Individual differences do not mask effects of unconscious processing

Itay Yaron[1], Nathan Faivre[2], Liad Mudrik[1,3], & Matan Mazor[4,5]

[1] Sagol School of Neuroscience, Tel Aviv University

[2] University Grenoble Alpes, University Savoie Mont Blanc, CNRS, LPNC

[3] School of Psychological Sciences, Tel Aviv University

[4] Department of Psychological Sciences, Birkbeck, University of London

[5] Wellcome Centre for Human Neuroimaging, UCL

Author note

Correspondence concerning this article should be addressed to Itay Yaron, Haim Levanon 55, Tel Aviv, Israel. E-mail: mufc.itay@gmail.com.

All datasets and analysis scripts used in this manuscript are publicly available online: https://github.com/mufcItay/NDT.

Abstract

A wave of criticisms and replication failures is currently challenging claims about the scope of unconscious perception and cognition. Such failures to find unconscious processing effects at the population level may reflect the absence of individual-level effects, or alternatively, the averaging out of individual-level effects with opposing signs. Importantly, only the first suggests that consciousness may be necessary for the tested process to take place. To arbitrate between these two possibilities, we tested previously collected data where unconscious processing effects were not found (26 effects from 470 participants), using four Bayesian and frequentist tests that are robust to individual differences in effect signs. By and large, we found no reliable evidence for unconscious effects being masked by individual differences. In contrast, when we examined 136 non-significant effects from other domains, a novel non-parametric sign consistency test did reveal effects that were hidden by opposing individual results, though as we show, some of them might be driven by design-related factors. Taken together, four analysis approaches provide strong evidence for the restricted nature of unconscious processing effects not only across participants, but also across different trials within individuals. We provide analysis code and best-practice recommendations for testing for non-directional effects.

Keywords: unconscious processing; individual differences; consciousness

Individual differences do not mask effects of unconscious processing

Introduction

Our brains simultaneously perform complex information processing functions and yet, at any given moment in time, only a small subset of these functions is accompanied by conscious experience. This raises the question: which brain functions depend on consciousness, and which functions can take place without it?

One approach to investigating the scope and limits of unconscious processing is to measure the effect of different stimulus features on behaviour, while making sure that the stimulus itself is not consciously perceived (for review, see Kouider & Dehaene, 2007; Reingold & Merikle, 1988). If a stimulus feature affects behaviour even when the participant is not aware of the stimulus, being conscious of the stimulus cannot be necessary for processing that feature.

For example, Vorberg and colleagues (2003) studied the role of consciousness in motor preparation. In a series of experiments, they presented an arrow stimulus (henceforth, the prime) which was followed by an arrow-shaped metacontrast target stimulus, rendering it invisible (Breitmeyer & Ganz, 1976; see Figure 1A). Unconscious motor preparation priming was demonstrated by showing that participants responded faster to the target stimulus when its direction was congruent with the direction of the prime. This suggests that the direction of the prime has been represented unconsciously, triggering a motor plan. In similar studies, participants were reported to unconsciously perform arithmetic operations (Sklar et al., 2012), extract and integrate word meanings (Damian, 2001; Gaal et al., 2014; Sklar et al., 2012), or scenes and objects (Mudrik et al., 2011), detect errors (Charles et al., 2013), and exert inhibition over responses (Gaal et al., 2008) to stimuli that were masked from awareness. Findings of high-level

processing in the absence of consciousness served to inform and reform theories of consciousness (Dehaene & Naccache, 2001; Lamme, 2020; Lau & Rosenthal, 2011; Oizumi et al., 2014).

However, more recent work has called into question some of these previous findings and their interpretations. First, many of the original results do not replicate when tested in independent samples of participants (using direct replications, e.g., Biderman & Mudrik, 2018; Moors & Hesselmann, 2019; Stein et al., 2020, or conceptual replications, e.g., Hesselmann et al., 2015, 2016; Rabagliati et al., 2018). Second, some of these findings might be driven by residual consciousness in a subset of trials due to unreliable awareness measures (Meyen et al., 2022; Moors & Hesselmann, 2018; Rothkirch & Hesselmann, 2017; Shanks, 2017; Zerweck et al., 2021). Indeed, when re-analyzed to properly control for this possibility, some of these effects disappear (Meyen et al., 2022; Shanks, 2017). As a result, the scientific pendulum seems to be receding back to a narrower account of unconscious processing, consistent with a functional role of consciousness in most aspects of cognition (Balota, 1986; Moors et al., 2017; Peters et al., 2017).

Yet, the field is still far from reaching a consensus regarding the scope and limits of unconscious processing. Although progress has already been made in recent years toward improving methodology in unconscious processing studies, revealing the functional role of consciousness in cognition and perception remains difficult. Here we consider a largely neglected limitation of unconscious processing studies: by focusing on the average of signed (i.e., directional) single-participant summary statistics (for example, subtraction of reaction times between two conditions), previous investigations require not only that unconscious processing should leave a trace on behaviour, but also that this trace should be qualitatively similar across different participants (i.e., that the experimental manipulation would affect most participants in

the same direction). We note that though this second requirement is intuitive, it is orthogonal with the theoretical question at stake; our main concern is whether a given stimulus feature can affect behaviour in the absence of consciousness, yet this does not necessarily imply that it affects all participants in the same way. This way, previous analyses of unconscious processing may have been too *conservative*, potentially missing effects that happen to vary between different participants.

On the face of it, pronounced individual differences in unconscious processing effects on cognition seem possible, even likely. Indeed, previous investigations revealed heterogeneity in susceptibility to the attentional blink (Martens et al., 2006), in the effects of stimulus onset asynchrony (SOA) on metacontrast masking (Albrecht et al., 2010), and in the effects of visual imagery on conscious perception in a binocular rivalry setting (Dijkstra et al., 2019). Some qualitative differences have been linked to variability in processing speed (Martens et al., 2006), genetics (Maksimov et al., 2013), and brain anatomy and physiology (Boy et al., 2010; Gaal et al., 2011). Critically, in other behavioural paradigms, unconscious stimuli had opposite effects on different participants. Bolger et al., (2019) showed that while most participants responded faster to upright faces in a breaking continuous flash suppression task (b-CFS; Jiang et al., 2007), some responded faster to upside-down faces. Other studies found that priming effects changed in magnitude and even flipped in sign as a function of SOA and visibility (Boy & Sumner, 2014; Schlaghecken & Eimer, 2004). On the other hand, in other studies individual differences in prime visibility were not correlated with the magnitude of priming effects (Albrecht et al., 2010; Boy & Sumner, 2014; but see Eimer & Schlaghecken, 2002). Taken together, it is not clear if, and to what extent, unconscious effects are subject to meaningful individual variability. If they do, then some previously reported null results might actually be true effects, masked by such variability.

