left-most end of the x-axis, and the largest SPN is at the right-most end. The blue line is comprised 249 data points, and the black lines show 95% confidence intervals. If the upper confidence interval does not rise above zero, we have a significant SPN (*P*<.05, two-tailed). The middle panel shows that the 134 unpublished SPNs in the file drawer (red) are smaller (i.e., less negative) than the 115 published SPNs in the literature (green). The bottom panel shows a funnel plot of 249 grand-average SPNs arranged by mean (x-axis) and standard error (y-axis). Red dots are unpublished SPNs,

Figure 3 continued on next page

Figure 3 continued

green dots are published SPNs. Dots to the left of the blue central triangle represent significant SPNs (inner edge, *P*<.05, outer edge, *P*<.01); if dots are inside the blue triangle, the effect is non-significant. (**B**) Equivalent set of plots for the peak amplitude P1 on regular trials. (**C**) Equivalent set of plots for the trough amplitude N1 on regular trials.

minimum effect that would still be of theoretical interest. ***Brysbaert, 2019***, suggests this may often be ~0.4 in experimental psychology, and this requires 52 participants. Indeed, the theoretically interesting effect of Task on SPN amplitude could be in this range.

Power of nonparametric tests

Of the 249 SPNs, 9.2% were not normally distributed about the mean according to the Shapiro-Wilk test (8.4% according to Kolmogorov-Smirnov test). Non-parametric statistics could thus be appropriate for some SPN analyses. For a non-parametric SPN, significantly more than half of the participants must have lower amplitude in the regular condition. We can examine this with a binomial test. Consider a typical 24 participant SPN: For a significant binomial test (*P*<.05, two tailed), we need at least 18/24=3/4 participants in the sample to show the directional effect (regular < random). Next, consider doubling sample size to 48: We now need only 32/48=2/3 participants in the sample to show the directional effect. ***Figure 4D–F*** illustrates the proportion of participants showing the directional effect as a function of SPN amplitude. Only 146 of the 249 grand-average SPNs (59%) were computed from a sample where at least 3/4 of the participants showed the directional effect. Meanwhile, 183 (73%) were from a sample where at least 2/3 of the participants showed the directional effect (blue horizontals in ***Figure 4D***). This analysis recommends increasing sample size to 48 in future experiments.

Power of SPN modulation effects

When there are more than two conditions, mean SPN differences may be tested with ANOVA. To assess statistical power, we reran 40 representative ANOVAs from published and unpublished work. This includes all those which support important theoretical conclusions (***Makin et al., 2015***; ***Makin et al., 2016***; ***Makin et al., 2020c***; see https://osf.io/hgncs/ for a full list of the experiments used in this analysis). Observed power was less than the desired 0.8 in 15 of 40 (***Figure 4H***). We note that several underpowered analyses are from Project 7 (***Makin et al., 2015***). This is still an active research area, and we will increase sample size in future experiments.

Increasing the number of trials

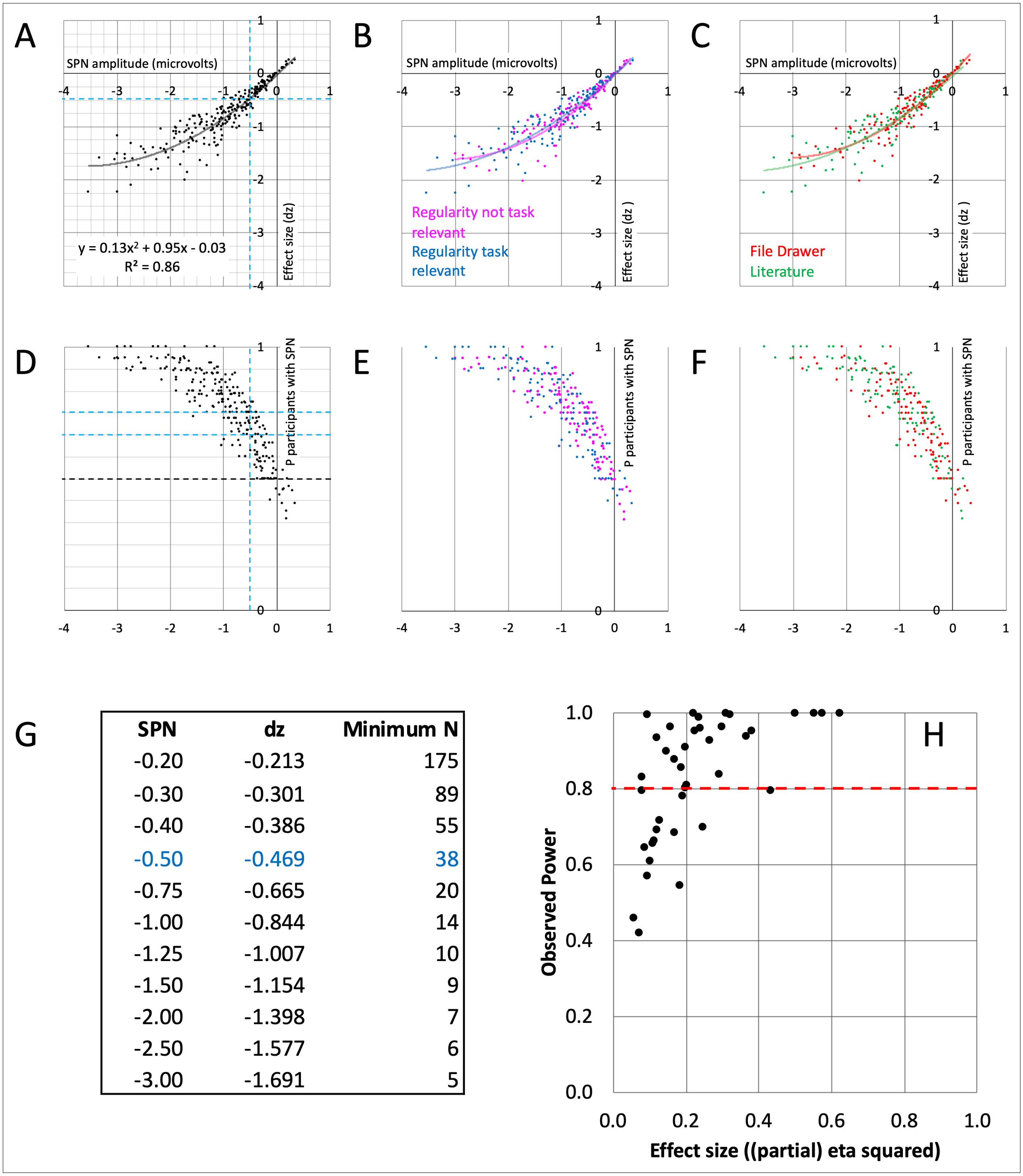
Another line of attack is to increase the number of trials per participant. ***Boudewyn et al., 2018*** argue that adding trials is alternative way to increase statistical power in ERP research, even when split-half reliability is apparently near ceiling (as it is in SPN research: ***Makin et al., 2020b***). In one highly relevant analysis, ***Boudewyn et al., 2018*** examined a within-participant 0.5 microvolt ERP modulation with a sample of 24 (our median sample). Increasing the number of trials from 45 to 90 increased the probability of achieving a significant effect from ~.54 to~.89 (see figure eight in ***Boudewyn et al., 2018***). These authors caution that simulations of other ERP components are required to establish generalizability. We typically include at least 60 trials in each condition. However, going up to 100 trials per condition could increase SPN effect size, and this may mitigate the need to increase sample size (***Baker et al., 2021***). Of course, too many trials could introduce participant fatigue and a consequent drop in data quality. There is likely a sample size X trial number 'sweet spot' and are unlikely to have hit it already by luck.

Typical sample sizes in other EEG research

When planning our experiments, we have often assumed that 24 is a typical sample size in EEG research. This can be checked objectively. We searched open-access EEG articles within the PubMed Central database using a text-mining algorithm. A total of 1,442 sample sizes were obtained. Mean sample size was 35 (±22.97) and a median was 28. The most commonly occurring sample size was 20. We also extracted sample sizes from 74 EEG datasets on the OpenNeuro BIDS compliant repository. The mean sample size was 39.34 (±38.56), the median was 24, and the mode was again 20. Our SPN experiments do indeed have typical sample sizes, as we had assumed.

