Coupling between the phase of a neural oscillation or bodily rhythm with behavior: Evaluation of different statistical procedures

Nicolai Wolpert, Catherine Tallon-Baudry

NeuroImage 236 (2021) 118050

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Coupling between the phase of a neural oscillation or bodily rhythm with behavior: Evaluation of different statistical procedures

Nicolai Wolpert[1,*], Catherine Tallon-Baudry[1]

Laboratoire de Neurosciences Cognitives et Computationnelles, Ecole Normale Supérieure, Inserm u960, PSL University, 24 rue Lhomond, Paris 75005, France

ARTICLE INFO

Keywords:
Oscillations
Rhythms
Behavior
Phase
Simulations
Circular statistics

ABSTRACT

Growing experimental evidence points at relationships between the phase of a cortical or bodily oscillation and behavior, using various circular statistical tests. Here, we systematically compare the performance (sensitivity, False Positive rate) of four circular statistical tests (some commonly used, i.e. Phase Opposition Sum, Circular Logistic Regression, others less common, i.e., Watson test, Modulation Index). We created semi-artificial datasets mimicking real two-alternative forced choice experiments with 30 participants, where we imposed a link between a simulated binary behavioral outcome with the phase of a physiological oscillation. We systematically varied the strength of phase-outcome coupling, the coupling mode (1:1 to 4:1), the overall number of trials and the relative number of trials in the two outcome conditions. We evaluated different strategies to estimate phase-outcome coupling chance level, as well as significance at the individual or group level. The results show that the Watson test, although seldom used in the experimental literature, is an excellent first intention test, with a good sensitivity and low False Positive rate, some sensitivity to 2:1 coupling mode and low computational load. Modulation Index, initially designed for continuous variables but that we find useful to estimate coupling between phase and a binary outcome, should be preferred if coupling mode is higher than 2:1. Phase Opposition Sum, coupled with a resampling procedure, is the only test retaining a good sensitivity in the case of a large unbalance in the number of occurrences of the two behavioral outcomes.

1. Introduction

Oscillations are ubiquitous in the sensory and cognitive brain (Buzsáki et al., 2013). The phase of neural oscillations modulates not only spike rate (Fries et al., 2007) and spike timing (Fiebelkorn and Kastner, 2020) but also behavior. For instance, the phase of infraslow, theta or alpha oscillations at which a near-threshold stimulus is presented correlates with its probability of detection, in the visual (Busch et al., 2009; Dugué et al., 2011; Helfrich et al., 2018; Mathewson et al., 2009), auditory (Ng et al., 2012; Rice and Hagstrom, 1989; Strauß et al., 2015) and somatosensory domain (Ai and Ro, 2013; Baumgarten et al., 2015; Monto et al., 2008). Phase has been also been related to other types of behavior such as reaction time (Callaway and Yeager, 1960; Dustman and Beck, 1965), decision-making (Wyart et al., 2012), visual search performance (Dugué et al., 2015) or auditory discrimination (Kayser et al., 2016; McNair et al., 2019). The timing of eye movements depends on brain alpha phase (Drewes and VanRullen, 2011; Gaarder et al., 1966; Hamm et al., 2012; Staudigl et al., 2017). Finally, the phase of various oscillatory bodily signals, such as the cardiac cycle, the gastric rhythm, or respiration, also influences both neural activity and behavior (for reviews, see Azzalini et al., 2019; Garfinkel and Critchley, 2016; Tort et al., 2018).

Most studies relating the phase of neural or bodily oscillations with behavior aim at establishing a statistical link between phase and a given binary outcome (e.g. “hit” or “miss” in a near-threshold detection experiment). However, just as there is an abundance of paradigms and cognitive variables studied, there is a large variety of statistical methods employed, hindering comparisons between studies. Besides, the reasons why a certain method is favored in a given experimental situation are usually not provided, and only few systematic investigations of the properties of statistical tests relating phase to behavior exist (VanRullen, 2016; Zoefel et al., 2019). Here, we systematically compare the performance of four statistical circular tests (Phase Opposition Sum, Circular Logistic Regression, Watson’s test and Modulation Index) to quantify relationships between oscillatory phase and a binary outcome with opposite preferred phases.

We created semi-artificial datasets based on real data acquired from 30 participants at rest, from which we extracted the phase of a neural rhythm (alpha rhythm, 8–12 Hz) and the phase of a bodily rhythm (gastric rhythm, ~0.05 Hz – Wolpert et al., 2020). We simulated a typical

* Corresponding author.
E-mail address: nicolaiwolpert@gmail.com (N. Wolpert).
[1] ORCIDs for authors as follows: Nicolai Wolpert: https://orcid.org/0000-0003-4381-0970 Catherine Tallon-Baudry: https://orcid.org/0000-0001-8480-5831

https://doi.org/10.1016/j.neuroimage.2021.118050
Received 24 November 2020; Received in revised form 12 February 2021; Accepted 6 April 2021
Available online 19 April 2021

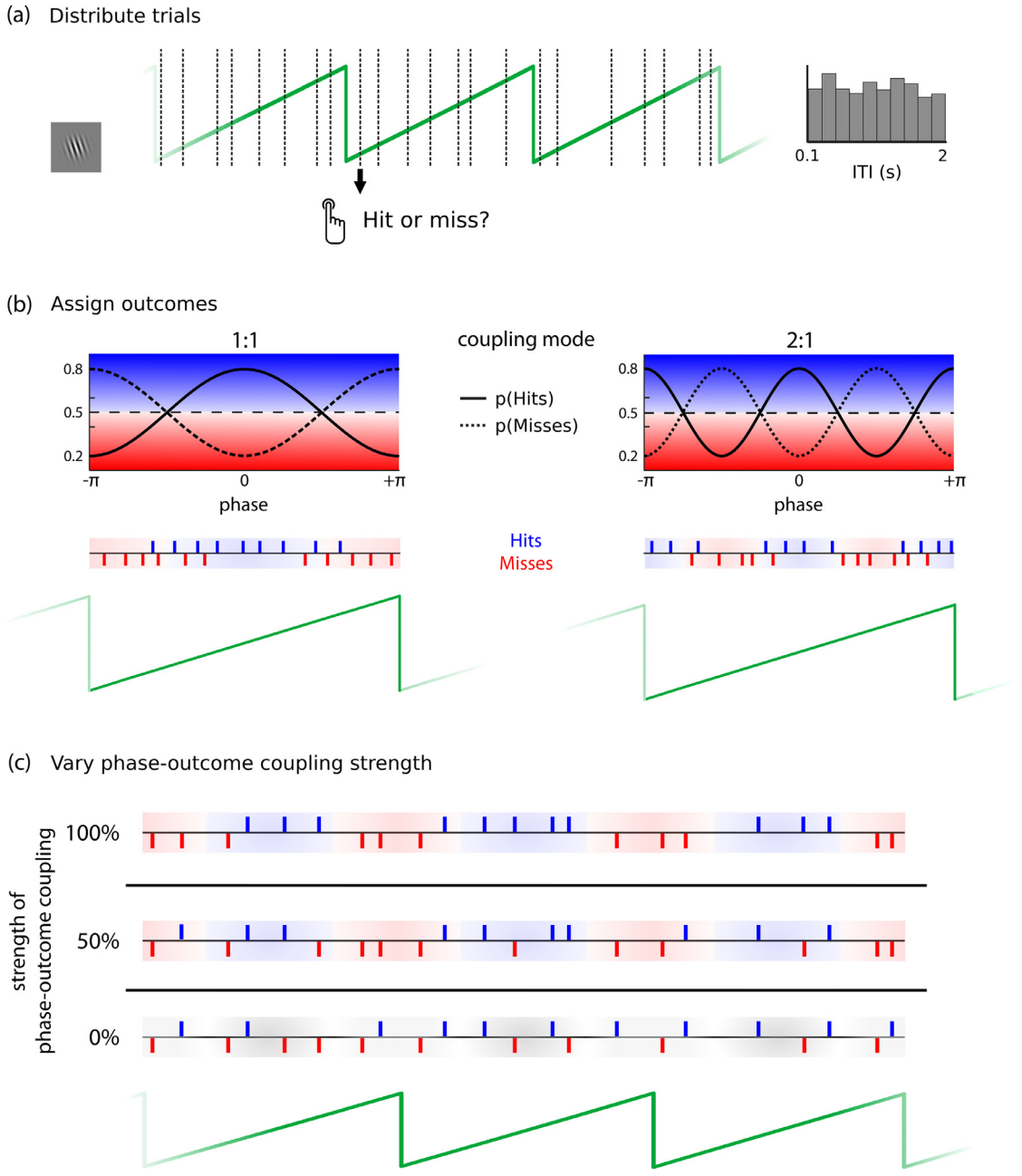

Fig. 1. Procedure for simulating behavior. (a) In a first step, a series of "trials" was distributed with a random Inter-Trial-Interval (ITI) selected from a flat distribution between 100 and 2000 ms, mimicking button presses uniformly distributed over time (b) Next, mutually exclusive behavioral outcomes, as in two-alternative forced choice experiments (hits and misses) were determined for each trial as a function of phase. A hit was assigned with mean probability of 0.5 (dashed lines), which was modulated over a cycle of the carrier frequency by a cosine function, such that hit probability (solid lines) ranged between 0.2 and 0.8. For a 1:1 coupling mode (left), p_{Hit} contained a single peak. For a 2:1 coupling mode (right), the probability function was rescaled to contain two peaks. The probability function for misses (dotted line) was defined as 1-p_{Hit}. By design, hits and misses where therefore distributed to occur at opposite phases. Middle rows show an example of resulting occurrences of hits and misses. Of note, the phase range at which a given outcome was more likely was fixed within a participant, but could vary between participants. (c) Phase-outcome coupling strength was varied by randomly reassigning labels (hits or misses) to a proportion of behavioral outcome. Top row: time series with 100% phase-outcome coupling, no label reassignment. Middle row: time series at 50% phase-outcome coupling strength (random label reassignment in 50% of the trials). Bottom row: time series with 0% phase-outcome coupling (random label reassignment in 100% of trials). Hits and misses are distributed randomly.

two-alternative forced choice experiment, where participants have to choose between two mutually exclusive options at each trial – stimulus seen vs. not in an experiment probing vision at threshold, or dog vs. cat in a categorization experiment with morphed images. Behavior was generated as transient events belonging to two categories ("hits" and "misses") with a flat distribution over time. We then imposed a statistical link between phase and behavior (Fig. 1), with hits more likely (resp. misses less likely) in one phase range, hence creating two behavioral outcomes with opposite preferred phases. The use of a semi-artificial dataset has the advantage of retaining all the complexity of real data, that can be difficult to model, such as the inter- and intra-individual variability in power law exponent (Podvalny et al., 2015; Voytek et al., 2015), individual differences in peak frequencies of oscillations of interest (Haegens et al., 2014), or the cycle duration variability necessary for some statistical procedures (Bahramisharif et al., 2013; Richter et al., 2017). While this approach is ideally suited to derive practical conclu-

sions on the performance of different statistical tests, which is the main aim of this article, it does not allow an in-depth assessment of the effect of data quality and features. We therefore also performed a selected control analysis on synthetic data.

We characterize each of the four tests not only by its sensitivity (the probability of finding an existing phase-outcome effect) but also its False Positive rate (the probability of a significant result despite the absence of a true effect). We systematically varied parameters such as the overall number of observations as well as the relative number of observations in the two outcome conditions. We also varied coupling mode. Indeed, most studies so far relied on the assumption that the phase-behavioral outcome relationship would be a 1:1 coupling with respect to an underlying carrier frequency in an a priori specified frequency band (e.g., 8–12 Hz alpha oscillation or 0.05 Hz gastric rhythm), i.e. with only one phase range associated with a given behavioral outcome. However, phase-behavioral outcome coupling might be more complex, with a given behavioral outcome being more frequent in several phase ranges of the band-specific oscillation, resulting in 2:1 coupling, over even higher coupling modes. Mathematically, a behavioral outcome with 2:1 coupling with an oscillation at frequency f would be equivalent to 1:1 coupling of behavior with an oscillation at frequency f^*2. However, data interpretation would be different. Indeed, the brain generates some specific rhythms at a given frequency, like the parieto-occipital rhythm. Bodily rhythms, such as respiration or the gastric rhythm, are defined by their central frequency (respectively ~0.3 Hz and ~0.05 Hz). Thus, from a biological perspective, 2:1 coupling at a (neural or bodily) carrier frequency is not equivalent to 1:1 coupling at twice the carrier frequency. Furthermore, it is possible that there exists inter-subject variability in coupling mode (i.e., with some subjects exhibiting 1:1 coupling, others 2:1 etc.), with regards to the *same* carrier frequency. We thus probed how the four tests compare in relation to such "higher" modes of coupling, and show that Modulation Index (Tort et al., 2010), originally devised for continuous variables, detects the link between (continuous) phase and a binary response variable.