The paper proceeds as follows. We first simulate a setting where a strong effect of unconscious processing on behaviour is entirely missed in standard analysis, due to pronounced inter-individual differences. We then show that the same effect is revealed when using three tests that are robust to population variability: Bayesian prevalence analysis (Ince et al., 2021, 2022), Bayesian hierarchical modelling (Haaf & Rouder, 2019), and a frequentist test based on analysis of variance (ANOVA; Miller & Schwarz, 2018). When applied to data gathered from eight unconscious processing studies (reporting 26 non-significant effects), the same three tests support the null hypothesis according to which the behaviour of individual participants is unaffected by unconscious perception. This strengthens claims for a true absence of an effect in these studies. Finally, we propose a non-parametric alternative that provides improved sensitivity and specificity, avoiding potentially unjustified statistical assumptions regarding the data-generating process. Our test successfully reveals effects on multisensory integration, visual search, and confidence ratings that could not be detected using standard directional analysis. However, similar to the three other approaches, it reveals no effects when applied to the studies of unconscious processing examined here. We conclude that existing data are most consistent with the absence of influences of unconscious stimuli on cognition and perception, not only at the population, but also at the single-participant level.

Simulating non-directional unconscious effects

To demonstrate how true causal effects of unconscious processing can be masked by inter-individual differences in effect signs, we simulated a typical experiment using a within-participants manipulation (Figure 1). Specifically, we generated trial-by-trial data from a standard unconscious priming experiment. For each simulated participant, we generated reaction time data from two conditions (corresponding to congruent and incongruent primes in unconscious

processing studies). Individual-level effect sizes (in milliseconds) were sampled from a normal distribution centred at zero ($e_i \sim \mathcal{N}(0, \sigma_b)$, where e_i denotes the true effect size of the i^{th} participant and σ_b the between-participant standard deviation. Then, the trial-by-trial reaction times (RTs) of each participant and condition were generated according to each participant's true effect score (e_i), the relevant condition ($c \in \{1,0\}$, where $c = 1$ denotes the incongruent condition, and $c = 0$ denotes the congruent condition), and the within-participant standard deviation (σ_w) ($RT_{i,c} \sim \mathcal{N}(0, \sigma_w) + c * e_i$).

In two simulations, we manipulated two factors: the between-participant standard deviation (SD) over effect sizes (σ_b), and the within-participant SD over RTs within each condition (σ_w). This resulted in two distinct scenarios under this framework: (1) a *qualitative* or *non-directional differences* scenario, where all individuals show an effect, but individual-level effects largely vary in magnitude and sign (σ_b=15, σ_w=30; Figure 1B)), and (2) a *global null* scenario (Allefeld et al., 2016; Nichols et al., 2005), where no single participant is affected by the experimental manipulation (σ_b=0, σ_w= 100; Figure 1C). We simulated N_t=200 trials from N_p=15 participants per scenario, noting that the general principle holds for other sample sizes and number of trials.

First, we analyzed this simulated data using a two-sided paired t-test on the differences in mean RTs between the two conditions. This is the standard protocol for testing if unconscious processing took place. In both simulations, we obtained a null result, revealing no evidence for a difference in RT between the congruent and incongruent conditions (*non-directional differences*: $M = 0.13$, 95% CI $[-9.28, 9.54]$, $t(14) = 0.03$, $p = .977$; *global null*: $M = -4.42$, 95% CI $[-13.75, 4.90]$, $t(14) = -1.02$, $p = .326$). Importantly, in the *non-directional differences* simulation, all participants were affected by the experimental manipulation (that is, their true

effect sizes were different from zero). Thus, this commonly used approach systematically misses true causal effects of the experimental manipulation whenever they are inconsistent between participants.

To reiterate, a standard t-test misses existing individual-level effects because, operating on individual-level summary statistics, it is oblivious to within-participant variability in the dependent variable. In recent years, researchers sought to address this limitation, advocating for the use of statistical methods that incorporate both within and between-participant variability. Specifically, three approaches were proposed. First, the *prevalence Bayesian test* approach (Ince et al., 2021; Ince et al., 2022; henceforth PBT) estimates the prevalence of individual-level effects in a given population (the proportion of individuals showing an effect). The prevalence approach relies on a two stages procedure in which effects are tested at the individual level using a standard hypothesis-testing approach, and the true proportion of significant effects is then estimated using a Bayesian parameter estimation procedure. Second, the *qualitative individual differences* approach (Rouder and Haaf, 2020; Rouder and Haaf, 2021; Haaf and Rouder, 2019; henceforth QUID) quantifies the relative support for the presence of "qualitative differences" in effects, that is, inter-individual differences in effect signs, by performing a Bayesian model comparison over a family of hierarchical models with different constraints (Haaf & Rouder, 2019). Third, Miller & Schwarz (2018) introduce a parametric and frequentist test, based on ANOVA. Specifically, their Omnibus ANOVA test (henceforth OANOVA) tests the joint null hypothesis that there are no systematic differences between experimental conditions across individuals, or within individuals and across trials. Together, this is equivalent to the *global null* scenario we presented above.

We applied the tests to our simulated data, using the default priors from the original publications (Ince et al., 2021; Rouder & Haaf, 2021), and an implementation of the ANOVA model underlying the OANOVA test (Miller and Schwarz, 2018; see https://github.com/mufcItay/NDT for an R based implementation of the test). For PBT, we examined whether the lower bound of the prevalence HDI_{95} exceeded zero. We considered $BF > 3$ as evidence for an effect, $BF < \frac{1}{3}$ as evidence for no effect (*global null*), and values between these thresholds ($\frac{1}{3} \leq BF \leq 3$) as inconclusive (Jeffreys, 1998), and used an α of 0.05 for the OANOVA test. Reassuringly, all tests were able to differentiate between the two simulated scenarios, providing very strong evidence for an effect in the *non-directional differences* scenario, but not in the *global null* one. Specifically, According to PBT, about half of the population was estimated to show an effect of congruency on RT in the *non-directional differences* simulation (using a two-sided t-test for the individual-level test; HDI_{95}= [26, 74], MAP=51%; the 95% highest density interval and maximum a posteriori, respectively), but this proportion was not reliably different from zero in the *global null* simulation (HDI_{95}= [0, 24], MAP= 2%). Using the QUID method, a random effects model with individual-level effects was overwhelmingly preferred in the *non-directional differences* simulation (BF= 1.51e+33), but a null model was preferred in the *global null* simulation (BF= 0.08). Similarly, the OANOVA test revealed significant results in the *non-directional differences* scenario (F(15, 2970) = 13.95, p < .001), and a nonsignificant effect in the *global null* simulation (F(15, 2970) = 1.33, p =.177).

The simulations above demonstrate that adopting a non-directional approach, that is, an approach that takes into account the potential for opposite true effect signs among different participants, has the potential to reveal individual-level effects that would otherwise be missed due to high between-participant variability. Equipped with these validated tools, in the next

section we use the QUID, PBT, and OANOVA tests to ask whether null results in the field of unconscious processing are driven by such inter-individual variability, or alternatively, whether they reflect the true absence of a causal effect.