Summary for horseman two: low statistical power

Low statistical power is an obstacle to cumulative SPN research. Before the COVID pandemic stopped all EEG research, we were completing



SPN	dz	Minimum N
-0.20	-0.213	175
-0.30	-0.301	89
-0.40	-0.386	55
-0.50	-0.469	38
-0.75	-0.665	20
-1.00	-0.844	14
-1.25	-1.007	10
-1.50	-1.154	9
-2.00	-1.398	7
-2.50	-1.577	6
-3.00	-1.691	5

Figure 4. SPN effect size and power. (**A–C**) The nonlinear relationship between SPN amplitude and effect size. The equation for the second order polynomial trendline is shown in panel A ($y = 0.13x^2+0.95\,x - 0.03$). This explains 86% of variance in effect size ($R^2=0.86$). Using the equation, we can estimate effect size for an SPN of a given amplitude. Dashed lines highlight –0.5 microvolt SPNs, with average effect size d_z of –0.469. For an 80% chance of finding this effect, an experiment requires 38 participants. The relationships were similar whether regularity was task relevant or not (**B**), and in

Figure 4 continued on next page

Figure 4 continued

published and unpublished work (**C**). (**D–F**) How many participants show the SPN? The larger (more negative) the SPN, the more individual participants show the effect (regular < random). Dashed lines highlight –0.5 microvolt SPNs, which are quite often present in 2/3 but not 3/4 of the participants. The relationships were similar whether regularity was task relevant or not (**E**), and in published and unpublished work (**F**). (**G**) Table of required N for 80% chance of obtaining an SPN of a given amplitude. (**H**) Observed power and effect size of 40 SPN modulations. 15/40 do not reach the 0.8 threshold (red line).

around 10 EEG experiments per year with a median sample of 24. When EEG research starts again, we may reprioritize, and complete fewer experiments per year with more participants in each. Furthermore, one can tentatively assume that other ERPs have comparable signal/noise properties to the SPN. If so, we can plausibly infer that many ERP experiments are underpowered for detecting 0.5 microvolt effects. ***Figure 4G*** thus provides a rough sample size guide for ERP researchers, although we stress that more ERP-specific estimates of effect size should always be treated as superior. We also stress that our sample size recommendations do not directly apply to multi-level or multivariate analyses, which are increasingly common in many research fields. Nevertheless, this investigation has strengthened our conviction that EEG researchers should collect larger samples by default. Several benefits more than make up for the extra time spent on data collection. Amongst other things, larger samples reduce Type 2 error, give more accurate estimates of effect size, and facilitate additional exploratory analyses on the same data sets.

Horseman three: P-hacking

Sensitivity of an effect to arbitrary analytical options is called the 'vibration of the effect' (***Button et al., 2013***). An effect that vibrates substantially is vulnerable to 'P-hacking': that is, exploiting flexibility in the analysis pipeline to nudge effects over the threshold for statistical significance (for example, *P*=.06 conveniently becomes *P*=.04, and the result is publishable). "Double dipping" is one particularly tempting type of P-hacking for cognitive neuroscientists because we typically have such large, multi-dimensional data sets (***Kriegeskorte et al., 2009***). Researchers can dip into a large dataset, observe where something is happening, then run statistical analysis on this data selection alone. For example, in one early SPN paper, ***Höfel and Jacobsen, 2007a*** state that "Time windows were chosen after inspection of difference waves" (page 25, section 2.8.3). It is commendable that Höfel and Jacobsen were so explicit: often double dipping is spread across months where researchers alternate between 'preliminary' data visualization and 'preliminary' statistical analysis so many times that they lose track of which came first. Double dipping beautifies results sections, but without appropriate correction, it inflates Type 1 error rate. We thus attempted to estimate the extent of P-hacking in our SPN research, with a special focus on double dipping.

Electrode choice

Post hoc electrode choice can sometimes be justified: Why analyse an electrode cluster that misses the ERP of interest? Post hoc electrode choice could be classed as a *questionable research practice* rather than flagrant malpractice (***Agnoli et al., 2017***; ***Fiedler and Schwarz, 2015***; ***John et al., 2012***). Nevertheless, we must at least assess the consequences of this flexibility. What would have happened if we had dogmatically stuck with the same electrodes for every analysis, without granting ourselves any flexibility at all?

To estimate this, we chose three a priori bilateral posterior electrode clusters and recomputed all 249 SPNs (Cluster 1 = [PO7 O1, O2 PO8], Cluster 2 = [PO7, PO8], Cluster 3 = [P1 P3 P5 P7 P9 PO7 PO3 O1, P2 P4 P6 P8 P10 PO8 PO4 O2]). The first two clusters were chosen because we have used them often in published research (Cluster 1: ***Makin et al., 2020c***; Cluster 2: ***Makin et al., 2016***). Cluster 3 was chosen because the 16 electrodes cover the whole bilateral posterior region. These three clusters are labelled on a typical SPN topoplot in ***Figure 5A***. Reassuringly, we found that SPN amplitude is highly correlated across clusters (Pearson's *r* ranged from .946 to .982, *P*<.001; ***Figure 5B***). This suggests vibration is low, and flexible electrode choice has not greatly influenced our SPN findings. In fact, mean SPN amplitude would have been slightly higher if we had used Cluster 2 for all projects. Average SPN amplitude was –0.98 microvolts [95% CI = –1.08 to –0.88] with the original cluster, –0.961 [–1.05; –0.87] microvolts for Cluster 1,–1.12 [–1.22; –1.01] microvolts for Cluster 2, and –0.61 [–0.66; –0.55] microvolts for Cluster 3.