Before presenting the results, we remind the reader of the rationale behind each of the four evaluation methods we test (Fig. 2), which are all *non-parametric* methods. In logistic regression, the phases of the two groups are used as circular predictors in a regression model to predict the outcome (e.g. choice in an auditory discrimination task – Kayser et al., 2016; McNair et al., 2019). Another method that has been proposed is *Phase Opposition Sum* (POS), which measures the extent to which phases of different groups cluster at different portions of a cycle (VanRullen, 2016). It is based on the Inter-Trial phase Coherence (ITC), which quantifies the extent of phase concentration across trials (Lachaux et al., 1999; Tallon-Baudry and Bertrand, 1999). Significance testing is done with non-parametric permutation statistics (VanRullen, 2016). The Watson test is the nonparametric version of the Watson-Williams two-sample test. It computes a test statistic U^2, which is based on the ordering of the phases and computing the cumulative relative frequency distributions. Last, we adapted the *Modulation Index* (MI, Tort et al., 2010), initially proposed to detect phase-amplitude coupling between continuous variables, to coupling between phase and a binary behavioral outcome. Here, MI is computed based on an event *rate* of one of the conditions (e.g., hit rate per phase bin). This method measures the extent to which an empirical distribution (here, hit rate per phase bin) differs from a uniform distribution. Significance is estimated by a surrogate procedure, as for POS.

2. Material and methods

2.1. Experimental data

We used real data to extract physiological phase time series, on which we simulated behavioral output. Data were obtained from 30 healthy participants (16 male, mean age 24, range 19–30) in resting-state with eyes open, 21 corresponding to already analyzed and published data (Richter et al., 2017; Wolpert et al., 2020) and the rest to an unpublished pilot study. All participants signed a written informed consent and were paid for participation. The procedures were approved by the Ethics Committee CPP Île de France III and were in accordance with the Helsinki declaration. Recordings were of 12–15 min length. Brain spontaneous activity was measured with an Elekta Neuromag® TRIUX magnetoencephalography (MEG) system with a sampling frequency of 1000 Hz. Signal Space Separation (tSSS) was performed using MaxFilter (Elekta Neuromag) to remove external noise. Subsequent analysis was conducted on magnetometer signals. The cardiac artifact was corrected using Independent Component Analysis (ICA), as implemented in the FieldTrip toolbox (Oostenveld et al., 2011). Briefly, MEG data were highpass-filtered at 0.5 Hz (zero phase shift 4th order butterworth filter) and epoched from 200 ms before to 200 ms after each R-peak. The number of independent components to be identified was the rank of the time × trial matrix. Continuous magnetometer data were then decomposed according to identified ICA components. The pairwise phase-consistency (PPC, Vinck et al., 2010) was computed between the ICA-decomposed signals and the ECG signal to isolate those components most reflective of ECG activity. Components with PPC values larger than 3 standard deviations than the mean were rejected iteratively from the continuous MEG data from each block until either no component exceeded 3 standard deviations or 3 components were rejected. In practice, this resulted in 3 components being rejected in each subject. Blink artifacts were defined by the EyeLink eyetracker system, padded by ±100 ms. Muscle and movement artifacts were identified automatically based on a z-value threshold on the MEG data filtered into a band of 110–140 Hz and 4–30 Hz respectively.

Concomitant to MEG, electrogastrogram (EGG) data were recorded by means of seven active electrodes placed on the abdominal skin (for details on EGG acquisition and preprocessing see Wolpert et al., 2020). Since we wanted to compare results using phase time series of two oscillations with very different frequencies, we extracted both the phase of MEG alpha oscillations (8–12 Hz) from the magnetometer channel with the largest alpha power, as well as the phase of the gastric slow rhythm (~0.05 Hz) from the abdominal electrode showing the largest EGG signal. To obtain the alpha phase time series, we first applied an 8–12 Hz bandpass 6th order Butterworth zero-phase shift filter using the Fieldtrip toolbox (Oostenveld et al., 2011). The EEG time series were filtered around gastric peak frequency (mean 0.049 ± 0.005 Hz) with a third-order frequency sampling designed finite impulse response filter (MATLAB: FIR2), with a bandwidth of ±0.015 Hz around gastric peak frequency. We then retrieved instantaneous phase applying the Hilbert transform to the filtered data.

2.2. Simulations of phase-behavior relationships

The general rationale for simulations was as follows. We simulated 1000 virtual "experiments" with 30 participants each. For each participant, we created an artificial time series of outcomes, "hits" and "misses" with a two-step procedure. First (Fig. 1a), we created a series of "events" (mimicking "trials" in a perceptual experiment) with a random time interval selected from a flat distribution between 100 and 2000 ms. In this way, trials were distributed uniformly with respect to phase. In a second step (Fig. 1b), the label "hit" or "miss" was assigned to each trial according to a probability function depending on phase. The outcome "hit" was assigned with a mean probability of 0.5, which was modulated as a cosine function of phase defined on $-\pi$ to $+\pi$, rescaled in amplitude to take values between 0.2 and 0.8. The probability function for misses was then defined as $p_{Miss} = 1 - p_{Hit}$. In other words, an event placed at the preferred phase for hits would have an 80% probability of being a hit and 20% probability of being a miss. Note also that because p_{Miss} and p_{Hit} sum up to 1 at each phase, hits and misses have opposite preferred phases. We rotated the probability function to a random degree, such that preferred phases varied across subjects within an experiment, as well as across experiments. To simulate higher cou-

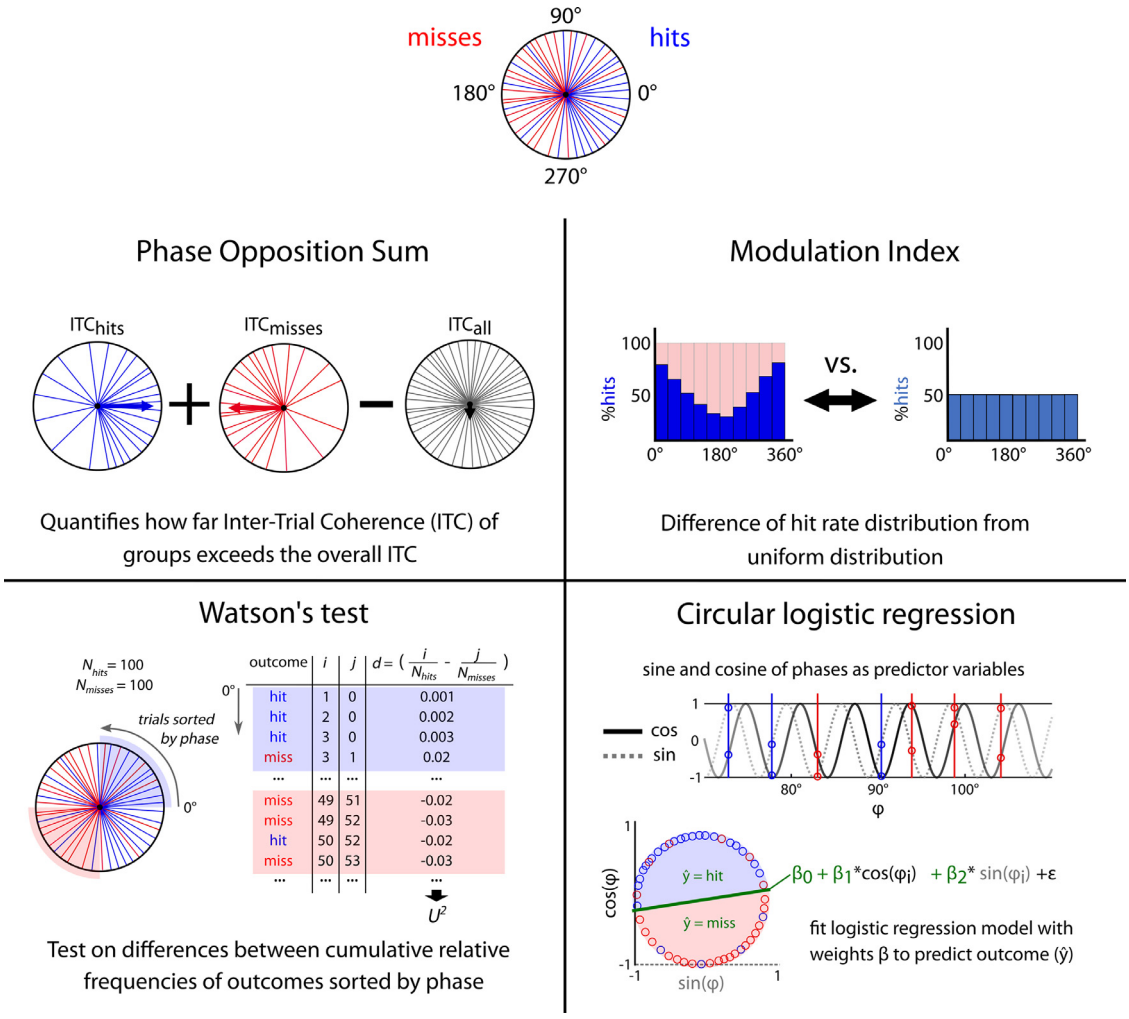

Fig. 2. Illustration of statistical tests compared in this paper. The aim is to assess whether hits (blue) and misses (red) are occurring at different oscillatory phases (polar circle, top). Phase Opposition Sum (VanRullen, 2016): This measure is based on the Inter-Trial Coherence (ITC), which quantifies the extent of phase concentration for a set of trials. Phase Opposition Sum combines the ITCs by subtracting the overall ITC from the separate ITC from each group. It thus becomes positive if the phases separated into hits and misses result in a higher ITC than the overall ITC. Modulation Index (Tort et al., 2010): The phase is binned into *N* phase bins of equal width, and the hit rate per phase bin computed, yielding a hit rate distribution. Note that the hit rate distribution is the mirror image of the miss rate distribution. If hits and misses occur at different portions of the cycle, the distribution will deviate from uniformity. MI measures the extent to which the empirical hit rate distribution deviates from a flat uniform distribution. Watson's test: Phases from hits and misses are sorted in ascending order, and for each trial, index *i* counts the cumulative number of hits and index *j* the cumulative number of misses. At each trial (row), the difference between the respective cumulative relative frequencies (*i*/#hits and *j*/#misses) is then computed. These differences are combined into a test statistic U^2 (for formula see 2.3). Circular logistic regression: The sine and cosine of phases for hits and misses are used as predictors in a circular logistic regression model with coefficients β_1 and β_2 and the intercept term β_0. To quantify the performance of the fit, a root-mean square is then computed using true outcomes and predictor coefficients.) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pling modes, probability functions were rescaled to contain two, three or four peaks and troughs per physiological phase cycle, thereby representing different "coupling modes", with either one, two, three or four preferred phases. We refer to these coupling modes as 1:1, 2:1, 3:1 and 4:1.

Finally, we introduced a parameter to vary the strength of the effect, which we call the *strength of phase-outcome coupling* (Fig. 1c). This was done by adding a certain amount of randomness or "noise" to the outcomes of events: A given proportion of trials was selected where a hit or miss was re-assigned with 50:50% chance. For example, with a strength of phase-outcome coupling of 30%, the outcome of trials would depend on phase in 30% of trials, whereas the remaining 70% of trials would be randomly selected, independently of phase. Finally, we randomly subsampled a set of a given size for hits and misses respectively, thereby controlling the number of trials for hits and misses and the relative number of trials in each group.

In sum, our simulations varied the following parameters: 1) *Strength of phase-outcome coupling*, or percentage of trials where outcome depended on phase; 2) *Coupling mode*, or number of peaks of the probability function for outcome by phase, reflecting the number of preferred phase ranges for each behavioral outcome. We refer to these coupling modes as 1:1, 2:1, 3:1 and 4:1, from 1 preferred phase range to 4. 3) Overall *number of trials* in the experiment, and 4) the *relative number of trials* for each behavioral outcome.