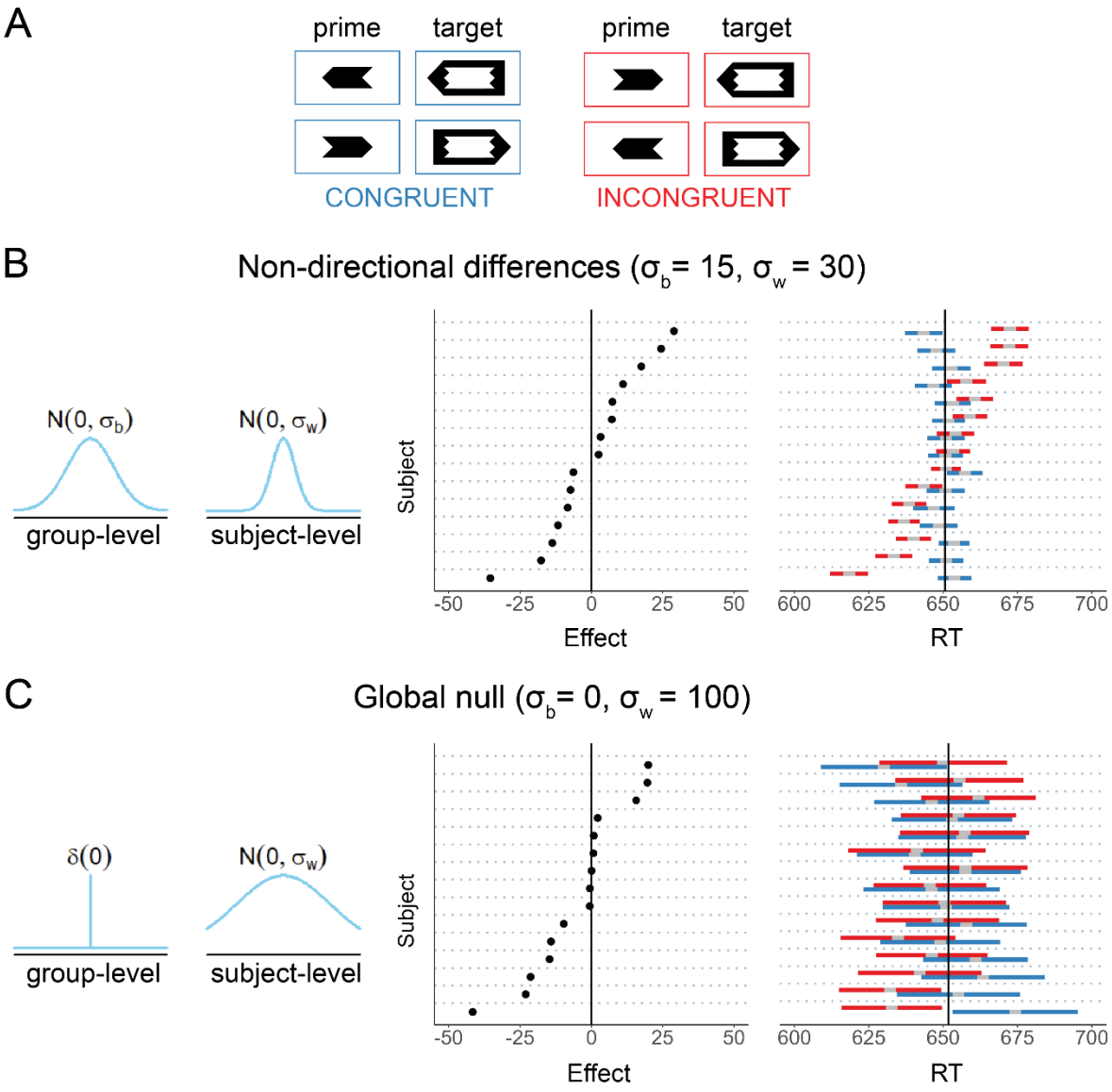


Figure 1. Simulated data demonstrating how true effects of unconscious priming can be masked by heterogeneity at the population level. Panel A: stimuli in a typical unconscious processing experiment (based on Vorberg and colleagues, 2003). Participants make speeded decisions about a consciously perceived target stimulus (for example, the direction of an arrow: right or left). The presentation of the

target stimulus is preceded by a prime stimulus, which is masked from awareness. Decision time is measured as a function of prime-target agreement: congruent (blue) or incongruent (red). Panels B, C: Left: simulation parameters controlling the within (σ_w) and between (σ_b) participants SD. Right: the results that were generated using the simulation parameters. Each point depicts the measured individual-level summary statistics for the difference between the mean RTs of each condition (congruent and incongruent), and the blue and red segments depict the 95% confidence interval (CI) around the average of RTs (the grey segment in the middle of each CI) in the congruent and incongruent conditions, respectively. A constant of 650ms was added to the RTs in both panels for presentation purposes. Panel B: a *non-directional differences* scenario (simulated using the parameters σ_b=15, σ_w=30). Panel C: a *global null* scenario (no effect of the experimental manipulation; simulated using the parameters σ_b=0, σ_w=100). Since standard directional tests rely on individual-level summary statistics, they cannot arbitrate between the scenarios described in the two panels.

Reexamining unconscious effects

To examine whether inter-individual differences masked true unconscious priming effects in previously reported studies, we collected and tested data from eight studies that reported null results (Benthien & Hesselmann, 2021; Biderman & Mudrik, 2018; Faivre et al., 2014; Hurme et al., 2020; Skora et al., 2021; Stein & Peelen, 2021; Zerweck et al., 2021; Chien et al. 2022; all datasets and analysis scripts are publicly available online: https://github.com/mufcItay/NDT).

We collected data associated with 26 null effects (see Supplementary Table 1 for details about all effects), 19 focusing on differences in RT and 7 on differences in signal detection sensitivity, d' (Green & Swets, 1966). We used the criteria set by the original authors for demonstrating unawareness (e.g., using objective and/or subjective measures of awareness), and a

two-sided non-parametric sign-flipping test for filtering out significant priming effects[1]. Finally, we excluded participants with fewer than five trials per experimental condition and/or zero variance in the dependent variable (e.g., when accuracy was measured). Together, these data allowed us to reexamine null unconscious processing effects using a non-directional approach that takes into account the potential for differences in effect signs when testing for group-level effects. We accordingly asked whether true effects of unconscious processing were masked by population heterogeneity in effect signs.

To that end, the effects of interest were tested using PBT, QUID, and the OANOVA tests (see Supplementary Figure 1 for an analysis of the significant directional effects which were excluded). PBT was applied to all 26 effects. In contrast, QUID and OANOVA were used on subsets of 20 and 21 of these effects, respectively (omitting five effects of signal detection sensitivity, d', from both tests, and one additional RT interaction from the QUID analysis, as its current implementation only supports simple RT effects). All tests agreed on finding no reliable evidence for non-directional unconscious effects. According to PBT, the MAP prevalence statistic was zero in 76.92% of the effects (maximal MAP = 11.36%; see Fig. 2A), and the 95% HDI included zero in all of them. Similarly, for both QUID and the OANOVA tests, no single BF or p-value revealed evidence for an effect (maximal BF_{10} = 0.80 and all p-values > 0.05; see Figure 2B, C). Notably, QUID obtained moderate evidence for the *global null* model in 70% of the cases (see Fig. 2B). The remaining effects were inconclusive. Hence, for the effects collected here, in the case of unconscious processing, the three tests revealed a highly similar pattern of

[1] Across all RT effects, our analysis used raw RT scores, and thus our results diverged from the original results when log transformations were used (see the notes column in Supplementary Table 1 for details).

results, consistent with a strong interpretation of previously reported null results as revealing the genuine absence of a causal effect of unconsciously perceived stimuli on behaviour.