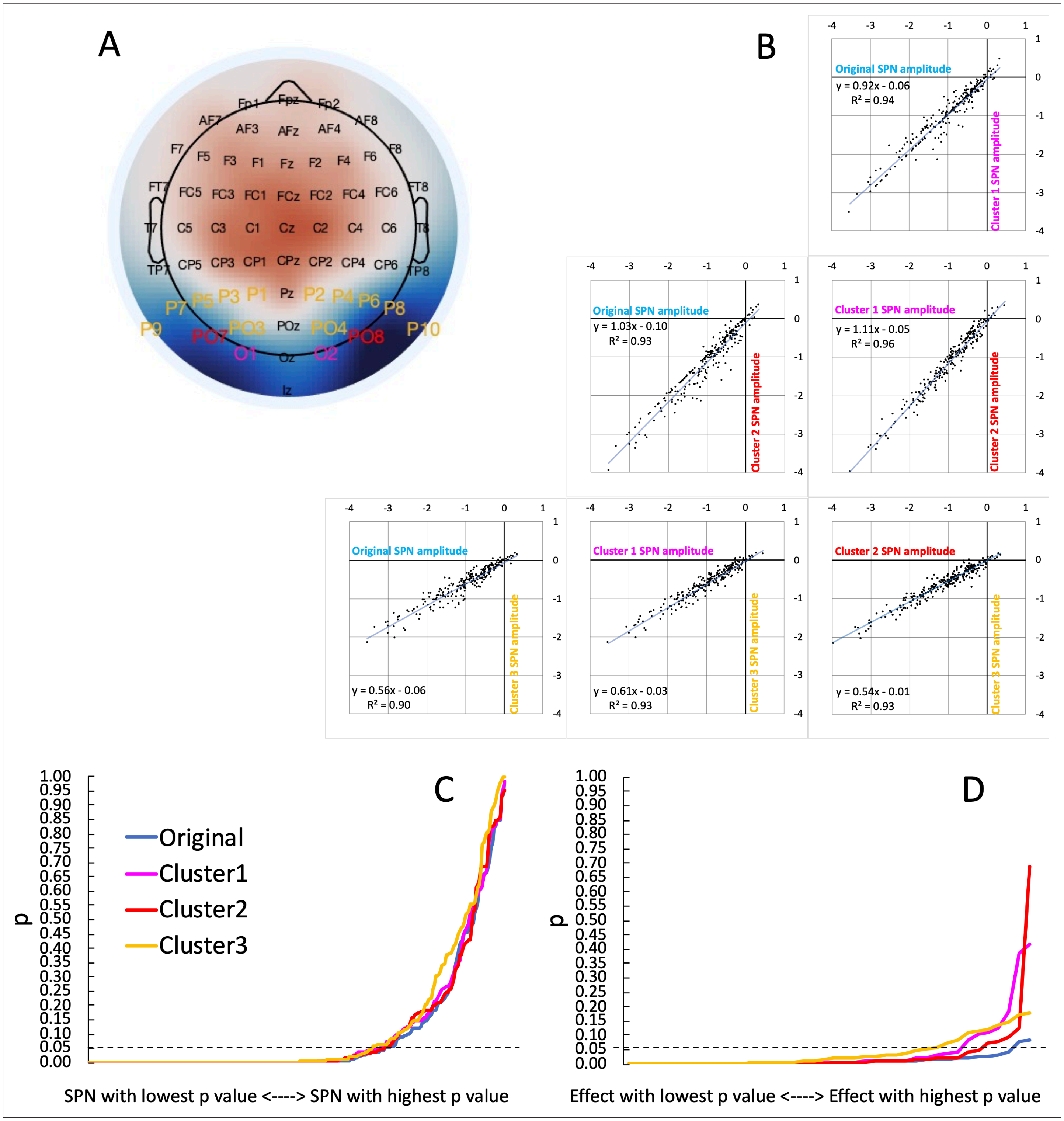
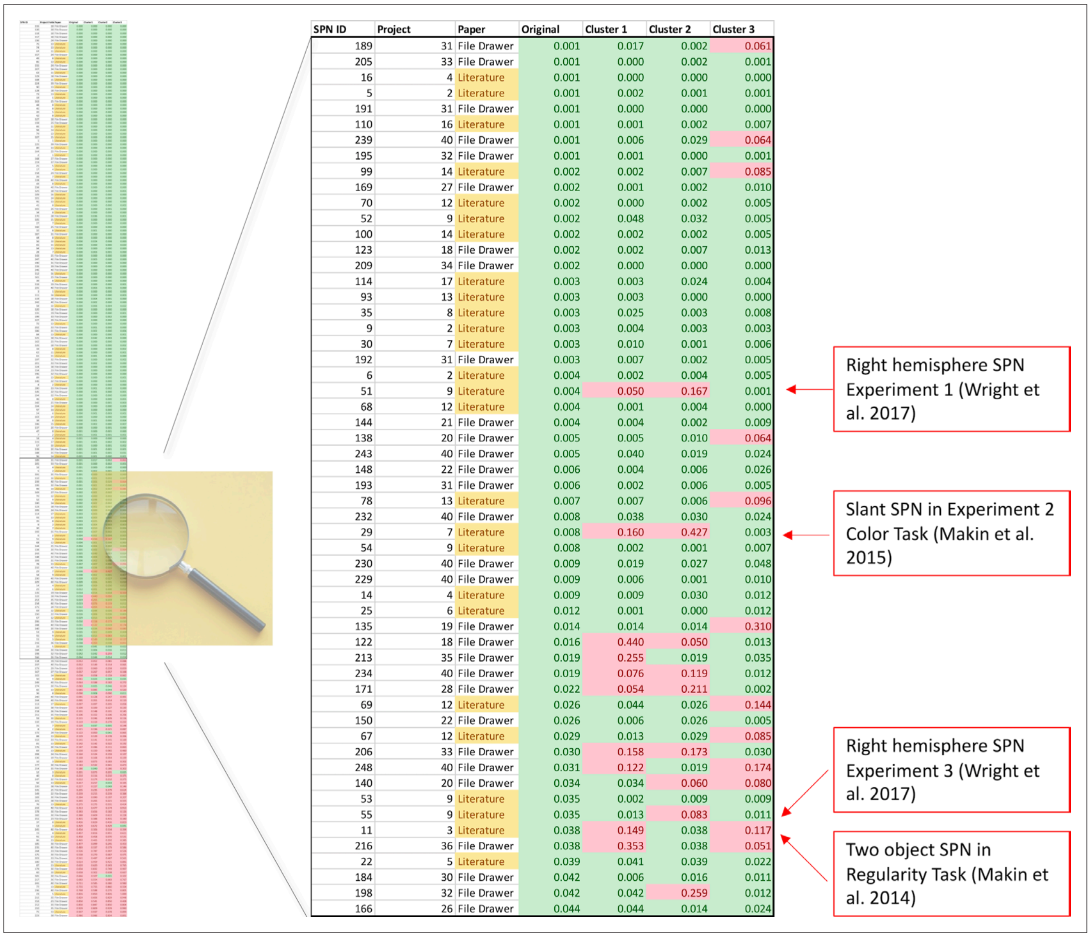


Figure 5. Vibration of the SPN effect. (**A**) Typical SPN topographic difference map with labels colour coded to show three alternative electrode clusters, which could have been used dogmatically in all analyses, whatever the observed topography. (**B**) Scatterplots show SPNs from the original cluster and the three alternatives, which are highly correlated. (**C**) One-sample t-tests were used to establish whether each SPN is significant. The cumulative distribution of p values is shown here. The smallest p value (from the most significant SPN) is at the left-most end of the x axis, and the largest p value (from the least significant SPN) is at the right-most end. The *p* values from the original cluster and the three alternatives were very similar. There was a similar number of significant SPNs (169–177). (**D**) ANOVAs are used to assess SPN modulations. The *p* values from 40 representative ANOVA effects do not overlap completely. There were more significant SPN modulations when the original electrode cluster was used than any alternative (38 vs 35–29).



SPN ID	Project	Paper	Original	Cluster 1	Cluster 2	Cluster 3
189	31	File Drawer	0.001	0.017	0.002	0.061
205	33	File Drawer	0.001	0.000	0.002	0.001
16	4	Literature	0.001	0.000	0.000	0.000
5	2	Literature	0.001	0.002	0.001	0.001
191	31	File Drawer	0.001	0.000	0.000	0.000
110	16	Literature	0.001	0.001	0.002	0.007
239	40	File Drawer	0.001	0.006	0.029	0.064
195	32	File Drawer	0.001	0.001	0.000	0.001
99	14	Literature	0.002	0.002	0.007	0.085
169	27	File Drawer	0.002	0.001	0.002	0.010
70	12	Literature	0.002	0.000	0.002	0.005
52	9	Literature	0.002	0.048	0.032	0.005
100	14	Literature	0.002	0.002	0.002	0.005
123	18	File Drawer	0.002	0.002	0.007	0.013
209	34	File Drawer	0.002	0.000	0.000	0.000
114	17	Literature	0.003	0.003	0.024	0.004
93	13	Literature	0.003	0.003	0.000	0.000
35	8	Literature	0.003	0.025	0.003	0.008
9	2	Literature	0.003	0.004	0.003	0.003
30	7	Literature	0.003	0.010	0.001	0.006
192	31	File Drawer	0.003	0.007	0.002	0.005
6	2	Literature	0.004	0.002	0.004	0.005
51	9	Literature	0.004	0.050	0.167	0.001
68	12	Literature	0.004	0.001	0.004	0.000
144	21	File Drawer	0.004	0.004	0.002	0.009
138	20	File Drawer	0.005	0.005	0.010	0.064
243	40	File Drawer	0.005	0.040	0.019	0.024
148	22	File Drawer	0.006	0.004	0.006	0.026
193	31	File Drawer	0.006	0.002	0.006	0.005
78	13	Literature	0.007	0.007	0.006	0.096
232	40	File Drawer	0.008	0.038	0.030	0.024
29	7	Literature	0.008	0.160	0.427	0.003
54	9	Literature	0.008	0.002	0.001	0.007
230	40	File Drawer	0.009	0.019	0.027	0.048
229	40	File Drawer	0.009	0.006	0.001	0.010
14	4	Literature	0.009	0.009	0.030	0.012
25	6	Literature	0.012	0.001	0.000	0.012
135	19	File Drawer	0.014	0.014	0.014	0.310
122	18	File Drawer	0.016	0.440	0.050	0.013
213	35	File Drawer	0.019	0.255	0.019	0.035
234	40	File Drawer	0.019	0.076	0.119	0.012
171	28	File Drawer	0.022	0.054	0.211	0.002
69	12	Literature	0.026	0.044	0.026	0.144
150	22	File Drawer	0.026	0.006	0.026	0.005
67	12	Literature	0.029	0.013	0.029	0.085
206	33	File Drawer	0.030	0.158	0.173	0.030
248	40	File Drawer	0.031	0.122	0.019	0.174
140	20	File Drawer	0.034	0.034	0.060	0.080
53	9	Literature	0.035	0.002	0.009	0.009
55	9	Literature	0.035	0.013	0.083	0.011
11	3	Literature	0.038	0.149	0.038	0.117
216	36	File Drawer	0.038	0.353	0.038	0.051
22	5	Literature	0.039	0.041	0.039	0.022
184	30	File Drawer	0.042	0.006	0.016	0.011
198	32	File Drawer	0.042	0.042	0.259	0.012
166	26	File Drawer	0.044	0.044	0.014	0.024

Right hemisphere SPN Experiment 1 (Wright et al. 2017)

Slant SPN in Experiment 2 Color Task (Makin et al. 2015)

Right hemisphere SPN Experiment 3 (Wright et al. 2017)

Two object SPN in Regularity Task (Makin et al. 2014)

Figure 6. Which SPNs are significant using alternative clusters? The left column shows a table of all 249 SPNs, colour coded (green, significant; red, non-significant), and sorted by *p* value obtained using the original cluster. The important part of the table is the centre, where significance thresholds are crossed by some clusters but not others. The central part is expanded, so text is now readable. The important cases are published SPNs that are not significant when either Cluster 1 or 2 is used instead. These 4 cases are all labelled (red boxes).