2.2.1. Sensitivity-analysis

We aimed to assess which statistical test would be most sensitive to detect phase-behavior relationships under a given coupling mode. For this, we generated a time series of hits and misses, keeping the total number of trials constant at 250 with as many hits as misses, while systematically increasing the strength of phase-outcome coupling. Phase-outcome coupling strength started from 0% (i.e., random behavior, no relation-

ship between behavioral outcome and phase) and was incremented in steps of 5% up to 40% (i.e., behavioral outcome depends on phase in 40% of the trials). We verified that the difference in the number of observations between the two conditions after imposing noise was not more than 10%. For each strength of phase-outcome coupling, we ran 1000 virtual experiments with 30 subjects each. For each virtual experiment, we distributed events in each subject separately, assigned the labels hits and misses to those events according to the probability functions by phase, and computed the phase-outcome statistics for the four different tests (see Section 2.3). We then assessed for each virtual experiment if there was a significant effect at the group-level ($p < \alpha = 0.05$; for how we assess significance at the group level, see Section 2.5). We repeated this procedure for each of the coupling modes investigated from 1:1 to 4:1.

We defined the False Positive rate as the percentage of experiments with 0% strength of phase-outcome coupling, i.e. no effect present, where significant group-level effect was (falsely) detected. Sensitivity (True Positive rate) for phase-outcome coupling strength larger than 0% was computed as the percentage of experiments correctly detecting an injected phase-outcome coupling. This allowed us to compare the performance of different statistical tests as the strength of phase-outcome coupling was gradually increased.

2.2.2. Relative trial number between groups

In a different set of simulations, we addressed how an imbalance in the number of observations for hits and misses would affect the statistical tests. We initially distributed 240 hits and 240 misses, with the strength of phase-outcome coupling fixed at either 15%, to estimate sensitivity, or 0%, to estimate False Positive rate. From this pool of 2×240 events, we subsampled a number of hits and misses, systematically varying the relative proportion in number of observations for hits vs. misses (i.e., 20:80, 30:70, 40:60, 70:30 and 80:20), while keeping the total number of trials constant at 300. To generate the sample with a ratio of 20:80, we subsampled 60 of the initial set of hits and kept all 240 misses. To generate the sample with a ratio of 30:70, we build on the 20:80 sample by adding 30 hits and removing 30 misses, and so on. This was done in 1000 virtual experiments with 30 subjects each.

To assess how a potential loss in sensitivity for an imbalanced number of trials can be recovered, we applied the resampling procedure proposed by (Dugué et al., 2015; Staudigl et al., 2017). This procedure works as follows: For each subject, one resamples (without replacement) as many trials from the group with more trials as there are trials in the smaller group, and recomputes the phase-outcome statistics. This procedure is repeated N times (in our case $N = 100$), resulting in a distribution of N resampled values. The true test statistic is then estimated as the mean of this resample distribution. To quantify the impact of this resampling procedure, we computed for each hit:miss proportion the phase-outcome statistics both with and without resampling. In sum, this yielded 2×2 conditions: Effect present or absent and resampling vs. no resampling.

2.2.3. Amplitude of the underlying oscillation

In an additional analysis, we investigated the impact of the amplitude of the oscillation modulating outcome. For this, we first created a synthetic 10 Hz oscillation as a sinewave of amplitude scaled to [−1;1], with a sampling frequency of 1000 Hz and 15 min duration (Fig. 3). For each of the 30 virtual subjects, hits and misses were assigned based on the synthetic 10 Hz sinewave. We then modulated the amplitude of the 10 Hz oscillation by a scaling factor ranging between 0 and 0.2 before adding it to background noise, generated as pink noise with an amplitude rescaled to [−1, 1]. The resulting combined signal was then filtered around 10 Hz (±1) using a 6th order Butterworth zero-phase shift filter, and the Hilbert transform was applied on the combined signal to extract instantaneous phase. The resulting phase time series thus represented the "empirical" phase time series whose signal-to-noise ratio depended on the amplitude of the true underlying oscillation. Phases for hits and misses were extracted, and the phase-outcome statistics computed for each amplitude. This was repeated in 1000 virtual experiments, to compute sensitivity and False Positive rate.

2.3. Statistical tests

We applied four circular statistical tests commonly used in the field of neuroscience (Fig. 2).

2.3.1. Phase Opposition Sum (POS)

The Phase Opposition Sum (POS) index is a non-parametric method assessing phase differences between conditions (Drewes and VanRullen, 2011; Dugué et al., 2011; VanRullen, 2016). It is based on a comparison of the phase concentration of hits and misses to a phase locking computed over all trials. The extent of phase concentration is quantified using the *Inter-Trial Coherence measure* (ITC – Tallon-Baudry et al., 1996; Lachaux et al., 1999), which is defined as:

$$ITC_{all} = \sum_{i=1}^{N_{all}} \Phi_i / N_{all} \quad (1)$$

$$ITC_{hits} = \sum_{i=1}^{N_{hits}} \Phi_i / N_{hits} \quad (2)$$

$$ITC_{misses} = \sum_{i=1}^{N_{misses}} \Phi_i / N_{misses} \quad (3)$$

where Φ_i is the phase angle at which the event i occurs, N_{all} the total number of trials, and N_{hits} and N_{misses} are the number of hits and misses. Inter-Trial Coherence quantifies the phase-locking of a circular distribution of phases by taking values between 0 (uniform phase distribution) and 1 (perfect phase-alignment). The *Phase Opposition Sum* (POS) is then defined as:

$$POS = ITC_{hits} + ITC_{misses} - 2ITC_{all} \quad (4)$$

POS is positive when the ITC of each group exceeds the overall ITC. The POS measure is a recent improvement (VanRullen, 2016) of the Phase Bifurcation Index, defined as $(ITC_{hits} - ITC_{all}) * (ITC_{misses} - ITC_{all})$, which has been the measure of choice for many studies on phase differences (e.g., Busch et al., 2009). Using the additive measure is motivated by the finding that POS is more robust to low trial numbers and differences in relative trial numbers between groups compared to the Phase Bifurcation Index and a number of other measures (Sherman et al., 2016; VanRullen, 2016).

Note that the raw POS value obtained is not yet informative whether significant phase-concentration is present or not. An additional step is required to quantify the deviance from a null distribution estimated by a permutation procedure, as will be detailed in Section 2.4.

2.3.2. Watson's test

The Watson test is the nonparametric version of the Watson-Williams two-sample test, the circular equivalent of a two-sample t-test for angular means (Baumgarten et al., 2015; Samaha et al., 2015; VanRullen, 2016), since it does not rely on the assumption that the sampled populations are unimodal. It is computed the following way (Zar, 2010): First, the phases of hits and misses are separately grouped in ascending order. Let N_{hits} and N_{misses} denote the number of samples in each group, and N the total number of samples ($N_{hits}+N_{misses}$). With i as the index of hits and j as index of misses, the cumulative relative frequencies for the observations in the two groups are then computed as i/N_{hits} and j/N_{misses}. Values of d_k (with k running from 1 to N) are defined as the differences between the two cumulative relative frequency distributions ($d_k = i/N_{hits} - j/N_{misses}$). The test statistics, called Watson's U^2, is then computed as:

$$U^2 = \frac{N_{hits} N_{misses}}{N^2} \left[\sum_{k=1}^{N} d_k^2 - \frac{\left(\sum_{k=1}^{N} d_k\right)^2}{N} \right] \quad (5)$$

Significance can be read from significance tables for U^2 (see Zar, 2010). We also estimate significance of U^2 using the same permutation procedure as for POS and MI, described in Section 2.4.

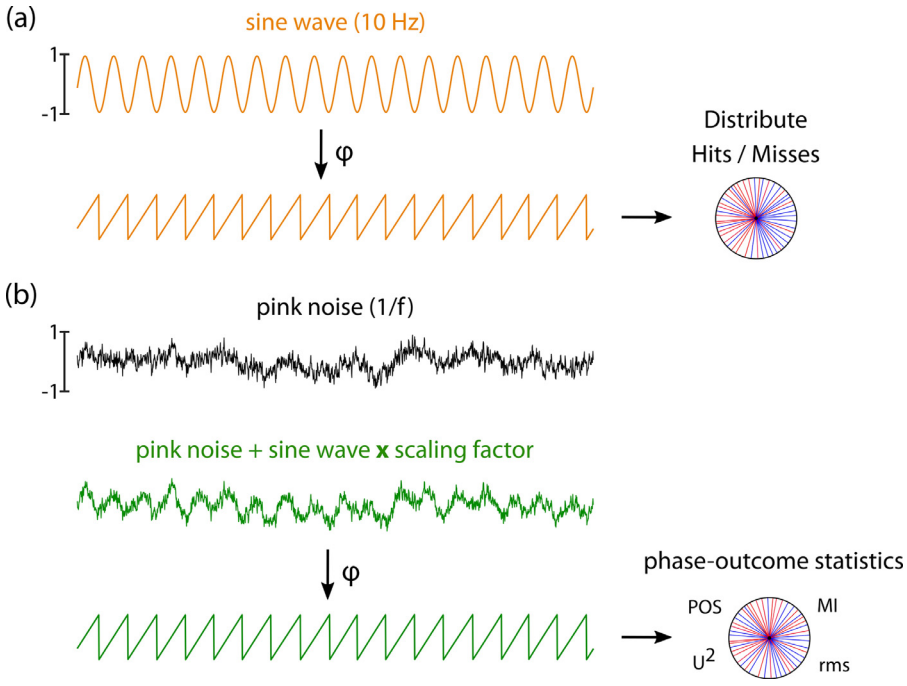

Fig. 3. Simulations on oscillatory amplitude. (a) A pure sine wave at 10 Hz was generated with an amplitude between −1 and 1, and hits and misses were distributed based on its instantaneous phase. (b) Background activity was simulated as pink noise with an amplitude between −1 and 1. The sine wave was multiplied with a scaling factor and added to background activity, and instantaneous phase retrieved. Phases for hits and misses based on the phase time series of this combined signal were then retrieved and the phase-outcome statistics computed for each amplitude.

2.3.3. Circular logistic regression

Circular logistic regression is used to test whether phase predicts outcome at the single-trial level. Phases are sine- and cosine transformed and used as circular predictors of the outcome in a regression model (Al-Daffaie and Khan, 2017):

$$\hat{y}_i = \beta_0 + \beta_1 \cos\Phi_i + \beta_2 \sin\Phi_i + \varepsilon \tag{6}$$

where $\hat{y}_i$ is the outcome for trial i, Φ_i is the phase at which the event occurred in trial i and ε the error term. A p-value for each participant can be directly obtained by comparing the full regression model with an intercept-only model using an F-Test, as described in (Zoefel et al., 2019). The root-mean-square of the obtained predictor coefficients ($\sqrt{\beta_1^2 + \beta_2^2}$) is used to quantify how well phase predicts behavioral outcome. As for the other tests, we here computed p-values using a permutation procedure (see Section 2.4) to assess whether the root-mean-square was higher than expected by chance.

2.3.4. Modulation Index (MI)

The Modulation Index (Tort et al., 2010), or MI, measures the extent to which an empirical distribution differs from a uniform distribution with the Kullback-Leibler distance. It was originally applied to detect phase-amplitude coupling (Tort et al., 2008, 2009), i.e. between phase and a *continuous* neural variable. We modified the method to quantify the relationship between phase and a *binary* response variable (hits vs. misses). We transform hits and misses into a hit *rate* per oscillatory phase bin. Phases are sorted into K bins spanning the $[-\pi, \pi]$ interval (here: $K = 10$), and the hit rate computed for each bin. MI measures how far the distribution of hit rate deviates from a uniform distribution with respect to phase bins. (Note that MI could also be computed based on miss rate. The two distributions are in fact complementary.) Formally, MI is defined as:

$$\mathrm{MI} = \frac{\log(K) + \sum_{j=1}^{K} P(j) * \log P(j)}{\log(K)} \tag{7}$$

Where P(j) is the standardized hit rate in phase bin j:

$$\mathrm{P(j)} = \frac{Hit\,rate\varphi(\mathrm{j})}{\sum_{i=1}^{K} Hit\,rate\varphi(\mathrm{i})} \tag{8}$$

And the hit rate per phase bin HR_{φ} is the number of hits in the phase bin φ divided by the total number of trials in the phase bin:

$$\mathrm{HR}_{\varphi} = \frac{N\,Hits\varphi}{N\,Hits\varphi + N\,Misses\varphi} \tag{9}$$

Note that MI thus differs from the other tests considered here since it is not directly based on the phases themselves but based on a *proportion* of hits relative to the number of trials in each bin. MI ranges from 0 if there is no phase-modulation of hit rate at all (meaning a perfectly uniform distribution) to 1 if there is perfect coupling (i.e. P(j)=1 for a given bin and 0 for all other bins). With a limited number of trials, it might happen that P(j) is zero, i.e. no hit in that phase bin. The bin can simply be ignored, because adding or removing an event with zero probability does not alter entropy.