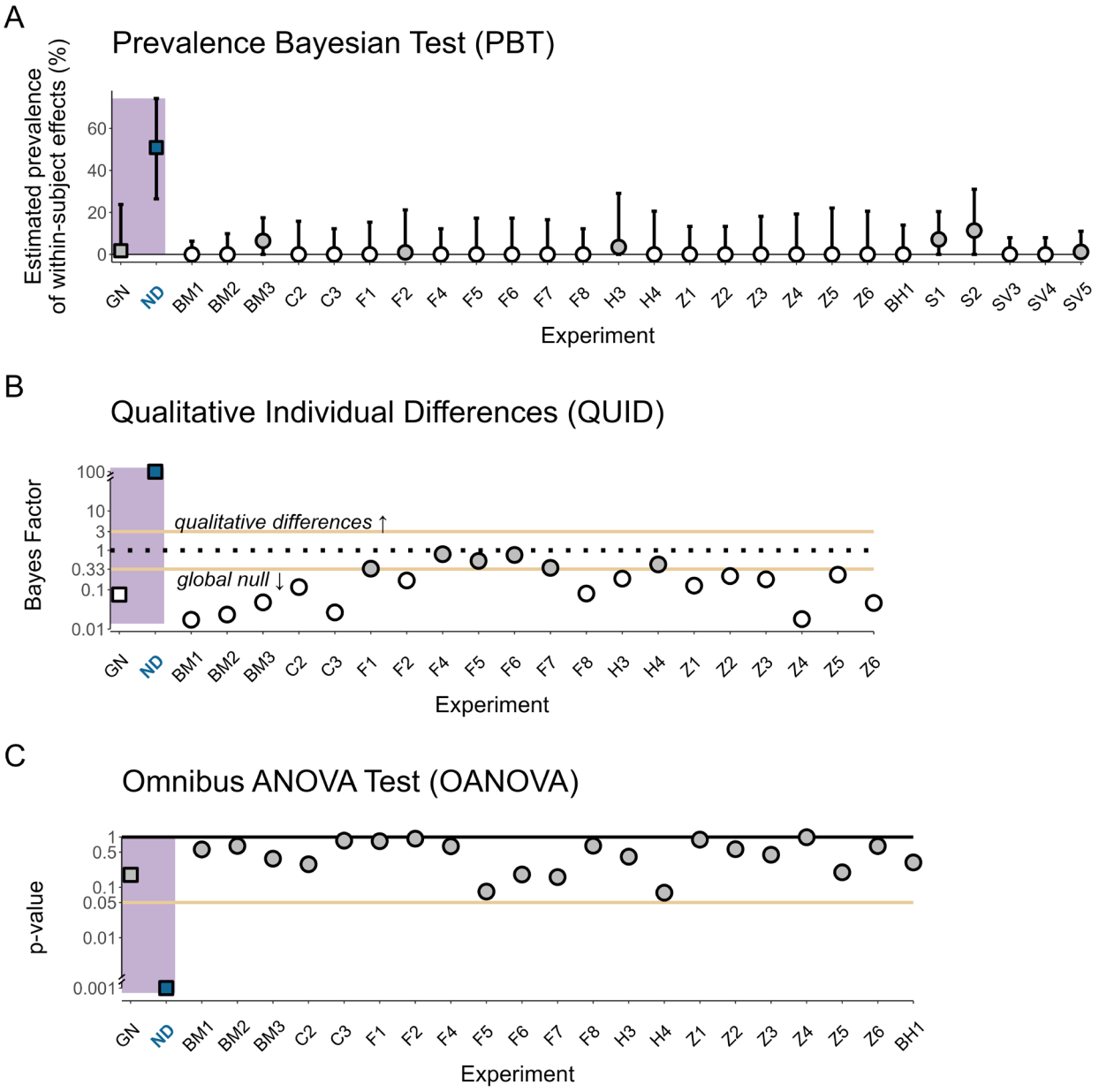


Figure 2. The results of applying the PBT (A), QUID (B), and OANOVA (C) tests to effects that produced null results in a non-parametric directional test and to simulated data (the Non-directional effect (ND)

and Global Null (GN) simulations described above, presented as square-shaped markers). Effect labels[2] appear on the x-axis. Panel A: the estimated prevalence of an unconscious effect in each of the cases, using PBT (Ince et al., 2021). Markers indicate the maximum a posteriori (MAP) for each of the effects, and segments depict the 95% highest density interval (HDI_{95}). Grey markers depict above-zero MAPs. Panel B: Bayes factors for the comparison between a random effects model that takes into account potential differences in effect signs and the global null model. White markers depict cases where moderate evidence for the global null model was found, while grey markers indicate inconclusive results. The dashed black line indicates a BF of 1 (no preference for either model), and the solid orange lines indicate a BF cutoff of 3. Panel C: p-values obtained by the OANOVA test (Miller & Schwarz, 2018). Blue and grey markers indicate significant and non-significant results, respectively. For illustration purposes, BF and significance values are presented on a logarithmic scale on the y-axis.

Yet, the reviewed approaches also have some limitations that make it harder to draw firm conclusions based on their results. First, in contrast to frequentist tests within the Null Hypothesis Statistical Testing (NHST) tradition, PBT and QUID provide no control over long-term error rates (the probability of finding a false positive result or missing a true result over an infinite number of tests, with the former being more critical to our point here). Such error control promises a much-needed 'fool-proof' method to infer the existence of unconscious processing effects without making too many mistakes in the long run (Lakens et al., 2020). To illustrate,

[2] Effect labels abbreviations (sorted alphabetically): BH = Benthien and Hesselmann (2021), BM = Biderman and Mudrik (2018), C = Chien et al. (2022), F = Faivre et al. (2014), H = Hurme et al. (2020), S = Skora et al. (2021), SV = Stein and Peelen (2021), Z = Zerweck et al. (2021). For all labels, numbers denote effect indices within each study (see Supplementary Table 1 for the full mapping between labels and effects).

while an alpha level of 0.05 guarantees that only one in 20 tests will generate a significant result when there is no true difference between the conditions, using a Bayes Factor cutoff of 3, or an HDI of 95%, provides no such guarantee.

Second, both the model comparison approach used in QUID and the OANOVA test necessarily assume a parametric model of the data, making specific assumptions of normality and equal within-individual variance. In simulations, we find that violations of this second assumption can have dramatic effects on the specificity and sensitivity of both tests (see Appendix A). This can be addressed by more complex models that are capable of handling different distribution families, but as model complexity grows, unwanted effects of assumption violations may become harder to spot and quantify. Hence, taking a non-parametric approach provides safer inferences when the form of the data-generating process is not fully known.