Next, we ran one-sample t-tests on the SPN as measured at the 4 alternative clusters. The resulting p values are shown cumulatively in ***Figure 5C***. Crucially, the area under the curve is similar for all cases. The significant SPN count varies only slightly, between 169 and 177 (χ^2 (3)=0.779, *P*=.854). We conclude that flexible electrode choice has not substantially inflated the number of significant SPNs in our research.

To illustrate this in another way, ***Figure 6*** shows a colour-coded table of *p* values, sorted by original cluster. At the top there are many rows which are significant whichever cluster is used (green rows). At the bottom there are many rows which are non-significant whichever cluster is used (red rows). The interesting rows are in the middle, where there is some disagreement indicating that the original effect was a false positive (some green and some red cells on each row). We can zoom in on the central portion: where, exactly, are the disagreements? Two cases come from ***Wright et al., 2017*** however, this project reported a contralateral SPN, and these are inevitably more sensitive to electrode choice because they only cover half the scalp surface area.

We applied the same reanalysis to our 40 representative ANOVA main effects and interactions. Here there is more cause for concern: 38 of the 40 effects were significant using the original electrode cluster, however this goes down to 33 with Cluster 1, 35 with Cluster 2, and to just 29 with Cluster 3 (***Figure 5D***). Flexible electrode choice has thus significantly increased the number of significant SPN modulations (χ^2 (3)=8.107, *P*=.044).

Spatio-temporal clustering

The above analysis examines consequences of choosing different electrode clusters a priori, while holding time window constant. Next, we

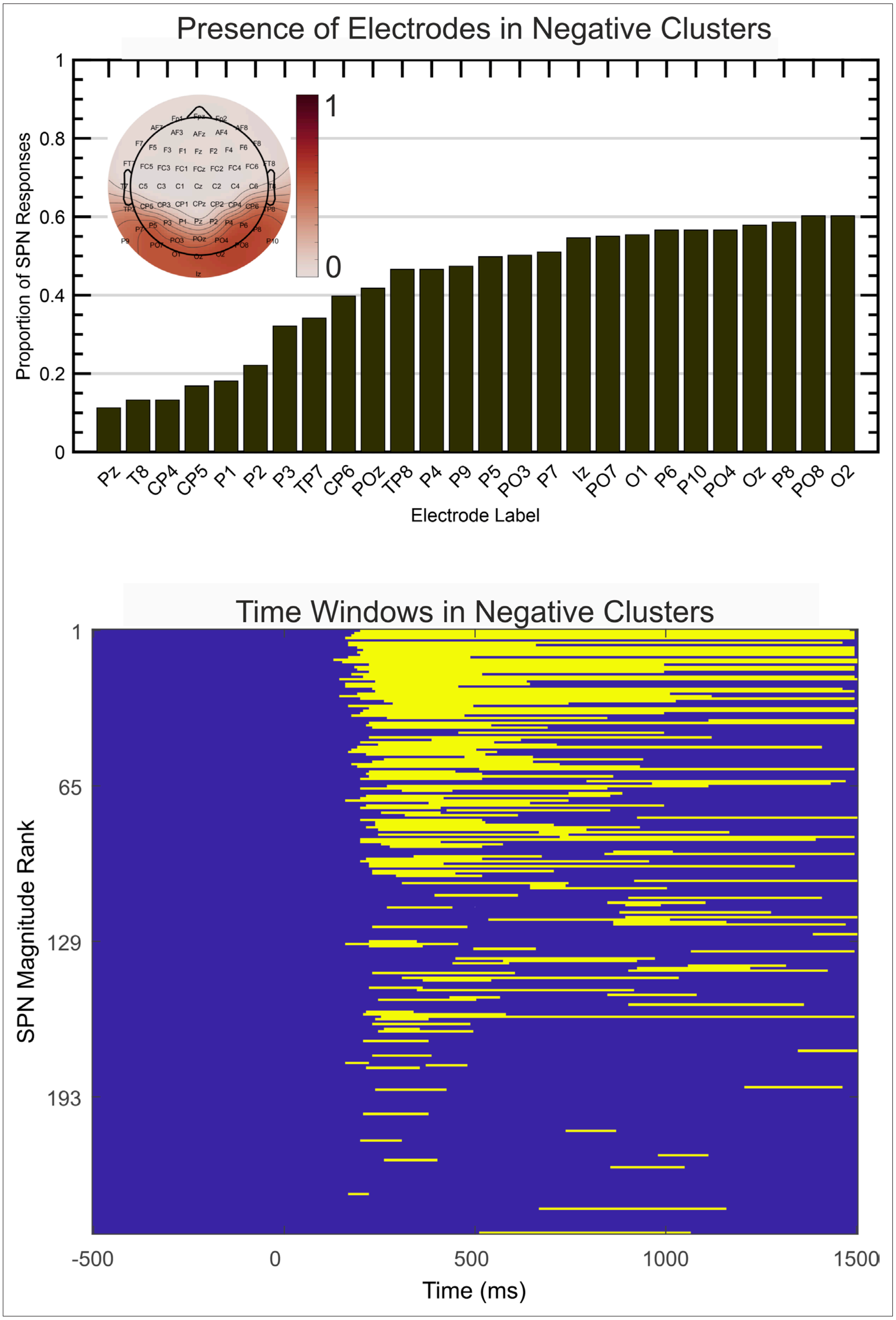


Figure 7. Spatio-temporal clustering results. The upper image illustrates the proportion of times each electrode appeared in the most significant negative cluster. Electrodes appearing in less than 10% of cases are excluded. The topoplot inset shows proportions on a colour scale for all electrodes. The lower image illustrates the time course over which the same negative clusters were active, ranked by SPN magnitude.

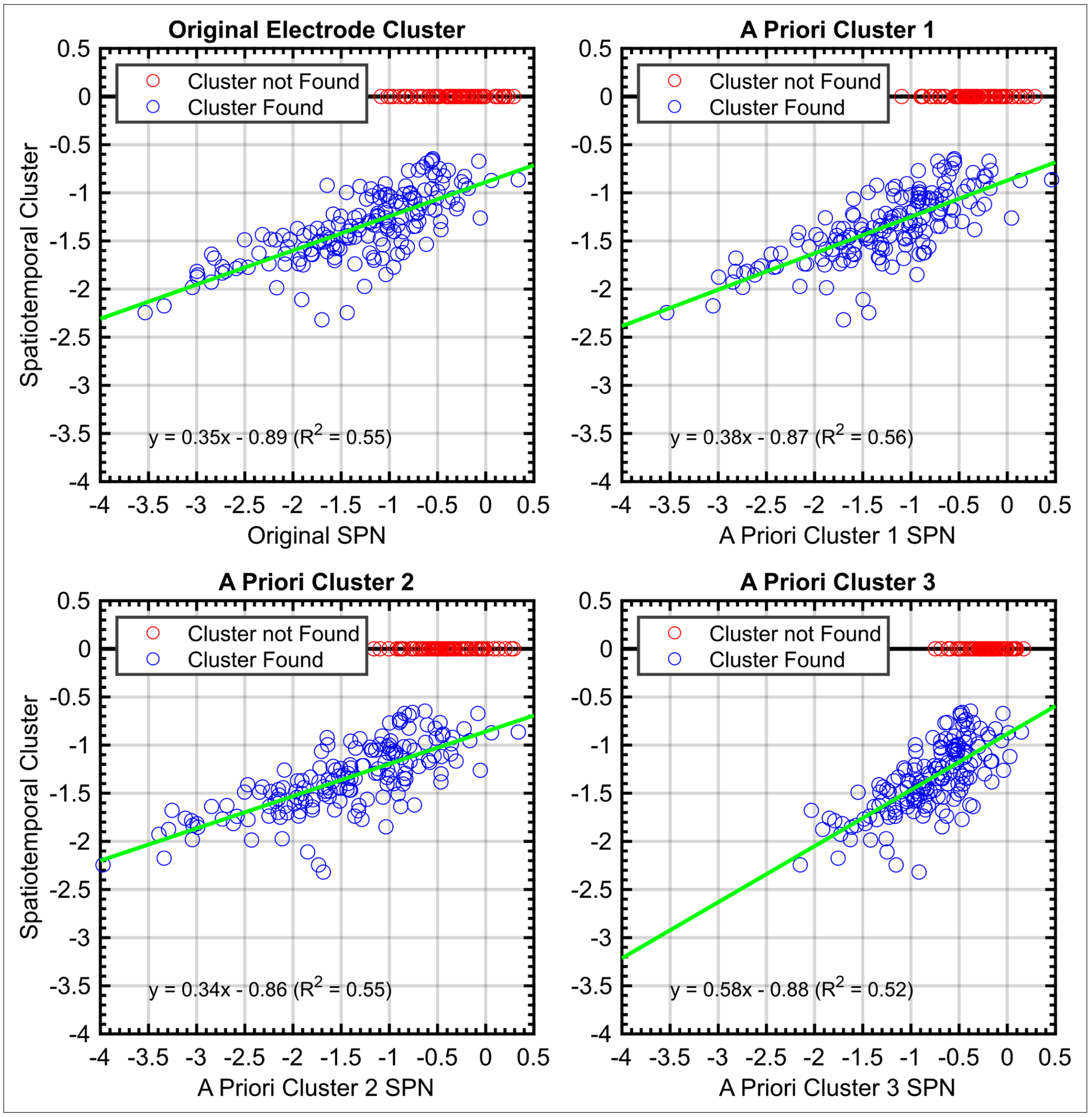


Figure 8. Correlations between SPNs from a priori and data-driven selections. Red data points indicate no significant negative cluster was found for that SPN. For these points, the mean SPN cluster is plotted as zero and does not influence the green least-squares regression line.