2.4. Significance at the single subject level

While Watson's test and circular logistic regression directly return a p-value, POS and MI yield only raw values which do not inform on statistical significance. We use a permutation-based approach to estimate a distribution of phase statistics under the null-hypothesis of no phase-outcome relationship (Fig. 4a). This step is necessary for POS or MI but can be applied to other phase statistics (Watson's U^2, root mean squares of the logistic regression).

The rationale for this procedure is to abolish the hypothetical effect of phase on behavioral outcome in the original data by randomly reassigning behavioral outcomes in single participants. For each given subject, the event assignment to hit or miss is randomly permuted 100 times, e.g. maintaining the hit/miss proportion and timings but randomizing the link with phase by assigning the hit/miss label randomly. Because we distributed trials uniformly with respect to phase, surrogate phase distributions would also converge to uniformity (with slight departures due to noise – Fig. 4b). Phase statistics are recomputed at each permutation, to generate the distribution of phase statistics under the null hypothesis. Chance level is defined as the mean (or median, but see 3.2) of this null distribution. The comparison of the empirical phase statistics with the null distribution yields a Monte-Carlo p-value at the single subject level (proportion of surrogate values larger than the empirical one). Additionally, the difference between the empirical phase

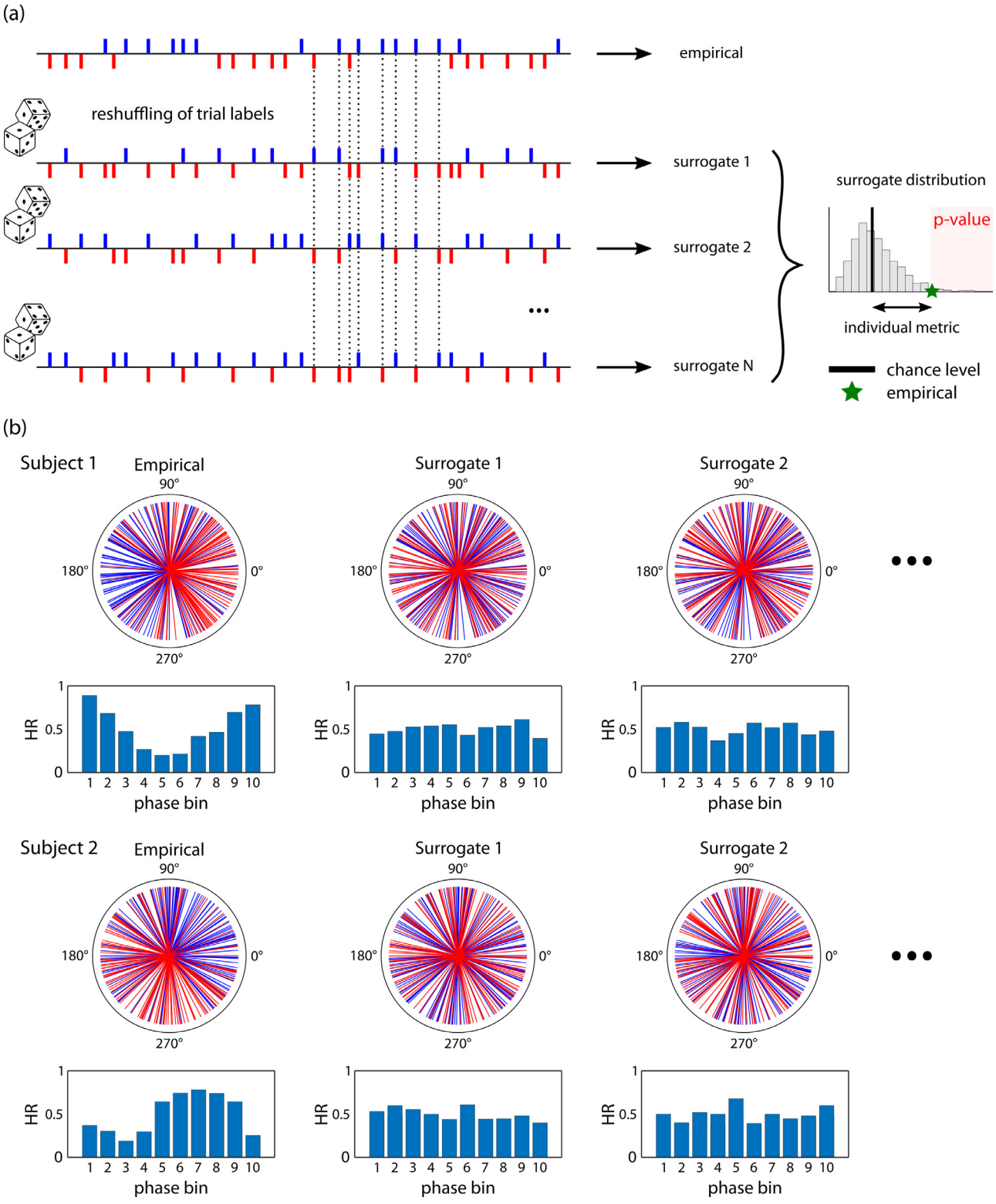

Fig. 4. (a) Permutation approach to estimate null distributions of no phase-outcome relationships. Top row: From the original time series of hits and misses, an empirical phase-outcome statistic is computed. The trial outcomes are then reshuffled (lower rows), with hit and miss labels randomly permuted, resulting in a new time series of hits and misses where the phase-outcome link is abolished while keeping the balance of relative number of observations and inter-stimulus intervals. This is repeated *N* times, and for each reshuffling, the phase-outcome statistic is computed. This results in a surrogate distribution (right). Chance level (black vertical line) is then defined as the mean of this distribution. The difference between the empirical phase-outcome statistic (green star) and chance level provides an individual metric of the strength of the phase-outcome effect. Additionally, an individual *p*-value can be computed as the proportion of surrogate values higher than empirical. (b) Example of empirical (left column) and surrogate (middle and right columns) phase distributions, for a phase-outcome coupling strength of 100% and 250 trials. Upper row: Polar representation, lower row: distributions of hit rate per phase bin.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

statistics and chance level provides an individual metric of the strength of the phase-outcome coupling.

2.5. Testing for significance at the group level

Significance at the group level can be assessed either by direct comparison between chance level statistics and empirical statistics, by computing surrogate averages, or by combining individual *p*-values.

2.5.1. Empirical vs. chance level

One option consists in comparing the empirical phase statistics to chance level estimates across participants, using a one-tailed paired-sample *t*-test, as in Richter et al., 2017. The test is one-tailed because the hypothesis is that there is more phase coupling than expected by chance.

2.5.2. Surrogate average

An alternative method is based on comparing the empirical group-average of phase statistics (e.g. POS values) against a null distribution of surrogate group-averages (Busch et al., 2009). One value is drawn randomly from the surrogate distribution from each subject, the average across subjects computed, and this is repeated 1000 times. This yields a distribution of 1000 surrogate averages under the null hypothesis. A Monte-Carlo *p*-value is then computed as the proportion of surrogate averages that are larger than the empirical average, and the result is considered as significant if this group-level *p*-value is below the threshold of significance (p=.05).

2.5.3. Combining p-values

Another option to calculate group-level significance is to combine the results of the individual subjects (VanRullen, 2016). We used the individual *p*-values corresponding to the proportion of permutations yielding a higher phase-outcome statistic than the empirical value. In case the *p*-value was smaller than $1 / N_{perm}$, we assigned the midpoint between zero and $1 / N_{perm}$, which is $1 / (2 * N_{perm})$. (vanRullen, 2016). To combine *p*-values, a wide range of different methods is available (Alves and Yu, 2014; Heard and Rubin-Delanchy, 2018; Loughin, 2004; Rosenthal, 1978), of which we selected three of the most frequently used.

2.5.3.1. Fisher's method. Fisher's method combines the individual *p*-values from K independent tests into the following test statistic (Fisher, 1938)

$$T = -2 * \sum_{i=1}^{K} \ln(p_i) \tag{10}$$

where p_i corresponds to the *p*-value of participant i. Under the null hypothesis, T follows a chi-square distribution with $2K$ degrees of freedom (Alves and Yu, 2014; Fisher, 1938; Zoefel et al., 2019). From this, a combined *p*-value can be obtained.

Fisher's method has been shown to be asymmetrically sensitive to small compared to large *p*-values (Whitlock, 2005). This might be a drawback depending on the context in which this test is used.

2.5.3.2. Stouffer's method. Stouffer's method (Stouffer, 1949) relies on a transformation of one-tailed *p*-values of K independent tests into equivalent z-scores, which are combined across observers into one z-score, which is finally turned back into a *p*-value:

$$p_{combined} = 1 - F(\sum_{i=0}^{K} \mathrm{F}^{-1}(1 - p_i) \; / \; \sqrt{N}) \tag{11}$$

Where F is the normal cumulative distribution function.

This test does not show any asymmetry with respect to *p*-values as mentioned for the Fisher method. It can be seen as a compromise between methods like Fisher's method with high sensitivity to small *p*-values and other methods with high sensitivity to large *p*-values (Heard and Rubin-Delanchy, 2018).

2.5.3.3. Edgington's method. Edgington proposed to combine *p*-values by a simple sum across K observations (Edgington, 1972):

$$S_E = \sum_{i=1}^{K} p_i \tag{12}$$

The combined *p*-value is then obtained from the cumulative distribution function for the resulting sum (Heard and Rubin-Delanchy, 2018; Zaykin et al., 2007).

2.6. Data and code availability statement

The custom code as well as a set of phase time series from real participants for performing the simulations for this article can be accessed online at the following address: https://github.com/niwolpert/Simulations_phase_statistics. Our scripts make use of Matlab's Circular Statistics Toolbox (Berens, 2009, available at: https://fr.mathworks.com/matlabcentral/fileexchange/10676-circular-statistics-toolbox-directional-statistics). In addition, we made use of Rufin van Rullen's code on Phase Opposition (VanRullen, 2016, available at: www.cerco.ups-tlse.fr/~rufin/PhaseOppositionCode/).

3. Results

3.1. Sensitivity of statistical tests across different coupling modes

We investigated the sensitivity of four statistical circular tests (circular logistic regression, Phase Opposition Sum (POS), Watson's test, the Raleigh test and Modulation Index (MI)) to coupling between oscillatory phase and behavioral outcome (i.e., hits vs. misses). We extracted physiological oscillations from real data of 30 participants, and created artificial series of behavioral outcomes where we controlled the statistical relationship between phase and outcome, systematically varying the percentage of events where the outcome probability depended on phase, which we call the phase-outcome coupling strength. We also varied coupling mode from 1:1 to 4:1 (Fig. 1b). For 30 virtual participants, we set the number of trials to 250 and distributed hits and misses with respect to phase, following a given coupling mode and strength of phase-outcome coupling, and applied the four statistical tests. For each subject we estimated a null distribution for the phase-outcome statistics of each test (POS, MI, Watson's U^2, and root-mean-square of circular logistic regression) by a reshuffling procedure (Fig. 4), and defined chance level as the mean of the surrogate distribution. We then tested across subjects if empirical values were significantly higher than chance levels by means of a one-tailed paired samples *t*-test (in the following denoted as the "empirical vs. chance method"). We performed 1000 of such virtual experiments with 30 participants each. Sensitivity was computed as the percentage of experiments detecting a significant effect ($p < .05$). Additionally, we estimated the False Positive rate for each of the tests by computing the number of experiments that yielded a significant difference when outcomes were purely randomly assigned, independently of phase.

We performed these simulations for two types of real oscillations of very different frequencies and origin (alpha 8–12 Hz oscillation measured with MEG and gastric slow wave at ~0.05 Hz, during resting state), to ensure that the results do not depend on the frequency or origin of oscillations. We also ran the same simulations on synthetic data (10 Hz sine wave superimposed on pink noise). As we observed very similar results for the different types of oscillations, we restrict the presentation of results to the real alpha oscillation.