Lastly, PBT begins with testing the significance of effects at the single subject level, thereby dichotomizing a continuous test statistic into one bit of information: significant or not. This dichotomization results in information loss and introduces an additional free parameter — the individual alpha level. This step is well justified when estimating population prevalence, but it is unnecessary for our purpose of detecting a non-directional effect at the population level. As we describe below, using a continuous participant-level statistic makes our test more sensitive (see Appendix B for a direct comparison between the two approaches).

In the next section, we introduce a novel non-directional test that takes into account population heterogeneity to infer group-level effects. The test is both frequentist and non-parametric, which addresses the above issues. Similarly to the OANOVA test it promises a control for long term error-rates, but unlike it, our test does not assume a parametric model of the data-generating process. Using a continuous within-participant summary statistic, it is also more

statistically powerful than approaches that focus on a dichotomous notion of effect prevalence (see Appendix B).

Sign Consistency: a non-parametric test that is robust to qualitative differences

Our test assumes that a within-participant effect is convincing if it is consistently evident across different trials. To test this, we can split the trials of an individual into two random halves, and ask whether both halves show the same qualitative effect (e.g., the performance in the congruent condition is higher than in the incongruent condition; see Fig. 3A). By doing this many times, we can measure how often the two halves agree. Following this strategy, we estimate the consistency of effect signs within each individual by measuring the frequency of consistent results across splits. Then, we compare the group-mean consistency score against a null distribution: 10,000 samples of group-level consistency scores, obtained after randomly shuffling the experimental condition labels within participants (here, to speed up the computational process, for each participant, 25 permutations were created, from which we randomly sampled a single permutation in each null distribution sample; Stelzer et al., 2013). Hence, our null distribution reflects the expected consistency of within-participant effects when there is no link between the experimental condition and the dependent variable. An easy-to-use implementation of the sign consistency test is available as part of the signcon R package (https://github.com/mufcItay/signcon).

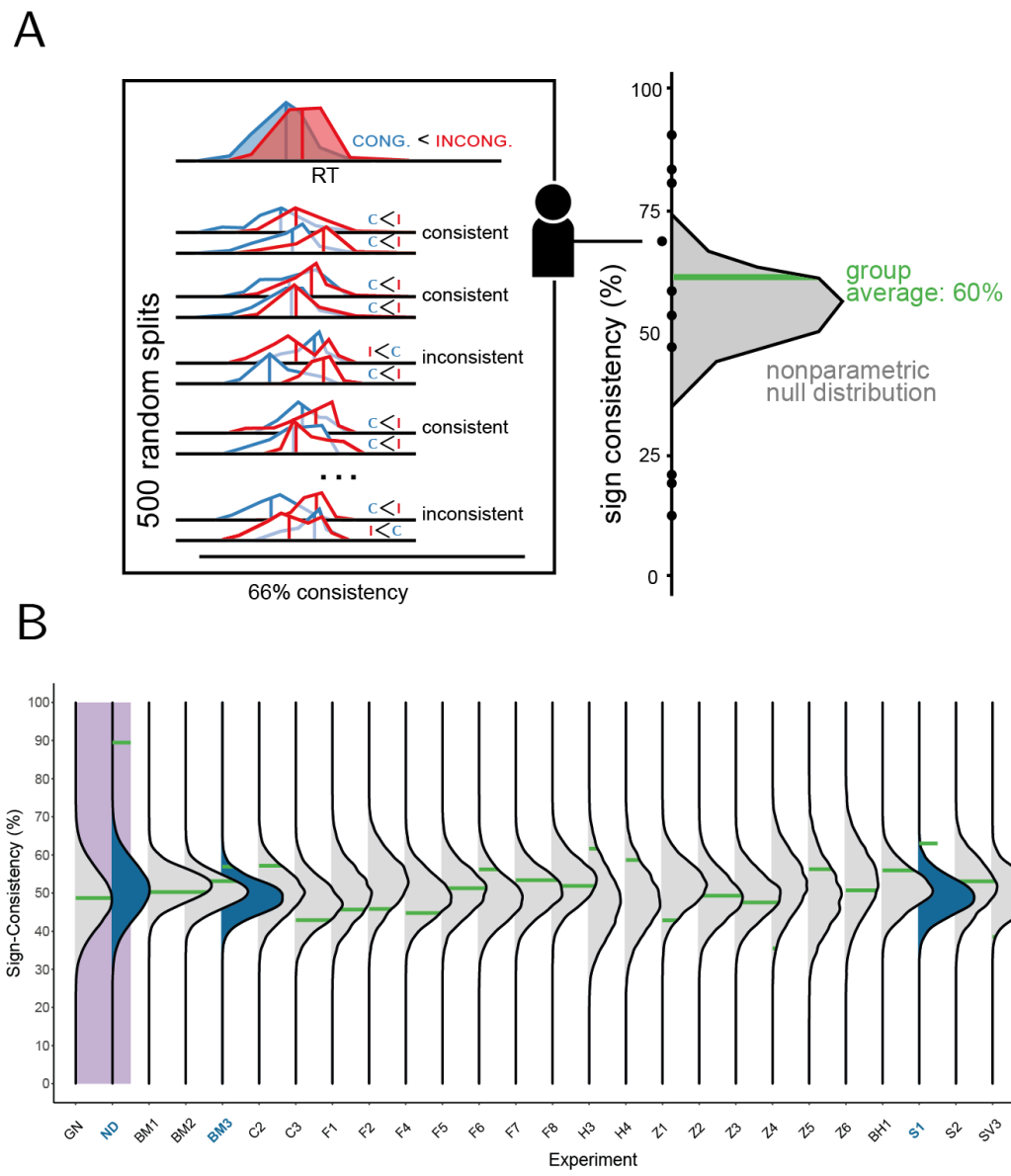


Figure 3. A frequentist, non-parametric, test for sign consistency. Panel A: a schematic illustration, using the same conventions as in Figure 1 (C = congruent, I = incongruent). Participant-wise sign consistency is quantified as the proportion of random splits of experimental trials, for which both halves display the same qualitative effect (C>I or I>C). Group-level sign consistency is compared against a non-parametric null distribution to obtain a significance value. The left panel illustrates a subset of random splits from a hypothetical participant. The upper row illustrates the overall RT data for that participant, and each row below shows one split of the data, where for each half we compare the mean of congruent and incongruent RT distributions, to test if the direction of the difference in the two halves is consistent or not. Then, to determine if the group shows evidence for non-directional effects, the averaged consistency score *across participants* (plotted in green), which is the proportion of consistent splits

across all splits, is compared to the null distribution (right panel). In this hypothetical case, the group does not show an effect, as the average score is well within the null distribution. Panel B: the results of applying the sign consistency test to effects that produced null results in a non-parametric directional test (N = 26). Significant results, for which the estimated mean sign consistency score is greater than 95% of the null distribution, are marked in blue. As in Figure 2, the x-axis lists effect labels.