used a spatio-temporal clustering technique (***Maris and Oostenveld, 2007***) to identify both electrodes and timepoints where the difference between regular and irregular conditions is maximal. Does this purely data driven approach lead to the same conclusions as the original analysis from a priori electrode clusters?

After obtaining all negative electrode clusters with significant effect ($P<.05$, two tailed) the single most significant cluster was extracted for each SPN. The proportion of times that each electrode appeared in this cluster is illustrated in ***Figure 7A*** (electrodes present in less than 10% of cases are excluded). Findings indicate that electrodes O1 and PO7 are most likely to capture the SPN over the left hemisphere, and O2 and PO8 are most likely to capture the SPN over the right hemisphere. This is consistent with our typical a priori electrode selections. ***Figure 7B*** shows activity was mostly confined to the 200 to 1000ms window, extending somewhat to 1500ms. This is consistent with our typical a priori time window selections. Apparently there were no effects at electrodes or at time windows we had previously neglected.

To quantify these consistencies, SPNs were recomputed using the electrode cluster and time window obtained with spatiotemporal cluster analysis. There was a strong correlation between this and SPN from each a priori cluster (Pearson's *r* ranged from .719 to .746, $P<.001$; ***Figure 8***).

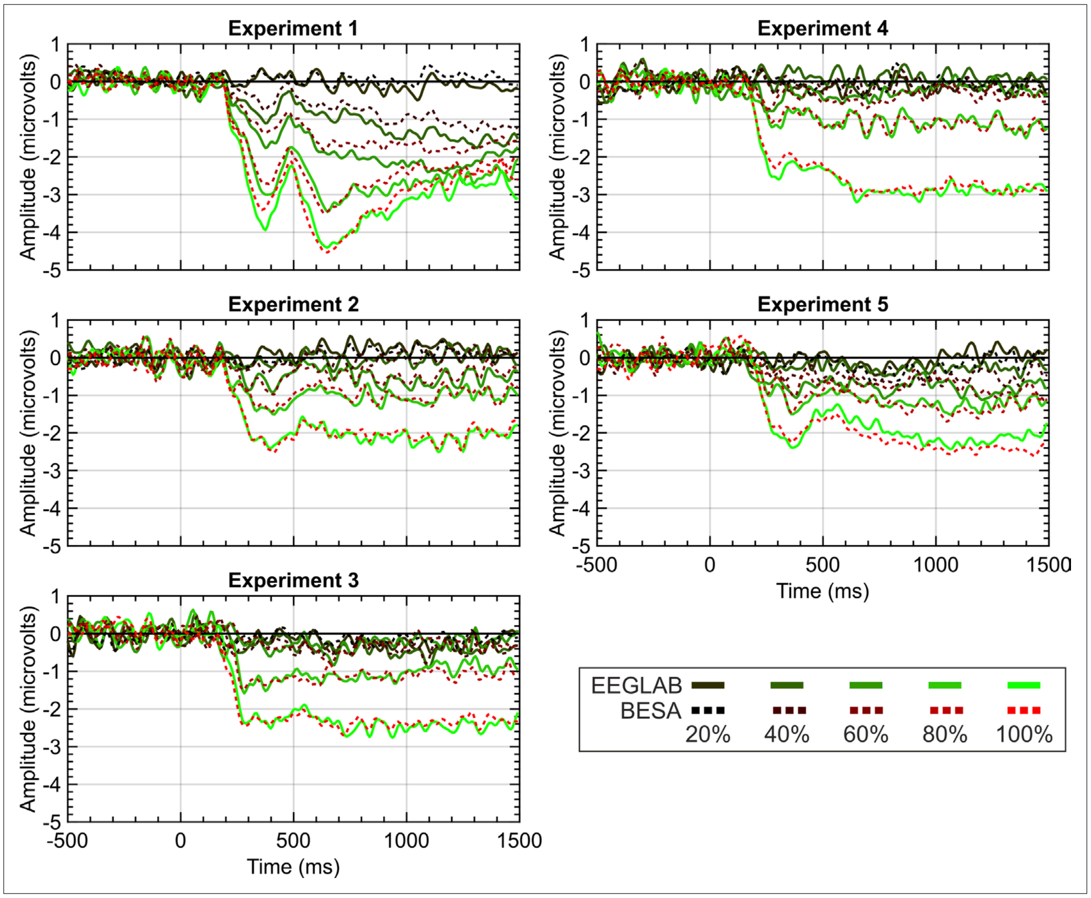


Figure 9. EEGLAB and BESA pipeline comparison. Panels show SPN waves for each of the 5 experiments in project 13. EEGLAB waves are the solid lines; the BESA waves are the dashed lines. 20%–100% refers to the proportion of symmetry in the stimulus (see ***Figure 1*** for example stimuli).

There are two further noteworthy results from spatiotemporal clustering analysis. First, 74% of published data sets yielded a significant cluster, while the figure was only 57% for unpublished data (as of May 2022). This is another estimate of publication bias. Second, we can explain some of the amplitude variation shown in ***Figure 8***. As described in ***Box 2***, 33% of variance in grand-average SPN amplitude can be predicted by two factors called W and Task (SPN (microvolts) = –1.669 W – 0.416Task + 0.071). We also ran this regression analysis on the spatiotemporal cluster SPNs (i.e., the blue data points in ***Figure 8***). The two predictors now explained just 16.8% of variance in grand-average SPN amplitude (SPN (microvolts) = –0.557 W – 0.170Task – 0.939). The reduced R^2, shallower slopes and lower intercept are largely caused by the fact that many data sets had to be excluded because they did not yield a cluster (i.e., the red data points in ***Figure 8***). This highlights one disadvantage of spatiotemporal clustering: For many purposes we want to include small and non-significant SPNs in pooled analysis of the whole catalogue. However, these relevant data points are missing or artificially set at zero if SPNs are extracted by spatiotemporal clustering.

Pre-processing pipelines

There are many pre-processing stages in EEG analysis (***Pernet et al., 2020***). For example, our data is often re-referenced to a scalp average, low pass filtered at 25 Hz, down-sampled to 128 Hz and baseline corrected using a –200 to 0ms pre-stimulus interval. We then remove some blink and other large artifacts with independent components analysis (***Jung et al., 2000***). We sometimes remove noisy electrodes with spherical interpolation. We then exclude trials where amplitude exceeds +/-100 microvolts at any electrode.

To examine whether these pre-processing norms are consequential, we reanalysed data from 5 experiments from Project 13 in BESA instead of Matlab and EEGLAB, using different cleaning and artifact removal conventions. When using BESA, we employed the recommended pipelines and parameters. Specifically, we used a template matching algorithm to identify eyeblinks and used spatial filtering to correct eyeblinks within the continuous data. Trials were removed that exceeded an amplitude of ±120 microvolts or a gradient of ±75 (with the gradient being defined as the difference in amplitude between two neighbouring samples). Although this trial exclusion takes place on filtered data, remaining trials are averaged across pre-filtered data. High and low pass filters were set to 0.1 and 25 Hz in both EEGLAB and BESA. While EEGLAB used zero-phase filters, filtering in BESA used a forward-filter for the high-pass filter but used a zero-phase filter for the low-pass. As seen in ***Figure 9***, similar grand-average SPNs fall out at the end of these disparate pipelines.

We conclude that post-hoc selection of electrodes and time windows are weak points that can be exploited by unscrupulous P-hackers. Earlier points in the pipeline are less susceptible to P-hacking because changing them does not predictably increase or decrease the desired ERP effect.