The results on sensitivity for a 1:1 coupling mode are presented in Fig. 5a,b. We observed that circular logistic regression, POS and the Watson test were similarly sensitive (Fig. 5a), with True Positive rate saturating at 25% phase locking strength. Circular logistic regression was slightly more sensitive than the other two, to the cost of a higher False Positive rate (Fig. 5b). The Watson test appears as a sensitive method with low False Positive rate. The sensitivity of MI was well below these

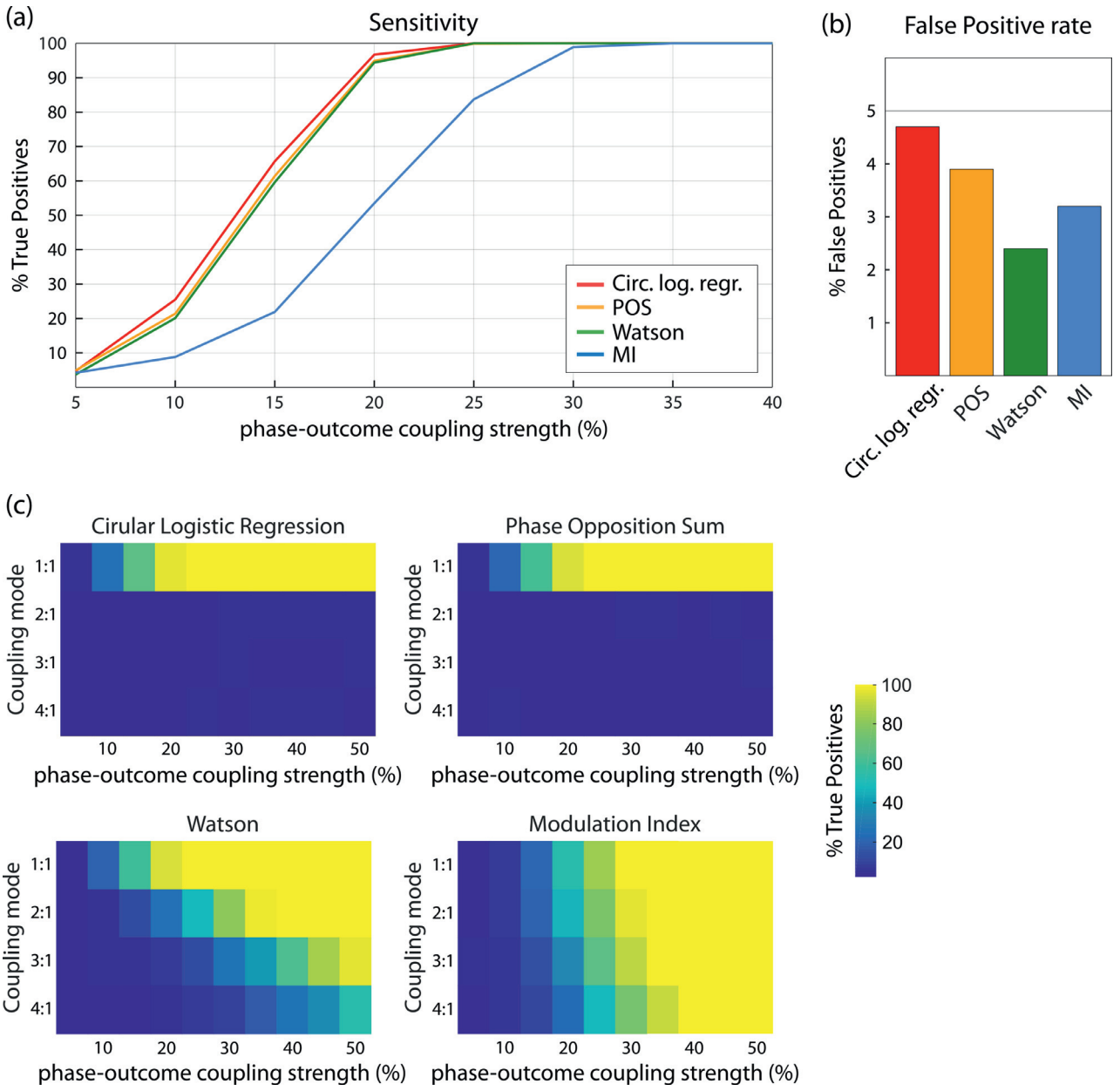

Fig. 5. Sensitivity of different tests to a 1:1 coupling mode (250 trials, 50% hits & misses). (a) Detection rate of True Positives as a function of phase locking strength. Circular logistic regression, POS and Watson's test clearly outperform MI. (b) False Positive rate computed based on outcomes randomly assigned, independently of phase. Red: Circular logistic regression, yellow: POS, green: Watson's test, blue: MI. (c) Sensitivity of the four tests as function of phase-outcome coupling strength and coupling mode. Color codes represent the percentage of True Positives. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

three tests, sensitivity saturating at 35% strength of phase-outcome coupling. The corresponding results for synthetic data are presented in Supplementary Figure 1.

Next, we investigated the influence of coupling mode on sensitivity by varying coupling mode from 1:1 to 4:1. Results are summarized in Fig. 5c. Circular logistic regression and POS do not detect coupling beyond the 1:1 coupling mode. The sensitivity of MI and the Watson test decreases when coupling mode increases, with a sharper decrease for Watson.

We also tested whether sensitivity depended on the overall number of trials for a 1:1 coupling mode by keeping the strength of phase-outcome coupling constant at 20% and gradually increasing the number of trials from 50 to 400 (Fig. 6). Sensitivity increased for all four statistical tests, and circular logistic regression, POS and the Watson test outperformed MI. False Positive Rate did not vary depending on the number of trials for any of the tests and was constantly below 5%.

In summary, we found clear differences in sensitivity between the statistical tests for different modes of coupling between phase and outcome: For 1:1 relationships, circular logistic regression, POS and the Watson test were all similarly sensitive, while MI was substantially less sensitive. In contrast, for higher forms of coupling, circular logistic regression and POS completely failed, with MI and Watson test as only sensitive tests. MI was the most powerful test for higher coupling modes. The Watson test was the only test being sensitive to all types of coupling modes, with a low False Positive rate.

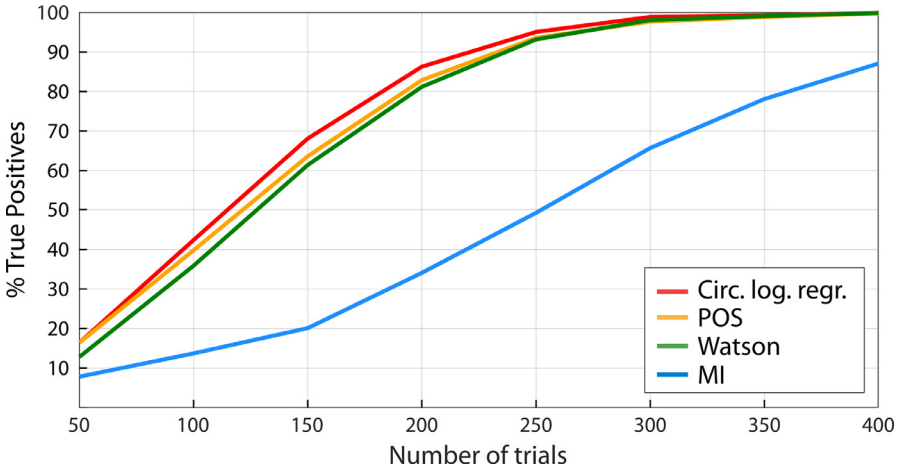

Fig. 6. Sensitivity by number of trials (1:1 coupling mode, 50% hits & misses). Phase-outcome coupling strength is kept constant at 20%, and the number of trials (hits & misses) is gradually increased. Sensitivity increases with number of trials. Circular logistic regression, POS and Watson outperform MI.

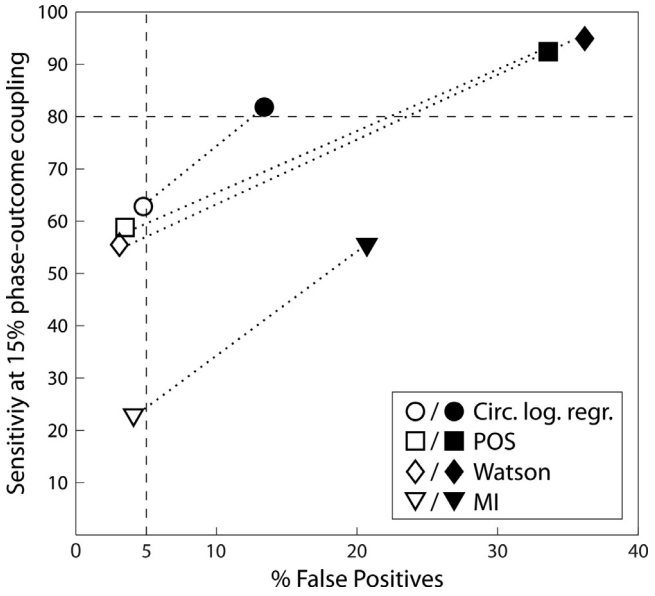

Fig. 7. False Positive rate at 0% phase-outcome coupling strength and sensitivity at 15% phase-outcome coupling strength (1:1 coupling mode, 300 trials, 50% hits & misses), when estimating chance level as the mean (open symbols) vs. the median (filled symbols) of surrogate distributions. Estimating chance level as the median of surrogate distributions increases sensitivity but also False Positive rate for all tests, especially for POS and the Watson test. False Positive rate remains below 5% for all tests when chance level is estimated as the mean of surrogate distributions.

3.2. Different ways of assessing significance at the group level

Significance of outcome-phase coupling at the group level can be assessed through various methods. In this section, we compare three different methods: empirical vs. chance (employed in the results described above), surrogate average, and combination of individual *p*-values.

In the *T-test empirical* vs. *chance level* approach, chance level is defined in each participant as the mean of the surrogate distribution. One can then test whether empirical values are higher than chance levels across participants with a one-tailed paired *t*-test. An advantage of this approach is that the difference between empirical and chance level coupling is summarized by one value (Richter et al., 2017). In the original proposal by Richter et al., chance level (for a continuous variable) was estimated as the median of the surrogate distribution. With the binary outcomes we test here, we found that using the median of the surrogate distribution inflates False Positive rate (Fig. 7) for all tests, while using the mean of the surrogate distribution produces False Positive rates below 5% for all tests. The reason for this is that our phase distributions resulted in surrogate distributions that were highly right-skewed (Fig. 4a). With right-skewed distributions, the mean is systematically larger than the median, resulting in higher estimate of chance level with the mean and hence a smaller False Positive rate. The skewness of surrogate distributions is thus critical for the resulting False Positive rate and sensitivity. Since skewness depends on data only, it is important to check the resulting False Positive rate and sensitivity for the data at hand to make an informed decision for the definition of chance level. Here, we decided to use the mean of the surrogate distribution as an estimation of chance level.

The *surrogate average* procedure directly generates a surrogate value at the group level. Significance is then expressed as the percentage of surrogate averages that are higher than the empirical average across participants. In practice, one first computes the average empirical value of the phase-outcome statistics (e.g. POS) across subjects. Then, a distribution of surrogate group-level averages is computed by randomly drawing one value from the surrogate distribution of each subject, computing the average over these random samples, and repeating this procedure a number of times (Busch et al., 2009). This method is computationally slightly more intensive since it requires an extra-step of surrogate statistics.

Finally, one can combine *p*-values obtained in each participant into a single group-level *p*-value (VanRullen, 2016). Individual *p*-values for each subject correspond to the percentage of surrogate values that are higher than the empirical individual value. Numerous methods exist for combining *p*-values – we here restrain our analyses to the methods of Fisher, Stouffer and Edgington.

For each of the 1000 virtual experiments, we tested for significance at the group level using each of these methods (empirical vs. chance, surrogate averages, and *p*-value combinations: Fisher, Stouffer and Edgington).

Results for a 1:1 coupling mode are presented in Fig. 8 for the Watson test and POS, with circular logistic regression and MI resulting in very similar profiles. All group-level statistics methods had a very similar sensitivity, except for the *p*-value combination using Stouffer's method, which consistently underperformed when strength of phase-outcome coupling was high. However, False Positive rate was consistently lower when using the *t*-test on empirical vs. chance level to assess significance at the group level.