We quantified the average within-participant sign consistency of effects for which a directional test did not produce significant results (the same effects reported in Fig. 2). For each individual, sign consistency was defined as the percentage of consistent signs across 500 random splits. Effect scores were calculated using a predefined summary function (i.e., taking the average RT or calculating signal detection sensitivity, d', in each condition, depending on the effect of interest). The results revealed a similar picture to the one provided by the Bayesian approaches (see Figure 3B). The vast majority of cases did not show significant sign consistency, with two exceptions: first, an effect of an unconsciously presented cue on wagering decisions (Skora et al., 2021; M = 63%, p = .001), and second, a scene-object congruency effect (Biderman and Mudrik, 2018; M = 57%, p = .027). Although the two effects were not detected by PBT, the prevalence MAP estimate was above zero for both (7.12% and 6.43%, respectively). Thus, despite some evidence for sign consistency, the overall picture remained the same, hinting at minimal qualitative inter-individual differences in unconscious processing.

Together, four different analysis methods support the conclusion that by and large, unconscious priming effects are not masked by individual differences. Yet one can still claim that these statistical tests are simply not sensitive enough to detect qualitatively variable, non-directional effects, even when those exist. To test this claim, and demonstrate how these methods can be used to reveal such hidden effects in other fields, we collected additional, openly accessible, datasets from studies conducted in different fields of research within experimental psychology. We then used our non-parametric test on these datasets, demonstrating its potential

benefit in determining whether a null result at the group level hides true, but variable, effects at the individual participant level.

Testing within-participant sign consistency across experimental psychology studies

We used the sign consistency test to expose hidden effects that were not revealed by standard directional tests in various fields of research. We collected data from different open-access databases (the Confidence Database (Rahnev et al., 2020), the Reproducibility Project (Open Science Collaboration, 2015), and the Classic Visual Search Effects open dataset (Adam et al., 2021)). We also used social media to ask for previously reported null effects. We had two inclusion criteria: first, to estimate subject-level sign consistency, the independent variable had to be manipulated within single participants. And second, we filtered out effects that were significant in a non-parametric, directional test. Overall, we collected data associated with 136 nonsignificant effects (121 from the Confidence Database, four from the Reproducibility Project, eight from the Classic Visual Search Effects open dataset and three from the social media query). In all cases, participants were excluded for having fewer than five trials per experimental condition and/or zero variance in the dependent variable.

We grouped the different effects into three categories, according to research topics and the analysis we used to test them: first, we tested for effects of participants' responses in 2-alternative forced choice tasks on their confidence ratings in all datasets from the Confidence Database (Rahnev et al., 2020; retrieved on 23/1/2023), by comparing the mean confidence ratings between two different responses. Second, we used the same Confidence Database datasets to test for metacognitive sensitivity effects of response. Metacognitive sensitivity was quantified as the area under the response-conditional type-2 Receiver Operating Characteristic curve (Meuwese et al.,

2014; here we also excluded datasets that did not include accuracy scores; the remaining 47 effects were analyzed). Third, we grouped effects from the Reproducibility Project (Open Science Collaboration, 2015), the Classic Visual Search Effects open dataset (Adam et al., 2021), and a single study from the social media query (Battich et al., 2021) under a more general "Cognitive Psychology" category. For these studies, we tested the sign consistency of the effect tested by the original authors (averaged difference or interaction effects).

Across the entire sample, including all analyzed effects (N = 136), most effects (63.24%) showed significant sign consistency. This trend was further explored within each category, revealing significant effects in 89.19% 27.66%, and 46.67%, of the Confidence, Metacognitive Sensitivity, and Cognitive Psychology effects (four visual search effects and all three effects from Battich et al., 2021), compared with only 7.69% of the unconscious processing effects, as reported above (see Figure 4). These results validate the potential of using sign consistency to reveal effects on cognition and perception. In striking contrast to the absence of hidden effects in the field of unconscious processing, we found compelling evidence for pronounced inter-individual differences that mask group-level effects in other domains.

However, special care should be taken when interpreting non-directional test results, and when designing experiments targeting non-directional effects (see Box A for best-practice recommendations). A case in point can be found in Battich et al. (2021), who examined the hypothesis that joint attention affects multisensory integration. Critically, this hypothesis was tested by comparing two social conditions that were counterbalanced across participants, such that for half of the participants a joint attention condition was performed before a baseline condition where participants performed the same task individually, and vice versa for the other half. As a result, contrasting the two conditions within participants is identical to contrasting

early and late trials. Thus, although the interaction between social condition and multisensory integration showed significant sign consistency ($M = 62.34\%$, $p < .001$, $M = 58.54\%$, $p < .001$, and $M = 66.62\%$, $p < .001$, for the three effects that showed sign consistency, paralleled by null results according to directional analysis), we cannot unambiguously interpret these results as suggesting a causal, non-directional, effect of the social manipulation. This is because, under this design, the social setting condition and the order of experimental conditions are perfectly correlated within individual participants, rendering both potential drivers behind the observed effect.

Similarly, the great majority of experiments in the Confidence Database showed significant non-directional effects of response on confidence, such that individual participants were more confident in making one response or the other. Here, order effects are not a concern, as the two responses are expected to be equally distributed within a block. However, since stimulus-response mapping was not counterbalanced within participants, we are unable to tell whether these effects reflect individual differences in stimulus preferences (e.g., enhanced sensory encoding for right-tilted gratings) or in response priming (e.g., reports of high confidence are primed by reporting a decision with the right finger).

As a general principle, counterbalancing of confounding experimental variables can be done either between participants (for example, using a different response-mapping for odd and even participants) or within participants (for example, changing the response-mapping between experimental blocks for all participants). While both approaches are effective in protecting against confounding of the mean tendency of the dependent measures, only within-subject counterbalancing is effective when testing for non-directional effects. Accordingly, unless all confounding variables (e.g., condition order or response-mapping) are randomized within

participants, the interpretation of non-directional effects cannot be uniquely linked to causal effects of the experimental manipulation.

Importantly, although we cannot conclusively attribute these non-directional effects to social setting versus condition order in the first example, or to response versus stimulus in the second, they both constitute examples of true effects that were masked by inter-individual differences. The absence of a directional effect in Battich et al is indicated by the fact that on average, participants showed similar levels of multisensory integration in the first and second parts of the experiment ($M_D = 0.03$, 95% CI $[-0.02, 0.09]$, $t(48) = 1.22$, $p = .228$, $M_D = 0.03$, 95% CI $[-0.01, 0.08]$, $t(48) = 1.47$, $p = .149$, and $M_D = -0.02$, 95% CI $[-0.06, 0.03]$, $t(48) = -0.69$, $p = .493$). In the case of confidence effects, response mapping was not counterbalanced across participants in many of the considered datasets. This way, the absence of a directional effect of response is also indicative of the absence of a directional effect of stimulus. Together, these previously hidden non-directional findings make the absence of significant non-directional effects in unconscious processing a more convincing indication of the true absence of such effects at the individual-participant level.