Summary for horseman three: P-hacking

It is easier to publish a simple story with a beautiful procession of significant effects. This stark reality may have swayed our practice in subtle ways. However, our assessment is that P-hacking is not a pervasive problem in SPN research, although some effects rely too heavily on post hoc data selection.

Reassuringly, we found most effects could be recreated with a variety of justifiable analysis pipelines: Data-driven and a priori approaches gave similar results, as did different EEGLAB and BESA pipelines. This was not a foregone conclusion – related analysis in fMRI has revealed troubling inconsistencies (***Lindquist, 2020***; ***Botvinik-Nezer et al., 2020***). It is advisable that all researchers compare multiple analysis pipelines to assess vibration of effects and calibrate their confidence accordingly.

Horseman four: HARKing

Hypothesizing After Results Known or HARKing (***Kerr, 1998***; ***Rubin, 2017***), is a potential problem in EEG research. Specifically, it is possible to conduct an exploratory analysis, find something unexpected, then describe the results *as if* they were predicted a priori. At worst, a combination of P-hacking and HARKing can turn noise into theory, which is then cited in the literature for decades. Even without overt HARKing, one can *beautify* an introduction section after the results are known, so that papers present a simple narrative (maybe this post hoc beautification could be called BARKing). Unlike the other horses, HARKing and BARKing are difficult to diagnose with reanalysis, which is why this section is shorter than the previous sections.

The main tool to fight HARKing is online pre-registration of hypotheses (for example, on aspredicted.org or osf.io). We have started using pre-registration routinely in the last four years, but we could have done so earlier. An even stricter approach is to use registered reports, where the introductions and methods are peer reviewed before data collection. This can abolish both HARKing and BARKing (***Chambers, 2013***; ***Munafò et al., 2017***), but we have only just started with this. Our recommendation is to use heavy pre-registration to combat HARKing. Perhaps unregistered EEG experiments will be considered unpublishable in the not-too-distant future.

Discussion

One of the most worrisome aspects of the replication crisis is that problems might be systemic, and not caused by a few corrupt individuals. It seems that the average researcher publishes somewhat biased research, without sufficient self-awareness. So, what did we find when we looked at our own research?

As regards publication bias – the first horseman of irreproducibility – we found that the 115 published SPNs were slightly stronger than the 134 unpublished ones. However, we are confident that there is no strong file drawer problem here. Even the unpublished SPNs are in the right direction (regular < random not random < regular). Furthermore, a complete SPN catalogue itself fights the consequences of publication bias by placing everything that was in the file drawer into the public domain.

We are more troubled by the second horseman: low statistical power. Our most negative conclusion is that reliable SPN research programs require larger samples than those we typically obtain (38 participants are required to reliably measure –0.5 microvolt SPNs and our median sample size is 24). This analysis has lessons for all researchers: It is evidently possible to 'get by' while routinely conducting underpowered experiments. One never notices a glaring problem: after all, underpowered research will often yield a lucky experiment with significant results, and this may support a new publication before moving on to the next topic. However, this common practice is not a strong foundation for cumulative research.

The costs of underpowered research might be masked by the third horseman: P-hacking. Researchers often exploit flexibility in the analysis pipeline to make borderline effects appear significant (**Simmons et al., 2011**). In EEG research, this often involves post-hoc selection of electrodes and time windows. Although some post-hoc adjustment is arguably appropriate, this double dipping certainly inflates false positive rate and requires scrutiny. We found that the same basic story would have emerged if we had rigidly used the same a priori electrode clusters in all projects or used a spatio-temporal clustering algorithm for selection. However, some of our SPN modulations were not so robust, and we have relied on post hoc time windows.

The fourth horseman, HARKing, is the most difficult dimension to evaluate because it cannot be diagnosed with reanalysis. Nevertheless, preregistration is the best anti-HARKing tool, and we could have engaged with this earlier. We are just beginning with pre-registered reports.

To summarize these evaluations, we would tentatively self-award a grade of A- for publication bias (75%, or lower first class, in the UK system), C+ for statistical power (58%), B+ for P-Hacking (68%), and B for Harking (65%). While some readers may not be interested in the validity of SPN research per se, they may be interested in this meta-scientific exercise, and we would encourage other groups to perform similar exercises on their own data. In fact, such exercises may be essential for cumulative science. It has been argued that research should be auditable (**Nelson et al., 2018**), but Researcher A will rarely be motivated to audit a repository uploaded by Researcher B, even if the datasets are FAIR. To fight the replication crisis, we must actively look in the mirror, not just passively let others snoop through the window.

Klapwijk et al., 2021 have performed a similar meta-scientific evaluation in the field of developmental neuroimaging and made many practical recommendations. All our themes are evident in their article. We also draw attention to international efforts to estimate the replicability of influential EEG experiments (**Pavlov et al., 2020**). These mass replication projects provide a broad overview. Here we provide depth, by examining all the data and practices from one representative EEG lab. We see these approaches as complementary: they both provide insight into whether a field is working in a way that generates meaningful results.

The focus on a single lab inevitably introduces some biases and limitations. Other labs may use different EEG apparatus with more channels. This would inevitably have some effect on the recordings. More subtle differences may also matter: For instance, other labs may use more practice trials or put more stress on the importance of blink suppression. However, the heterogeneity of studies and pipelines explored here ensures reasonable generalizability. We are confident that our conclusions are relevant for SPN researchers with different apparatus and conventions.

Curated databases are an extremely valuable resource, even if they are not used for meta-scientific evaluation. Public catalogues are an example of large-scale neuroscience, the benefits of which have been summarized by the neuroscientist Jeremy Freeman as follows: "Understanding the brain has always been a shared endeavour. But thus far, most efforts have remained individuated: labs pursuing independent research goals, slowly disseminating information via journal publications, and when analyzing their data, repeatedly reinventing the wheel" (**Freeman, 2015**). We tried to make some headway here with the SPN catalogue.

Perhaps future researchers will see their role as akin to expert museum curators, who oversee and update their public catalogues. They will obsessively tidy and perfect the analysis scripts, databases and metafiles. They will add new project folders every year, and judiciously determine when previous theories are no longer tenable. Of course, many researchers already dump raw data in online repositories, but this is not so useful. Instead, we need FAIR archives which are actively maintained, organized, and promoted by curators. The development of software, tools and shared repositories within the open science movement is making this feasible for most labs. We are grateful to everyone who is contributing to this enterprise.

It took more than a year to find all the SPN data, organize it, reformat it, produce uniform scripts, conduct rigorous double checks, and upload material to public databases. However, we anticipate that it will save far more than a year in terms of improved research efficiency. It is also satisfying that unpublished data sets are not merely lost. Instead, they are now contributing to more reliable estimates of SPN effect size and power. It is unlikely that any alternative activity could have been more beneficial for SPN research.

Acknowledgements

Many people contributed to the complete Liverpool SPN catalogue. One notable past contributor was Dr Letizia Palumbo (who was a postdoc 2013–2016). Many undergraduate, postgraduate and intern students have also been involved in data collection.

Alexis DJ Makin is in the Department of Psychological Sciences, University of Liverpool, Liverpool, United Kingdom
alexis.makin@liverpool.ac.uk
0000-0002-4490-7400

John Tyson-Carr is in the Department of Psychological Sciences, University of Liverpool, Liverpool, United Kingdom
0000-0003-3364-2184

Giulia Rampone is in the Department of Psychological Sciences, University of Liverpool, Liverpool, United Kingdom
0000-0002-2710-688X

Yiovanna Derpsch is in the Department of Psychological Sciences, University of Liverpool, Liverpool, and the School of Psychology, University of East Anglia, Norwich, United Kingdom

Damien Wright is in the Patrick Wild Centre, Division of Psychiatry, Royal Edinburgh Hospital, University of Edinburgh, Edinburgh, United Kingdom
0000-0002-9105-3559

Marco Bertamini is in the Department of Psychological Sciences, University of Liverpool, Liverpool, United Kingdom, and the Dipartimento di Psicologia Generale, Università di Padova, Padova, Italy
0000-0001-8617-6864

Author contributions: Alexis DJ Makin, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing; John Tyson-Carr, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – review and editing; Giulia Rampone, Investigation, Project administration, Resources, Writing – review and editing; Yiovanna Derpsch, Investigation, Resources, Writing – review and editing; Damien Wright, Investigation, Resources, Writing – review and editing; Marco Bertamini, Conceptualization, Investigation, Project administration, Resources, Supervision, Writing – original draft, Writing – review and editing

Competing interests: The authors declare that no competing interests exist.