3.3. Relative number of observations

In all previous simulations, the two behavioral outcomes were overall equally probable in each participant. Real experiments typically depart from this ideal balance in numbers of observations between con-

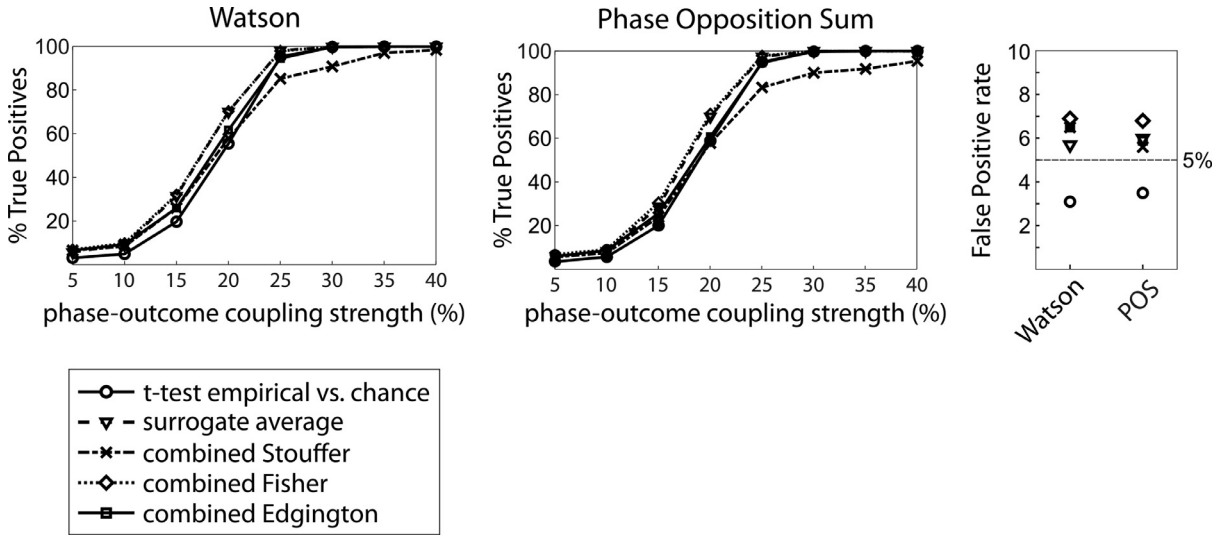

Fig. 8. Comparison of sensitivity and False Positive rate for the four different methods to test for significance at the group-level, with the examples of the Watson test and POS and for a 1:1 coupling mode (250 trials, 50% hits & misses). Circles: *t*-test on empirical vs. chance; triangles: surrogate average; stars: *p*-value combination using the Stouffer method; diamonds: *p*-value combination using the Fisher-method; squares: *p*-value combination using Edgington's method. Left and middle panel: Sensitivity; right panel: False Positive rate. Most methods perform very similarly, although the *p*-value combination using Stouffer's method performs comparably poorly for high strength of phase-outcome coupling, which was consistent across statistical tests. Using the paired *t*-test on empirical vs. chance resulted in the lowest False Positive rate for all the tests.

ditions, and differences in the relative number of observations might in turn influence the statistical power of the tests, as already demonstrated for POS (van Rullen, 2016). A "resampling procedure" has been proposed to correct for an imbalance in number of observations (Dugué et al., 2015; Staudigl et al., 2017). In each participant, a random subsample of *N* observations is drawn from the group with more observations, *N* being the number of observations in the condition with fewer observations. POS is recomputed, and the process repeated a 100 times, resulting in a distribution of resampled POS values, with the empirical POS being estimated as the mean of this resample distribution.

We assessed to which degree the different statistical tests are impaired by an imbalance in number of observations between conditions and whether statistical power can be recovered using the resampling procedure described above. In the following, we created a time series of hits and misses by setting the total number of trials to 300 and fixing the strength of phase-outcome coupling at 15%. We systematically varied the relative number of observations in each condition from 0.2:0.8 to 0.8:0.2 in steps of 0.1. We computed phase-outcome statistics for each relative number of observations, both with and without a resampling approach, in 1000 virtual experiments with 30 subjects each, with each of the methods for phase-outcome statistics, and estimated group level statistics in each virtual experiment using one sided paired *t*-test between empirical and chance level phase-outcome statistics. True Positive and False Positive rate were computed for each trial balance, with vs. without resampling procedure.

The results are presented in Fig. 9 for a 1:1 coupling mode. For a balanced number of observations (50:50), we replicate the results of Section 3.1 (Fig. 5), with circular logistic regression, POS and the Watson test being the most sensitive tests, and MI performing less well. We observed that all tests suffer from a loss in statistical power with larger imbalance in number of observations between conditions. Loss in power was most pronounced for POS. The resampling procedure (dotted lines) did not change the result for any of the tests except for POS, with a large gain in statistical power at large imbalance. Indeed, although POS performed less well than logistic regression and the Watson test for large imbalance, the use of the resampling procedure increased the statistical power of POS to the point that it performed better than all other methods for large imbalance. Even for an extreme imbalance of 20:80%, the sensitivity was only about 15% lower than for a 50:50% ratio. Note that since resampling removes trials to equalize number of observations, it results in an *overall* lower number of trials, which itself decreases the sensitivity of the test. We observed no systematic change in False Positive rate with relative trial number, which was below 5% for each statistical test and balance in trial number.

With varying ratios in trial numbers, MI showed some asymmetry in the sense that it performed better when the trial imbalance went into the direction of more hits than misses than vice versa. This is because MI is based on *hit* rate (the opposite pattern was observed when computing MI based on miss rate).

To conclude, these results demonstrated that all four statistical tests suffer from an imbalance in number of observations between conditions, with POS being most vulnerable. However, the resampling procedure restores POS sensitivity, which then exceeds the sensitivity of circular logistic regression, Watson and MI.

3.4. Amplitude of the underlying oscillation

In the previous simulations, we presented results using empirical data, and hence could not analyze how the amplitude of the underlying oscillation affects results. To analyze the influence of signal-to-noise ratio, we generated synthetic data. We first generated a synthetic 10 Hz sine wave, representing a "true" underlying oscillation, and assigned behavioral outcomes based on its instantaneous phase. To simulate background neural activity, we generated pink noise time series for 30 virtual subjects, with an amplitude in the range of −1 to 1. We scaled the amplitude of the sinewave by a factor varying between 0 and 0.2 before adding it to pink noise. The resulting signal was filtered around 10 Hz and instantaneous phase computed. We retrieved phases for hits and misses and computed phase-outcome statistics. This procedure was repeated in 1000 virtual experiments, and sensitivity and False Positive rate computed.

The results are presented in Fig. 10 for a 1:1 coupling mode. At zero amplitude (pure pink noise), the sensitivity of all the tests was below 5%, corresponding to their baseline False Positive rate. Sensitivity sharply increased between 0.02 and 0.05 and saturated at an amplitude around 0.15 for all tests. As expected, circular logistic regression, POS and Watson's test showed highest sensitivity while MI performed more poorly.

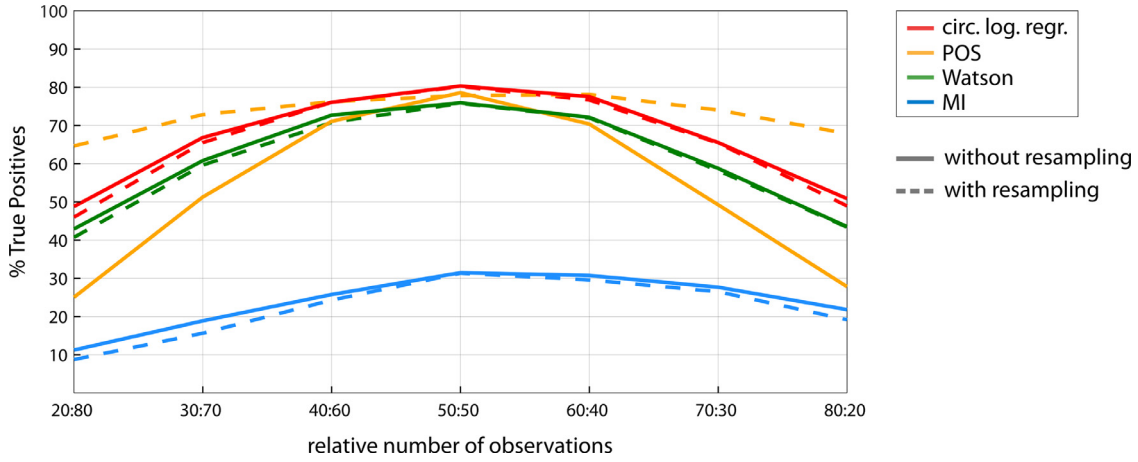

Fig. 9. Sensitivity of the different phase-outcome tests as a function of the relative number of observations for hits vs. misses (1:1 coupling mode, 300 trials, 15% phase-outcome coupling strength). Solid lines: without resampling; dotted lines: with resampling.

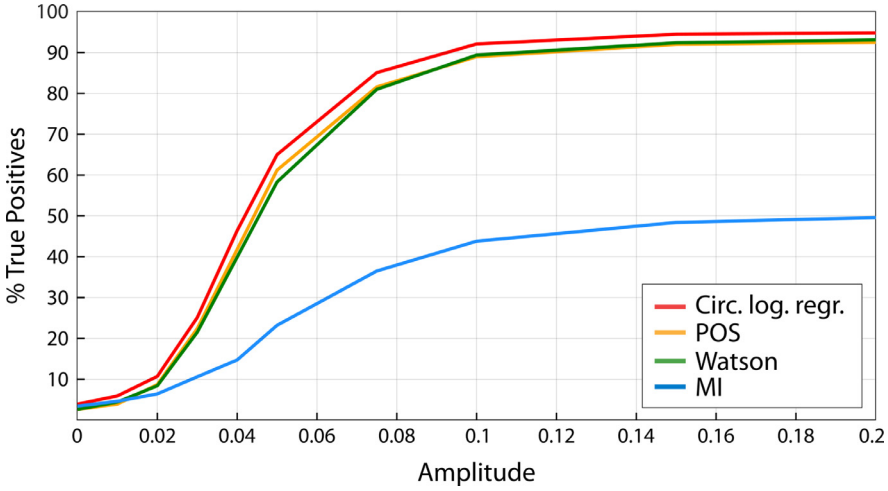

Fig. 10. Sensitivity of the different phase-outcome tests as a function of the amplitude of the 10 Hz oscillation relative to pink noise (1:1 coupling mode, 250 trials, 20% phase-outcome coupling strength).

3.5. Comparing permutation statistics with tabulated statistics

Two tests also directly output a *p*-value: The Watson test yields a U^2-statistic from which a *p*-value can be obtained from significance tables. For circular logistic regression, a *p*-value can be obtained by an *F*-test comparing the full regression model to an intercept-only model. However, in the results presented so far, we combined the Watson test and circular logistic regression with a permutation procedure to estimate a null distribution in each participant for consistency across all methods.

We compared the performance of these two approaches to compute a *p*-value on the individual level. For each of the 1000 virtual experiments, we computed individual *p*-values for each of the 30 virtual subjects using both the permutation *p*-value and the tabular *p*-value. For each strength of phase-outcome coupling, we computed the proportion of subjects with a significant *p*-value, to infer True Positive rate in the case of a strength of phase-outcome coupling above zero and a False Positive rate in the case of zero phase locking strength. Both strategies resulted in roughly equivalent sensitivity and False Positive rate (Fig. 11).

4. Discussion

We compared the performance of four statistical circular tests (POS, circular logistic regression, Watson's test and MI) at detecting relationships between phase and behavioral outcome. We created artificial data sets where we injected a statistical link between oscillatory phase and outcome (hit or miss) to compare the tests in terms of sensitivity and False Positive rate. We systematically varied the strength of the phase-outcome coupling and the coupling mode, as well as the total and relative numbers of observations. We observed that circular logistic regression, POS and the Watson test are similarly sensitive to a unimodal coupling mode (one preferred phase for each behavioral outcome). In comparison, MI performed poorly. The Watson test had the lowest False Positive rate, followed by MI, POS and logistic regression. In contrast, when going to higher coupling modes (groups have multiple opposed preferred phases), MI and Watson were the only sensitive tests, while all the other tests completely failed at detecting the effect. For those higher coupling modes, MI showed a higher sensitivity than the Watson test, especially for 3:1 and 4:1 coupling.