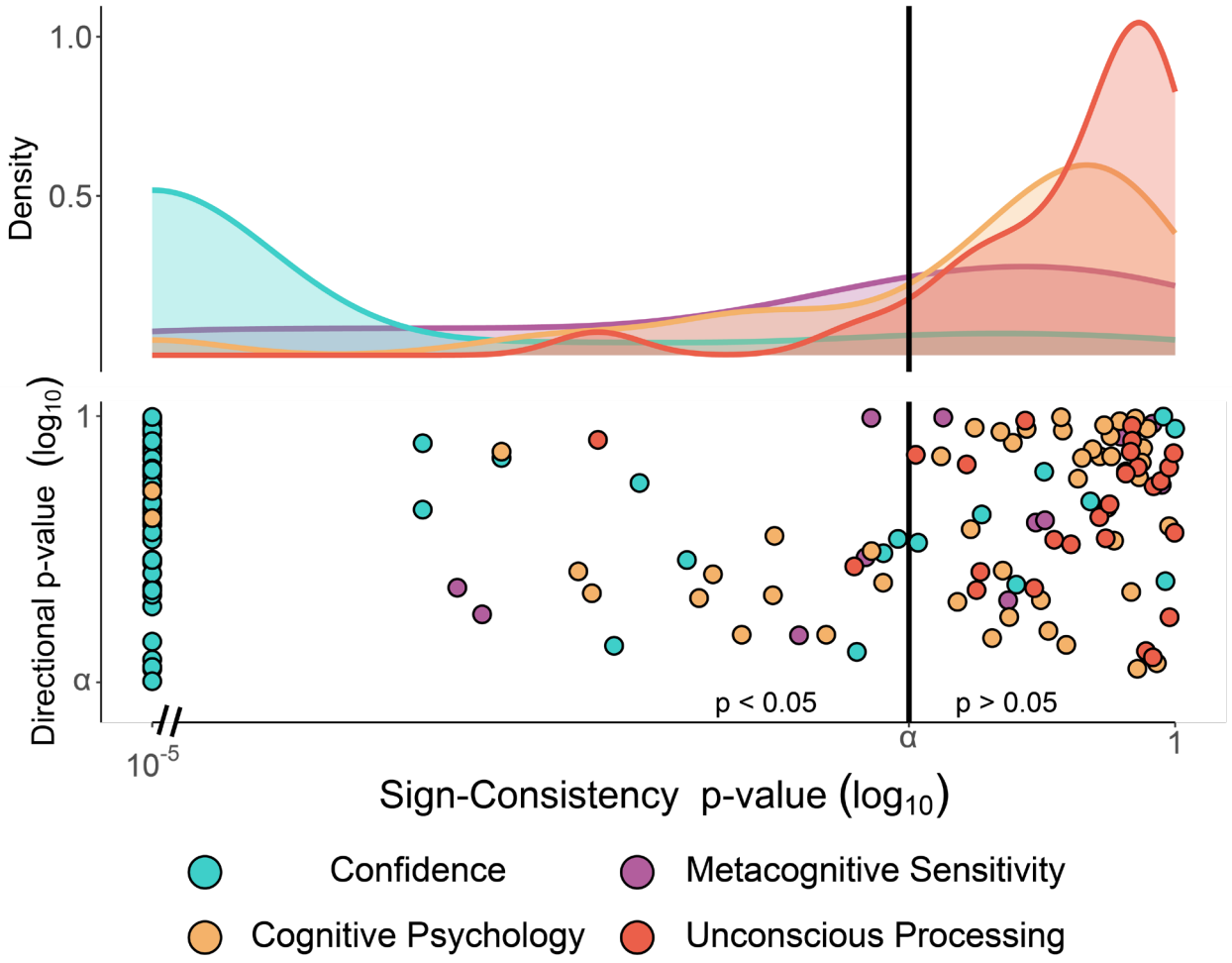


Figure 4. Sign consistency test results for null directional effects from different cognitive psychology fields. Turquoise and purple markers indicate the outcomes for datasets from the Confidence Database (Rahnev et al., 2020) that were analysed to reveal differences in confidence and metacognitive sensitivity between responses, respectively. Orange markers indicate the outcomes for effects from various cognitive psychology studies. Finally, for comparison purposes, we also plot here in red the results of the studies on unconscious processing (N=26), reported in the previous section. Lower panel: each point depicts the log_{10} transformed p-values obtained by the sign consistency test (x-axis) and a directional sign-flipping test (y-axis; datasets were filtered to exclude significant directional effects, hence the minimal directional p-value for all datasets is $\alpha = .05$). Upper panel: The p-values density distributions that summarize the results in the lower panel for datasets in each field.

Discussion

What is the scope and depth of unconscious processing? Previous claims about high-level unconscious processing effects have recently been criticized for methodological reasons (Meyen et al., 2022; Rothkirch & Hesselmann, 2017; Shanks, 2017), and for lack of replicability (Hesselmann et al., 2015; Moors et al., 2016; Moors & Hesselmann, 2018; Stein et al., 2020). Here, we point out that testing for effects that are consistent across individuals may be overly conservative for the question at stake. Instead, we examined if these null results might still be underlied by an effect, yet a non-directional one. That is, we tested the hypothesis that individual differences in unconscious processing mask true unconscious effects in individual participants. Adopting a non-directional approach that is robust to inter-individual differences in effects, we used two Bayesian tests (Ince et al., 2021; Rouder & Haaf, 2021), an ANOVA-based test (Miller & Schwarz, 2018), and a novel non-parametric frequentist test. We examined previously reported non-significant results (N = 26), and showed they cannot be explained by inter-individual differences in effects. All tests converged on a similar picture: besides two effects that were picked up by one of the four methods, unconscious processing effects were not masked by substantial inter-individual differences.

It is important to note that our claim here is not about the presence of individual differences in unconscious processing in general, but about the likelihood that such differences in effect signs may be responsible for null group-level findings. Indeed, previous studies revealed inter-individual differences in the magnitude, not the sign, of unconscious processing effects (Boy et al., 2010; Cohen et al., 2009; Gaal et al., 2011). For example, Van Gaal et al. (2011) used fMRI and a meta-contrast masked arrows-priming task, to show that grey matter density is correlated with the size of unconscious motor priming effects. Yet importantly, in this experiment

effects were defined according to the assumption that *incongruent* trials are performed slower than *congruent* trials (trials where primes and targets pointed to opposing and the same direction, respectively). Here, in contrast, we asked whether relaxing the assumption of effect sign uniformity could reveal unconscious effects that remain undetected using standard directional approaches.

Overall, in two out of 26 effects, a sign consistency test detected an effect that was missed by a standard, directional test. However, even these two effects should be examined cautiously. The effect found for the third experiment in Biderman & Mudrik (2018) ($M = 57\%$, $p = .027$) was not detected by the three other tests and did not survive a correction for false discovery rate among unconscious processing effects (Benjamini & Hochberg, 1995). Hence, it is likely that this effect reflects a type-1 error. Similarly, while we found significant sign consistency ($M = 63\%$, $p = .001$) for a d' effect in the first experiment of Skora et al., (2021), the authors expressed concerns regarding possible contamination of their measured effect by conscious processing due to regression to the mean (Shanks, 2017). Thus, we suggest that our findings should be interpreted as suggesting no masking of unconscious processing effects by population heterogeneity.