Received 08 January 2021
Accepted 14 April 2022
Published 15 June 2022

Ethics: This was meta-analysis of previously published work (and unpublished work). All datasets were electrophysiological recordings taken from human participants. All experiments had local ethics committee approval and were conducted according to the Declaration of Helsinki (Revised 2008). Therefore all participants gave informed consent.

Funding

Funder	Grant reference number	Author
Economic and Social Research Council	ES/S014691/1	Alexis DJ Makin John Tyson-Carr Giulia Rampone Marco Bertamini

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.66388.sa1
Author response https://doi.org/10.7554/eLife.66388.sa2

Additional files

Supplementary files

- Transparent reporting form

Data availability

All data that supports analysis and Figures, along with codes for analysis, are available on open science framework (https://osf.io/2sncj/). To make it possible for anybody to analyze our data, we developed an app that allows users to: (i) view the data and summary statistics as they were originally published; (ii) select data subsets, electrode clusters, and time windows; (iii) visualize the patterns; (iv) export data for further statistical analysis. This repository and app will be expanded to accommodate data from future projects. The app is available to download for Windows users at https://github.com/JohnTyCa/The-SPN-Catalogue (copy archived at swh:1:rev:75e729f867c275433b68807bc3f2228c57a3ccac).

The following dataset was generated:

Author(s)	Year	Dataset URL	Database and Identifier
Makin A	2021	https://osf.io/2sncj/	Open Science Framework, 2sncj

References

Agnoli F, Wicherts JM, Veldkamp CLS, Albiero P, Cubelli R, Pietschnig J. 2017. Questionable research practices among italian research psychologists. *PLOS ONE* **12**:e0172792. DOI: https://doi.org/10.1371/journal.pone.0172792, PMID: 28296929

Albers C, Lakens D. 2018. When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology* **74**:187–195. DOI: https://doi.org/10.1016/j.jesp.2017.09.004

Baker DH, Vilidaite G, Lygo FA, Smith AK, Flack TR, Gouws AD, Andrews TJ. 2021. Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods* **26**:295–314. DOI: https://doi.org/10.1037/met0000337, PMID: 32673043

Barlow HB, Reeves BC. 1979. The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research* **19**:783–793. DOI: https://doi.org/10.1016/0042-6989(79)90154-8, PMID: 483597

Bertamini M, Makin ADJ. 2014. Brain activity in response to visual symmetry. *Symmetry* **6**:975–996. DOI: https://doi.org/10.3390/sym6040975

Bertamini M, Silvanto J, Norcia AM, Makin ADJ, Wagemans J. 2018. The neural basis of visual symmetry and its role in mid- and high-level visual processing. *Annals of the New York Academy of Sciences*. DOI: https://doi.org/10.1111/nyas.13667, PMID: 29604083

Bishop D. 2019. Rein in the four horsemen of irreproducibility. *Nature* **568**:435. DOI: https://doi.org/10.1038/d41586-019-01307-2, PMID: 31019328

Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA, Avesani P, Baczkowski BM, Bajracharya A, Bakst L, Ball S, Barilari M, Bault N, Beaton D, Beitner J, Benoit RG, et al. 2020. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**:84–88. DOI: https://doi.org/10.1038/s41586-020-2314-9, PMID: 32483374

Boudewyn MA, Luck SJ, Farrens JL, Kappenman ES. 2018. How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology* **55**:e13049. DOI: https://doi.org/10.1111/psyp.13049, PMID: 29266241

Brysbaert M. 2019. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition* **2**:16. DOI: https://doi.org/10.5334/joc.72, PMID: 31517234

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience* **14**:365–376. DOI: https://doi.org/10.1038/nrn3475, PMID: 23571845

Cattaneo Z. 2017. The neural basis of mirror symmetry detection: A review. *Journal of Cognitive Psychology* **29**:259–268. DOI: https://doi.org/10.1080/20445911.2016.1271804

Chambers CD. 2013. Registered reports: A new publishing initiative at Cortex. *Cortex* **49**:609–610. DOI: https://doi.org/10.1016/j.cortex.2012.12.016, PMID: 23347556

Chen CC, Kao KLC, Tyler CW. 2007. Face configuration processing in the human brain: the role of symmetry. *Cerebral Cortex* **17**:1423–1432. DOI: https://doi.org/10.1093/cercor/bhl054, PMID: 16923779

Cowan N, Belletier C, Doherty JM, Jaroslawska AJ, Rhodes S, Forsberg A, Naveh-Benjamin M, Barrouillet P, Camos V, Logie RH. 2020. How do scientific views change? Notes from an extended adversarial collaboration. *Perspectives on Psychological Science* **15**:1011–1025. DOI: https://doi.org/10.1177/1745691620906415

Derpsch Y, Rampone G, Piovesan A, Bertamini M, Makin ADJ. 2021. The extrastriate symmetry response is robust to variation in visual memory load. *Psychophysiology* **58**:e13941. DOI: https://doi.org/10.1111/psyp.13941, PMID: 34592790

Enquist M, Johnstone RA. 1997. Generalization and the evolution of symmetry preferences. *Proceedings of the Royal Society of London. Series B* **264**:1345–1348. DOI: https://doi.org/10.1098/rspb.1997.0186

Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA. 2021. Investigating the replicability of preclinical cancer biology. *eLife* **10**:e71601. DOI: https://doi.org/10.7554/eLife.71601, PMID: 34874005

Fiedler K, Schwarz N. 2015. Questionable research practices revisited. *Social Psychological and Personality Science* **7**:45–52. DOI: https://doi.org/10.1177/1948550615612150

Freeman J. 2015. Open source tools for large-scale neuroscience. *Current Opinion in Neurobiology* **32**:156–163. DOI: https://doi.org/10.1016/j.conb.2015.04.002, PMID: 25982977

Grammer K, Fink B, Møller AP, Thornhill R. 2003. Darwinian aesthetics: Sexual selection and the biology of beauty. *Biological Reviews of the Cambridge Philosophical Society* **78**:385–407. DOI: https://doi.org/10.1017/s1464793102006085, PMID: 14558590

Höfel L, Jacobsen T. 2007a. Electrophysiological indices of processing aesthetics: Spontaneous or intentional processes? *International Journal of Psychophysiology* **65**:20–31. DOI: https://doi.org/10.1016/j.ijpsycho.2007.02.007, PMID: 17400317

Höfel L, Jacobsen T. 2007b. Electrophysiological indices of processing symmetry and aesthetics: A result of judgment categorization or judgment report? *Journal of Psychophysiology* **21**:9. DOI: https://doi.org/10.1027/0269-8803.21.1.9

Jacobsen T, Höfel L. 2003. Descriptive and evaluative judgment processes: Behavioral and electrophysiological indices of processing symmetry and aesthetics. *Cognitive, Affective & Behavioral Neuroscience* **3**:289–299. DOI: https://doi.org/10.3758/cabn.3.4.289, PMID: 15040549

Jacobsen T, Klein S, Löw A. 2018. The posterior sustained negativity revisited—An SPN reanalysis of Jacobsen and Höfel (2003). *Symmetry* **10**:27. DOI: https://doi.org/10.3390/sym10010027

John LK, Loewenstein G, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* **23**:524–532. DOI: https://doi.org/10.1177/0956797611430953, PMID: 22508865

Jung TP, Makeig S, Humphries C, Lee TW, McKeown MJ, Iragui V, Sejnowski TJ. 2000. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* **37**:163–178. DOI: https://doi.org/10.1111/1469-8986.3720163, PMID: 10731767

Keefe BD, Gouws AD, Sheldon AA, Vernon RJW, Lawrence SJD, McKeefry DJ, Wade AR, Morland AB. 2018. Emergence of symmetry selectivity in the visual areas of the human brain: fMRI responses to symmetry presented in both frontoparallel and slanted planes.

Human Brain Mapping **39**:3813–3826. DOI: https://doi.org/10.1002/hbm.24211, PMID: 29968956

Kerr NL. 1998. HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review* **2**:196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4, PMID: 15647155

Klapwijk ET, van den Bos W, Tamnes CK, Raschle NM, Mills KL. 2021. Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Developmental Cognitive Neuroscience* **47**:100902. DOI: https://doi.org/10.1016/j.dcn.2020.100902, PMID: 33383554

Koffka K. 1935. Principles of Gestalt Psychology. New York: Harcourt, Brace & Company.