4.1. Advantages and limitations of each test for the detection of phase-outcome locking at the participant's level

Phase Opposition Sum has frequently been used to test phase-outcome coupling (e.g., Busch et al., 2009; Drewes and VanRullen, 2011; Dugué et al., 2011; Hamm et al., 2012; McLelland et al., 2016; Ruzzoli et al., 2019; Staudigl et al., 2017). Two methodological studies (VanRullen, 2016; Zoefel et al., 2019) identified POS as a powerful method for detecting (unimodal) coupling. Our findings are in line with these results, but we add to this literature that the sensitivity of POS is on par with the Watson test and circular logistic regression, and that the sensitivity of POS comes at the cost of a higher False Positive rate

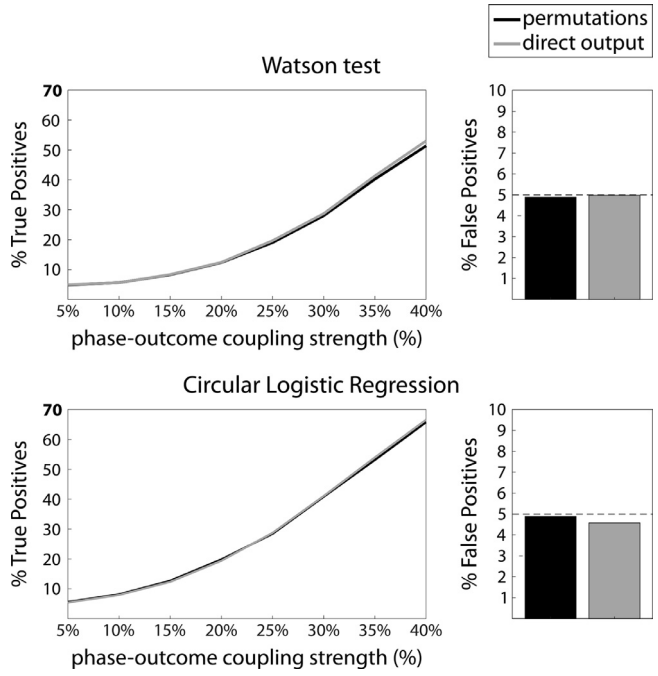

Fig. 11. Computing *p*-value based on permutations or on the direct output of the statistical test, for the Watson test and circular logistic regression, all for a 1:1 coupling mode (250 trials, 50% hits & misses). Left: Sensitivity (%True Positives), right: False Positive rate. Black: Permutation approach; gray: direct output. Both the Watson test and circular logistic regression are equally sensitive with either approach.

compared to the Watson test. Moreover, POS is not sensitive to higher coupling modes. POS is based on the Inter-Trial Coherence, which is the norm of the mean vector of all phases of a group of behavioral outcomes. If a behavioral outcome has two (or more) preferred phases, the resulting mean vector will be small, the two preferred phases tending to cancel each other. Compared to the other tests investigated here, POS is the most vulnerable to an imbalance in relative number of observations. This drop in sensitivity results from two factors: First, an insufficient number of trials in one of the two groups, which equally affects all the methods tested here (Fig. 6). Second, with high imbalance in relative number of observations, ITC_{all} becomes biased to toward either ITC_{hits} or ITC_{misses} (the one with more observations), which increases ITC_{all} and results in a reduced empirical POS (VanRullen, 2016). The resampling procedure compensates for this second factor, which is specific to POS, and has no effect on the other tests analyzed here. As a drawback, note though that resampling comes at the cost of longer computation time. Permutations can become computationally expensive especially if time-frequency data are analyzed (VanRullen, 2016), which increases exponentially with resampling.

Our results highlight the Watson test as an interesting method with several advantages that has remained underused in the experimental literature. First, the Watson test is among the three most sensitive methods for detecting unimodal coupling, and it also comes with the lowest False Positive rate. Second, it is the only method among the three winning methods for 1:1 coupling mode that was also sensitive to higher coupling modes. It could therefore be described as an "allrounder" method with a good tradeoff between sensitivity and False Positive rate and the potential to detect 2:1 coupling. Finally, the Watson test is computationally cheap: As it yields the same results whether the permutation procedure is used or *p*-values are directly computed, permutations are not strictly necessary for this test. It is also quite robust to moderate imbalances in relative number of observations.

Our findings concerning circular logistic regression as one of the most sensitive methods are in line with the results of Zoefel et al., 2019, who found it to be the best performing method. Still, by distinguishing between sensitivity and False Positive rate, we also observed that it came with the highest False Positive rate among the methods tested here, and it does not detect coupling modes higher than 1:1. The latter is due to the fact that circular logistic regression is fitting weights for an optimal linear separation between groups (Fig. 2), while two distributions clustering at more than one circular portion are not linearly separable.

We observed that among the four methods tested, MI was least sensitive to 1:1 coupling. This observation might be explained by an overestimation of chance level. Indeed, MI measures the departure of a distribution from uniformity (Tort et al., 2010). To determine whether a given MI value could be obtained by chance, we compare empirical MI with chance level MI, where chance level is estimated by reshuffling the behavioral outcomes, while keeping the same timing of behavioral events. Because the number of behavioral outcomes is finite, surrogate hit rate distributions only approximate uniformity, leading to a potential over-estimation of chance level. Additionally, MI values might be biased if behavioral events, irrespective of behavioral outcome, are not distributed evenly across all phase bins. Last, MI does not make any assumption on the *type* of departure from uniformity. What results in the good performance of MI at higher coupling modes comes at a cost at 1:1 coupling mode. In particular, a surrogate phase distribution might by chance display a bimodal distribution, which would be measured as a departure from non-uniformity and lead to an inflated chance level estimation. In contrast to 1:1 coupling, MI was clearly the most sensitive method for detecting higher coupling modes. Note that MI was originally devised for detecting relationships between phase and a continuous variable (e.g., phase-amplitude coupling – Tort et al., 2009). We show here that MI is also valuable to detect coupling between phase and a transient event such as a button press.

4.2. Testing for significance at the group level

We compared five different strategies to estimate significance on the group level: Running a paired *t*-test on empirical vs. chance level, creating a surrogate average distribution, and combining individual *p*-values with Stouffer's, Fisher's and Edgington's method. All methods had very similar sensitivity, except for Stouffer's *p*-value combination, which was substantially less sensitive when the strength of phase-outcome coupling was high. Note that there is an extensive literature investigating the power and properties of the different ways to combine *p*-values (e.g., Heard and Rubin-Delanchy, 2018; Whitlock, 2005). For example, it is considered that Fisher's method is asymmetrically sensitive to small *p*-values, while Stouffer's and Edgington's methods are seen as compromises between higher sensitivity to smaller vs. higher sensitivity to larger *p*-values. We here did not investigate in further detail how these methods compare, but observe that the Stouffer may be not the optimal choice in this context.

Using a *t*-test on empirical vs. chance resulted in a lower False Positive rate than the other four methods. Another advantage of relying on empirical vs. chance level is that one can quantify the strength of coupling on the individual level by computing empirical minus chance. This gives a continuous measure that can be regressed against other parameters of interest (e.g., to identify individual factors like age or daytime of recording that explain interindividual differences in strength of coupling). For those two reasons, the empirical vs. chance test seems a good option. Importantly, we here found that estimating chance level as the mean, rather than the median, of surrogate distributions should be preferred to avoid large False Positive rates. More generally, this result points to the importance of the method retained to estimate chance level when computing statistics on phase. We demonstrate this importance for phase-behavior coupling, but similar issues probably also arise for phase-phase or phase-amplitude coupling.

One important aspect in the design of our simulations is that we randomly vary the preferred phase for hits and misses from one sub-

ject to the other. Several previous studies have relied on the assumption that preferred phase would be constant across subjects (e.g., Mathewson et al., 2009; Monto et al., 2008; Rice and Hagstrom, 1989). However, there are different reasons for this to not hold true. For neural data, the measured phase at the scalp level might differ between participants due to factors like conduction relays (VanRullen, 2016). Depending on the context, it might therefore be advisable to focus on relative phase difference instead of absolute measured phase. To circumvent this issue, studies analyzing effects of phase on hit rate (Baumgarten et al., 2015; Zoefel and Heil, 2013) or continuous outcomes like evoked responses (Busch and VanRullen, 2010; Chakravarthi and Vanrullen, 2012; Neuling et al., 2012) have frequently realigned phase bins to the "preferred phase" for each participant (e.g., the phase with highest hit rate), to then run tests (e.g., ANOVA, Rayleigh test, or circular-linear correlation) on the phases pooled across subjects. Here, the tests we are using do not rely on phases being consistent across participants, and therefore do not require this additional realignment step.

Conclusions

In conclusion, we advocate the use of the Watson test, especially if the imbalance between observations in each condition is not larger than 40:60, and one wants to be open to higher coupling modes. POS becomes the measure of choice for 1:1 coupling when there is a large imbalance in the relative number of observations. In case one wants to investigate more complex coupling modes, MI seems as the optimal choice. To estimate significance on the group level, a good strategy is to compare empirical vs. chance levels, which comes with a low False Positive rate and provides an individual metric for the strength of the effect.

Note that we here constrained to the scenario of a binary outcome and did not consider the case of only one condition (e.g., clustering of saccades at a specific portion of the cardiac cycle – Ohl et al., 2016). Among the tests considered here, only MI can be directly used to assess the phase-dependency of events of only one type. MI has the advantage of detecting higher coupling modes, but requires a sufficient number of events to be present in all phase bins, which might not be the case for events whose onset is not in the control of the experimenter. Alternatively, other one-sample tests such as the Rayleigh test can be applied in this scenario (Ai and Ro, 2013; Galvez-Pol et al., 2019; Wyart et al., 2012). In addition, because we modelled a two-alternative forced choice experiment where the two behavioral outcomes are by design of opposite phase, we did not consider other possibilities, such as one outcome clustered at a specific phase, and the other homogenously distributed.

Data and code availability statement

The custom code as well as a set of phase time series from real participants for performing the simulations for this article can be accessed online at the following address: https://github.com/niwolpert/Simulations_phase_statistics. Our scripts make use of Matlab's Circular Statistics Toolbox (Berens, 2009, available at: https://fr.mathworks.com/matlabcentral/fileexchange/10676-circular-statistics-toolbox-directional-statistics). In addition, we made use of Rufin van Rullen's code on Phase Opposition (VanRullen, 2016, available at: www.cerco.ups-tlse.fr/~rufin/PhaseOppositionCode/).

Credit author statement

Catherine Tallon-Baudry: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing - review & editing **Nicolai Wolpert**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Roles/Writing - original draft

Declaration of Competing Interest

The authors declare that no competing interests exist.

Acknowledgments

This work was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 670325, Advanced grant BRAVIUS) and by a senior fellowship of the Canadian Institute for Advanced Research (CIFAR) program in Brain, Mind and Consciousness to C.T.-B., as well as from ANR-17-EURE-0017. The authors thank Rufin Van Rullen for providing code for Phase Opposition Sum, and Adriano Tort for providing Modulation Index code.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118050.

References

Ai, L., Ro, T., 2013. The phase of prestimulus alpha oscillations affects tactile perception. J. Neurophysiol. 111, 1300–1307. doi:10.1152/jn.00125.2013.
Al-Daffaie, K., Khan, S., 2017. Logistic regression for circular data. AIP Conf. Proc. 1842, 030022. doi:10.1063/1.4982860.
Alves, G., Yu, Y.-K., 2014. Accuracy evaluation of the unified P-value from combining correlated P-values. PLoS ONE 9, e91225. doi:10.1371/journal.pone.0091225.
Azzalini, D., Rebollo, I., Tallon-Baudry, C., 2019. Visceral signals shape brain dynamics and cognition. Trends Cogn. Sci. 23, 488–509. doi:10.1016/j.tics.2019.03.007.
Bahramisharif, A., Gerven, M.A.J.van, Aarnoutse, E.J., Mercier, M.R., Schwartz, T.H., Foxe, J.J., Ramsey, N.F., Jensen, O, 2013. Propagating neocortical gamma bursts are coordinated by traveling alpha waves. J. Neurosci. 33, 18849–18854. doi:10.1523/JNEUROSCI.2455-13.2013.
Baumgarten, T.J., Schnitzler, A., Lange, J., 2015. Beta oscillations define discrete perceptual cycles in the somatosensory domain. Proc. Natl. Acad. Sci. USA 112, 12187–12192. doi:10.1073/pnas.1501438112.
Berens, P., 2009. CircStat: a MATLAB toolbox for circular statistics. J. Stat. Softw. 31, 1–21. doi:10.18637/jss.v031.i10.
Busch, N.A., VanRullen, R., 2010. Spontaneous EEG oscillations reveal periodic sampling of visual attention. Proc. Natl. Acad. Sci. USA 107, 16048–16053. doi:10.1073/pnas.1004801107.
Busch, N.A., Dubois, J., VanRullen, R., 2009. The phase of ongoing EEG oscillations predicts visual perception. J. Neurosci. 29, 7869–7876. doi:10.1523/JNEUROSCI.0113-09.2009.
Buzsáki, G., Logothetis, N., Singer, W., 2013. Scaling brain size, keeping timing: evolutionary preservation of brain rhythms. Neuron 80, 751. doi:10.1016/j.neuron.2013.10.002.
Callaway, E., Yeager, C.L., 1960. Relationship between reaction time and electroencephalographic alpha phase. Science 132, 1765–1766. doi:10.1126/science.132.3441.1765.
Chakravarthi, R., VanRullen, R., 2012. Conscious updating is a rhythmic process. Proc. Natl. Acad. Sci. USA 109, 10599–10604. doi:10.1073/pnas.1121622109.
Drewes, J., VanRullen, R., 2011. This is the rhythm of your eyes: the phase of ongoing electroencephalogram oscillations modulates saccadic reaction time. J. Neurosci. 31, 4698–4708. doi:10.1523/JNEUROSCI.4795-10.2011.
Dugué, L., Marque, P., VanRullen, R., 2011. The phase of ongoing oscillations mediates the causal relation between brain excitation and visual perception. J. Neurosci. 31, 11889–11893. doi:10.1523/JNEUROSCI.1161-11.2011.
Dugué, L., Marque, P., VanRullen, R., 2015. Theta oscillations modulate attentional search performance periodically. J. Cogn. Neurosci. 27, 945–958. doi:10.1162/jocn_a_00755.
Dustman, R.E., Beck, E.C., 1965. Phase of alpha brain waves, reaction time and visually evoked potentials. Electroencephalogr. Clin. Neurophysiol. 18, 433–440. doi:10.1016/0013-4694(65)90123-9.
Edgington, E.S., 1972. An additive method for combining probability values from independent experiments. J. Psychol. 80, 351–363. doi:10.1080/00223980.1972.9924813.
Fiebelkorn, I.C., Kastner, S. (2020). Spike timing in the attention network predicts behavioral outcome prior to target selection. BioRxiv 2020.04.03.024109. doi:10.1016/j.neuron.2020.09.039
Fisher, R.A., 1938. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
Fries, P., Nikolić, D., Singer, W., 2007. The gamma cycle. Trends Neurosci. 30, 309–316. doi:10.1016/j.tins.2007.05.005.
Gaarder, K.R., Koresko, R.L., Kropfl, W., 1966. The phasic relation of a component of alpha rhythm to fixation saccadic eye movements. Electroencephalogr. Clin. Neurophysiol. 21, 544–551. doi:10.1016/0013-4694(66)90173-8.
Galvez-Pol, A., McConnell, R., Kilner, J.M., 2019. Active sampling in visual search is coupled to the cardiac cycle. Cognition 196, 104149. doi:10.1016/j.cognition.2019.104149.
Garfinkel, S.N., Critchley, H.D., 2016. Threat and the body: how the heart supports fear processing. Trends Cogn. Sci. (Regul. Ed.) 20, 34–46. doi:10.1016/j.tics.2015.10.005.

Haegens, S., Cousijn, H., Wallis, G., Harrison, P.J., Nobre, A.C., 2014. Inter- and intra-individual variability in alpha peak frequency. Neuroimage 92, 46–55. doi:10.1016/j.neuroimage.2014.01.049.

Hamm, J.P., Dyckman, K.A., McDowell, J.E., Clementz, B.A., 2012. Pre-cue fronto-occipital alpha phase and distributed cortical oscillations predict failures of cognitive control. J. Neurosci. 32, 7034–7041. doi:10.1523/JNEUROSCI.5198-11.2012.

Heard, N., Rubin-Delanchy, P., 2018. Choosing between methods of combining p-values. Biometrika 105, 239–246. doi:10.1093/biomet/asx076.

Helfrich, R.F., Fiebelkorn, I.C., Szczepanski, S.M., Lin, J.J., Parvizi, J., Knight, R.T., Kastner, S., 2018. Neural mechanisms of sustained attention are rhythmic. Neuron 99, 854–865. doi:10.1016/j.neuron.2018.07.032, e5.

Kayser, S.J., McNair, S.W., Kayser, C., 2016. Prestimulus influences on auditory perception from sensory representations and decision processes. Proc. Natl. Acad. Sci. USA 113, 4842–4847. doi:10.1073/pnas.1524087113.

Lachaux, J.-P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. Hum. Brain Mapping 8, 194–208. doi:10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C.

Loughin, T.M., 2004. A systematic comparison of methods for combining p-values from independent tests. Comput. Stat. Data Anal. 47, 467–485. doi:10.1016/j.csda.2003.11.020.

Mathewson, K.E., Gratton, G., Fabiani, M., Beck, D.M., Ro, T., 2009. To see or not to see: prestimulus alpha phase predicts visual awareness. J. Neurosci. 29, 2725–2732. doi:10.1523/JNEUROSCI.3963-08.2009.

McLelland, D., Lavergne, L., VanRullen, R., 2016. The phase of ongoing EEG oscillations predicts the amplitude of peri-saccadic mislocalization. Sci. Rep. 6, 29335. doi:10.1038/srep29335.

McNair, S.W., Kayser, S.J., Kayser, C., 2019. Consistent pre-stimulus influences on auditory perception across the lifespan. Neuroimage 186, 22–32. doi:10.1016/j.neuroimage.2018.10.085, Epub 2018 Nov 2.

Monto, S., Palva, S., Voipio, J., Palva, J.M., 2008. Very slow EEG fluctuations predict the dynamics of stimulus detection and oscillation amplitudes in humans. J. Neurosci. 28, 8268–8272. doi:10.1523/JNEUROSCI.1910-08.2008.

Neuling, T., Rach, S., Wagner, S., Wolters, C.H., Herrmann, C.S., 2012. Good vibrations: oscillatory phase shapes perception. Neuroimage 63, 771–778. doi:10.1016/j.neuroimage.2012.07.024.

Ng, B.S.W., Schroeder, T., Kayser, C, 2012. A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. J. Neurosci. 32, 12268–12276. doi:10.1523/JNEUROSCI.1877-12.2012.

Ohl, S., Wohltat, C., Kliegl, R., Pollatos, O., Engbert, R., 2016. Microsaccades are coupled to heartbeat. J. Neurosci. 36, 1237–1241. doi:10.1523/JNEUROSCI.2211-15.2016.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J., 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci. doi:10.1155/2011/156869.

Podvalny, E., Noy, N., Harel, M., Bickel, S., Chechik, G., Schroeder, C.E., Mehta, A.D., Tsodyks, M., Malach, R., 2015. A unifying principle underlying the extracellular field potential spectral responses in the human cortex. J Neurophysiol 114, 505–519. doi:10.1152/jn.00943.2014.

Rice, D.M., Hagstrom, E.C., 1989. Some evidence in support of a relationship between human auditory signal-detection performance and the phase of the alpha cycle. Percept. Mot. Skills 69, 451–457. doi:10.2466/pms.1989.69.2.451.

Richter, C.G., Babo-Rebelo, M., Schwartz, D., Tallon-Baudry, C., 2017. Phase-amplitude coupling at the organism level: the amplitude of spontaneous alpha rhythm fluctuations varies with the phase of the infra-slow gastric basal rhythm. Neuroimage 146, 951–958. doi:10.1016/j.neuroimage.2016.08.043.

Rosenthal, R., 1978. Combining results of independent studies. Psychol. Bull. 85, 185–193. doi:10.1037/0033-2909.85.1.185.

Ruzzoli, M., Torralba, M., Morís Fernández, L., Soto-Faraco, S., 2019. The relevance of alpha phase in human perception. Cortex 120, 249–268. doi:10.1016/j.cortex.2019.05.012.

Samaha, J., Bauer, P., Cimaroli, S., Postle, B.R., 2015. Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. Proc. Natl. Acad. Sci. USA 112, 8439–8444. doi:10.1073/pnas.1503686112.

Sherman, M.T., Kanai, R., Seth, A.K., VanRullen, R., 2016. Rhythmic influence of top-down perceptual priors in the phase of prestimulus occipital alpha oscillations. J. Cogn. Neurosci. 28, 1318–1330. doi:10.1162/jocn_a_00973.

Staudigl, T., Hartl, E., Noachtar, S., Doeller, C.F., Jensen, O., 2017. Saccades are phase-locked to alpha oscillations in the occipital and medial temporal lobe during successful memory encoding. PLoS Biol. 15, e2003404. doi:10.1371/journal.pbio.2003404.

Stouffer, S.A., 1949. The American Soldier, Vol.1: Adjustment during Army Life. Princeton University Press, Princeton.

Strauß, A., Henry, M.J., Scharinger, M., Obleser, J., 2015. Alpha phase determines successful lexical decision in noise. J. Neurosci. 35, 3256–3262. doi:10.1523/JNEUROSCI.3357-14.2015.

Tallon-Baudry, C., Bertrand, O., 1999. Oscillatory gamma activity in humans and its role in object representation. Trends Cogn. Sci. 3, 151–162. doi:10.1016/S1364-6613(99)01299-1.

Tallon-Baudry, C., Bertrand, O., Delpuech, C., Pernier, J., 1996. Stimulus specificity of phase-locked and non-phase-locked 40Hz visual responses in human. J. Neurosci. 16 (13), 4240–4249. doi:10.1523/JNEUROSCI.16-13-04240.1996.

Tort, A.B.L., Kramer, M.A., Thorn, C., Gibson, D.J., Kubota, Y., Graybiel, A.M., Kopell, N.J, 2008. Dynamic cross-frequency couplings of local field potential oscillations in rat striatum and hippocampus during performance of a T-maze task. Proc. Natl. Acad. Sci. USA 105, 20517–20522. doi:10.1073/pnas.0810524105.

Tort, A.B.L., Komorowski, R.W., Manns, J.R., Kopell, N.J., Eichenbaum, H, 2009. Theta-gamma coupling increases during the learning of item-context associations. PNAS 106, 20942–20947. doi:10.1073/pnas.0911331106.

Tort, A.B.L., Komorowski, R., Eichenbaum, H., Kopell, N, 2010. Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. J. Neurophysiol. 104, 1195–1210. doi:10.1152/jn.00106.2010.

Tort, A.B.L., Brankačk, J., Draguhn, A, 2018. Respiration-entrained brain rhythms are global but often overlooked. Trends Neurosci. 41, 186–197. doi:10.1016/j.tins.2018.01.007.

VanRullen, R., 2016. How to evaluate phase differences between trial groups in ongoing electrophysiological signals. Front. Neurosci. 10, 426. doi:10.3389/fnins.2016.00426.

Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., Pennartz, C.M.A, 2010. The pairwise phase consistency: a bias-free measure of rhythmic neuronal synchronization. Neuroimage 51, 112–122. doi:10.1016/j.neuroimage.2010.01.073.

Voytek, B., Kramer, M.A., Case, J., Lepage, K.Q., Tempesta, Z.R., Knight, R.T., Gazzaley, A., 2015. Age-related changes in 1/f neural electrophysiological noise. J Neurosci 35, 13257–13265. doi:10.1523/JNEUROSCI.2332-14.2015.

Whitlock, M.C., 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J. Evol. Biol. 18, 1368–1373. doi:10.1111/j.1420-9101.2005.00917.x.

Wolpert, N., Rebollo, I., Tallon-Baudry, C., 2020. Electrogastrography for psychophysical research: practical considerations, analysis pipeline, and normative data in a large sample. Psychophysiology 57, e13599. doi:10.1111/psyp.13599.

Wyart, V., de Gardelle, V., Scholl, J., Summerfield, C., 2012. Rhythmic fluctuations in evidence accumulation during decision making in the human brain. Neuron 76, 847–858. doi:10.1016/j.neuron.2012.09.015.

Zar, J.H., 2010. Biostatistical Analysis. (Prentice Hall).

Zaykin, D.V., Zhivotovsky, L.A., Czika, W., Shao, S., Wolfinger, R.D., 2007. Combining p-values in large scale genomics experiments. Pharm. Stat. 6, 217–226. doi:10.1002/pst.304.

Zoefel, B., Heil, P., 2013. Detection of near-threshold sounds is independent of EEG phase in common frequency bands. Front. Psychol. 4. doi:10.3389/fpsyg.2013.00262.

Zoefel, B., Davis, M.H., Valente, G., Riecke, L., 2019. How to test for phasic modulation of neural and behavioral responses. Neuroimage 202, 116175. doi:10.1016/j.neuroimage.2019.116175.