Kohler PJ, Clarke A, Yakovleva A, Liu Y, Norcia AM. 2016. Representation of maximally regular textures in human visual cortex. *The Journal of Neuroscience* **36**:714–729. DOI: https://doi.org/10.1523/JNEUROSCI.2962-15.2016, PMID: 26791203

Kohler PJ, Cottereau BR, Norcia AM. 2018. Dynamics of perceptual decisions about symmetry in visual cortex. *NeuroImage* **167**:316–330. DOI: https://doi.org/10.1016/j.neuroimage.2017.11.051, PMID: 29175495

Kohler PJ, Clarke ADF. 2021. The human visual system preserves the hierarchy of two-dimensional pattern regularity. *Proceedings. Biological Sciences* **288**:20211142. DOI: https://doi.org/10.1098/rspb.2021.1142, PMID: 34284623

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. 2009. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience* **12**:535–540. DOI: https://doi.org/10.1038/nn.2303, PMID: 19396166

Lindquist M. 2020. Pipeline choices alter neuroimaging findings. *Nature* **582**:36–37. DOI: https://doi.org/10.1038/d41586-020-01282-z, PMID: 32433631

Makin ADJ, Wilton MM, Pecchinenda A, Bertamini M. 2012. Symmetry perception and affective responses: A combined EEG/EMG study. *Neuropsychologia* **50**:3250–3261. DOI: https://doi.org/10.1016/j.neuropsychologia.2012.10.003, PMID: 23063934

Makin ADJ, Rampone G, Bertamini M. 2015. Conditions for view invariance in the neural response to visual symmetry. *Psychophysiology* **52**:532–543. DOI: https://doi.org/10.1111/psyp.12365, PMID: 25345662

Makin ADJ, Wright D, Rampone G, Palumbo L, Guest M, Sheehan R, Cleaver H, Bertamini M. 2016. An electrophysiological index of perceptual goodness. *Cerebral Cortex* **26**:4416–4434. DOI: https://doi.org/10.1093/cercor/bhw255, PMID: 27702812

Makin ADJ, Rampone G, Bertamini M. 2020a. Symmetric patterns with different luminance polarity (anti-symmetry) generate an automatic response in extrastriate cortex. *The European Journal of Neuroscience* **51**:922–936. DOI: https://doi.org/10.1111/ejn.14579, PMID: 31529733

Makin ADJ, Rampone G, Karakashevska E, Bertamini M. 2020b. The extrastriate symmetry response can be elicited by flowers and landscapes as well as abstract shapes. *Journal of Vision* **20**:11. DOI: https://doi.org/10.1167/jov.20.5.11, PMID: 32455428

Makin ADJ, Rampone G, Morris A, Bertamini M. 2020c. The formation of symmetrical gestalts is task-independent, but can be enhanced by active regularity discrimination. *Journal of Cognitive Neuroscience* **32**:353–366. DOI: https://doi.org/10.1162/jocn_a_01485, PMID: 31633466

Maris E, Oostenveld R. 2007. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* **164**:177–190. DOI: https://doi.org/10.1016/j.jneumeth.2007.03.024, PMID: 17517438

Martinovic J, Jennings BJ, Makin ADJ, Bertamini M, Angelescu I. 2018. Symmetry perception for patterns defined by color and luminance. *Journal of Vision* **18**:1–24. DOI: https://doi.org/10.1167/18.8.4, PMID: 30098176

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JPA. 2017. A manifesto for reproducible science. *Nature Human Behaviour* **1**:0021. DOI: https://doi.org/10.1038/s41562-016-0021, PMID: 33954258

Nelson LD, Simmons J, Simonsohn U. 2018. Psychology's renaissance. *Annual Review of Psychology* **69**:511–534. DOI: https://doi.org/10.1146/annurev-psych-122216-011836, PMID: 29068778

Norcia AM, Candy TR, Pettet MW, Vildavski VY, Tyler CW. 2002. Temporal dynamics of the human response to symmetry. *Journal of Vision* **2**:132–139. DOI: https://doi.org/10.1167/2.2.1, PMID: 12678588

Oka S, Victor JD, Conte MM, Yanagida T. 2007. VEPs elicited by local correlations and global symmetry. *Vision Research* **47**:2212–2222. DOI: https://doi.org/10.1016/j.visres.2007.03.020

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* **349**:aac4716. DOI: https://doi.org/10.1126/science.aac4716, PMID: 26315443

Pavlov YG, Adamian N, Appelhoff S, Arvaneh M, Benwell C, Beste C, Bland A, Bradford DE, Bublatzky F, Busch N, Clayson PE, Cruse D, Czeszumski A, Dreber A, Dumas G, Ehinger BV, Ganis G, He X, Hinojosa JA, Huber-Huber C, et al. 2020. #EEGManyLabs: Investigating the Replicability of Influential EEG Experiments. *PsyArXiv*. DOI: https://doi.org/10.31234/osf.io/528nr

Pernet C, Garrido MI, Gramfort A, Maurits N, Michel CM, Pang E, Salmelin R, Schoffelen JM, Valdes-Sosa PA, Puce A. 2020. Issues and recommendations from the OHBM COBIDAS MEEG committee for reproducible EEG and MEG research. *Nature Neuroscience* **23**:1473–1483. DOI: https://doi.org/10.1038/s41593-020-00709-0, PMID: 32958924

Rubin M. 2017. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology* **21**:308–320. DOI: https://doi.org/10.1037/gpr0000128

Sasaki Y, Vanduffel W, Knutsen T, Tyler CW, Tootell R. 2005. Symmetry activates extrastriate visual cortex in human and nonhuman primates. *PNAS* **102**:3159–3163. DOI: https://doi.org/10.1073/pnas.0500319102, PMID: 15710884

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**:1359–1366. DOI: https://doi.org/10.1177/0956797611417632, PMID: 22006061

Treder MS. 2010. Behind the looking-glass: A review on human symmetry perception. *Symmetry* **2**:1510–1543. DOI: https://doi.org/10.3390/sym2031510

Tyler CW. 1995. Empirical aspects of symmetry perception. *Spatial Vision* **9**:1–7. DOI: https://doi.org/10.1163/156856895x00089, PMID: 7626541

Tyler CW, Baseler HA, Kontsevich LL, Likova LT, Wade AR, Wandell BA. 2005. Predominantly extra-retinotopic cortical response to pattern symmetry. *NeuroImage* **24**:306–314. DOI: https://doi.org/10.1016/j.neuroimage.2004.09.018, PMID: 15627573

Tyson-Carr J, Bertamini M, Rampone G, Makin A. 2021. Source dipole analysis reveals a new brain response to visual symmetry. *Scientific Reports* **11**:285. DOI: https://doi.org/10.1038/s41598-020-79457-x, PMID: 33431986

Tyson-Carr J. 2022. The-SPN-Catalogue. swh:1:rev:75e729f867c275433b68807bc3f2228c57a3ccac. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:5c24336c033b5dd02f1022f9da3f0aee670c5b2f;origin=https://github.com/JohnTyCa/The-SPN-Catalogue;visit=swh:1:snp:ab29e39ad42a1793f38448ac018dcf4918935013;anchor=swh:1:rev:75e729f867c275433b68807bc3f2228c57a3ccac

van der Helm PA, Leeuwenberg EL. 1996. Goodness of visual regularities: A non-transformational approach. *Psychological Review* **103**:429–456. DOI: https://doi.org/10.1037/0033-295x.103.3.429, PMID: 8759043

Van Meel C, Baeck A, Gillebert CR, Wagemans J, Op de Beeck HP. 2019. The representation of symmetry in multi-voxel response patterns and functional connectivity throughout the ventral visual stream. *NeuroImage* **191**:216–224. DOI: https://doi.org/10.1016/j.neuroimage.2019.02.030, PMID: 30771448

Wagemans J. 1995. Detection of visual symmetries. *Spatial Vision* **9**:9–32. DOI: https://doi.org/10.1163/156856895x00098, PMID: 7626549

Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, Singh M, von der Heydt R. 2012. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin* **138**:1172–1217. DOI: https://doi.org/10.1037/a0029333, PMID: 22845751

Wright D, Makin ADJ, Bertamini M. 2017. Electrophysiological responses to symmetry presented in the left or in the right visual hemifield. *Cortex* **86**:93–108. DOI: https://doi.org/10.1016/j.cortex.2016.11.001, PMID: 27923173

Wright D, Mitchell C, Dering BR, Gheorghiu E. 2018. Luminance-polarity distribution across the symmetry axis affects the electrophysiological response to symmetry. *NeuroImage* **173**:484–497. DOI: https://doi.org/10.1016/j.neuroimage.2018.02.008, PMID: 29427849

Zwetsloot PP, Van Der Naald M, Sena ES, Howells DW, IntHout J, De Groot JA, Chamuleau SA, MacLeod MR, Wever KE. 2017. Standardized mean differences cause funnel plot distortion in publication bias assessments. *eLife* **6**:e24260. DOI: https://doi.org/10.7554/eLife.24260, PMID: 28884685