While our focus here was on unconscious processing, a non-directional analysis approach can be useful in many fields of investigation where individual differences are expected. A null finding in a standard t-test or an ANOVA may indicate the true absence of an effect or a lack of statistical power, but it may also be driven by qualitative heterogeneity in participant-level effect signs. In the field of neuroimaging, the adoption of information-based, non-directional approaches famously revealed such effects that were otherwise masked by heterogeneity in neural activation patterns and fine brain structure (Gilron et al., 2017; Kriegeskorte & Kievit,

2013; Norman et al., 2006). In the context of this investigation, we found considerable evidence for cases where inter-individual differences mask group-level effects. These cases carry theoretical significance both in uncovering previously missed effects, and in revealing aspects of human cognition that are subject to considerable population variability (Bolger et al., 2019; Rouder & Haaf, 2020, 2021).

Previously, Rouder & Haaf (2021) suggested that such qualitative individual differences may be expected in preference or bias-based effects (e.g., Schnuerch et al., 2021; Rouder and Haaf, 2021), but not in effects that are driven by low-level perceptual and attentional processes. Consistent with this proposal, the absence of substantial evidence for variability in effect signs in unconscious processing was paralleled with strong evidence for such qualitative inter-individual differences in subjective confidence ratings (e.g., some participants are more confident in classifying a grating as oriented to the right, while others show the opposite preference)[3]. However, robust participant-level effects were masked by qualitative individual differences in other domains too, not all of them relate to higher-level preferences or biases. For example, non-directional effects of distractor presence were found in a series of visual search experiments (Adam et al., 2021; sign consistency > 59.34%, $p < .040$, for four out of eight measured effects). These findings echo the non-directional effects of distractor-target compatibility on action planning that were revealed using the OANOVA test (Miller & Schwarz, 2018), and were not

[3] As we note above, since in most of these studies responses and stimuli are closely correlated, these effects cannot be unambiguously attributed to stimulus preferences or response priming effects. Relatedly, more recent work reveals that such inter-individual differences in preference for specific responses or stimuli can be traced back to heterogeneity in sensory encoding Rahnev (2021)

detected when using standard directional analysis (in both Machado et al.,2007, and Machado et al., 2009). Thus, aside from shedding light on previous non-significant results, our preliminary findings inform previous claims regarding the plausibility of population heterogeneity in effect signs in perceptual and attentional effects in general, providing some indication that such effects may be more prevalent than previously assumed.

To facilitate the adoption of this non-directional approach in experimental psychology, we release with this paper an R package with a simple-to-use implementation of our error-controlled and non-parametric sign consistency test (https://github.com/mufcItay/signcon). We note that unlike directional tests, the validity of the sign consistency test (and more generally, non-directional tests) depends on counterbalancing of confounding variables not only across participants, but also across trials within a single participant. We recommend using this test to complement standard, directional tests, especially in interpreting null findings at the group level (see Box A for a more detailed description of best-practice recommendations for non-directional testing). This seems especially important in the field of unconscious processing, where null results are becoming more prevalent, and carry theoretical significance as hinting at possible functional roles for conscious processing.

Conclusions

Experimental demonstrations of unconscious processing have been reported for nearly 150 years now (e.g., Peirce and Jastrow, 1884), yet their reliability and robustness has repeatedly been put into question (e.g., Holender, 1986 and Shanks, 2017). Here, we examined the possibility that some of the findings against such processing, reporting null results, might hide effects at the individual level, yet in opposing directions. We employed four non-directional tests to re-

examine 26 null effects. Our findings suggest no role for individual differences in explaining non-significant effects at the group level. Furthermore, by expanding our exploration outside the domain of unconscious processing, we found compelling evidence for effects that were shadowed by individual differences in effect signs, nuancing views about the universality of cognitive and perceptual effects. We provide a user-friendly implementation of the non-directional sign consistency test, and recommend its use for interpreting null results.

References

Adam, K. C., Patel, T., Rangan, N., & Serences, J. T. (2021). Classic visual search effects in an additional singleton task: An open dataset. *Journal of Cognition*, *4*(1).

Albrecht, T., Klapötke, S., & Mattler, U. (2010). Individual differences in metacontrast masking are enhanced by perceptual learning. *Consciousness and Cognition*, *19*(2), 656–666.

Allefeld, C., Görgen, K., & Haynes, J. D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *Neuroimage*, *141*, 378–392.

Balota, D. A. (1986). Unconscious semantic processing: The pendulum keeps on swinging. *Behavioral and Brain Sciences*, *9*(1), 23–24.

Battich, L., Garzorz, I., Wahn, B., & Deroy, O. (2021). The impact of joint attention on the sound-induced flash illusions. *Attention, Perception, & Psychophysics*, *83*, 3056–3068.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological*, *57*(1), 289–300.

Benthien, F. M., & Hesselmann, G. (2021). Does location uncertainty modulate unconscious processing under continuous flash suppression? *Advances in Cognitive Psychology*, *17*(1), 3.

Biderman, N., & Mudrik, L. (2018). Evidence for implicit—but not unconscious—processing of object-scene relations. *Psychological Science*, *29*(2), 266–277.

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601.

Boy, F., Evans, C. J., Edden, R. A., Singh, K. D., Husain, M., & Sumner, P. (2010). Individual differences in subconscious motor control predicted by GABA concentration in SMA. *Current Biology*, *20*(19), 1779–1785.

Boy, F., & Sumner, P. (2014). Visibility predicts priming within but not between people: A cautionary tale for studies of cognitive individual differences. *Journal of Experimental Psychology: General*, *143*(3).

Breitmeyer, B. G., & Ganz, L. (1976). Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological Review*, *83*(1), 1.

Charles, L., Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, *73*, 80–94.

Chien, S. E., Chang, W. C., Chen, Y. C., Huang, S. L., & Yeh, S. L. (2022). *The limits of unconscious semantic priming* (pp. 1–12). Current Psychology.

Cohen, M. X., Gaal, S., Ridderinkhof, K. R., & Lamme, V. (2009). Unconscious errors enhance prefrontal-occipital oscillatory synchrony. *Frontiers in Human Neuroscience*, *54*.

Damian, M. F. (2001). Congruity effects evoked by subliminally presented primes: Automaticity rather than semantic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 154.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.

Dijkstra, N., Hinne, M., Bosch, S. E., & Gerven, M. A. J. (2019). Between-subject variability in the influence of mental imagery on conscious perception. *Scientific Reports*, *9*(1), 1–10.

Eimer, M., & Schlaghecken, F. (2002). Links between conscious awareness and response inhibition: Evidence from masked priming. *Psychonomic Bulletin & Review*, *9*(3), 514–520.

Faivre, N., Mudrik, L., Schwartz, N., & Koch, C. (2014). Multisensory integration in complete unawareness: Evidence from audiovisual congruency priming. *Psychological Science*, *25*(11), 2006–2016.

Gaal, S., Naccache, L., Meuwese, J. D., Loon, A. M., Leighton, A. H., Cohen, L., & Dehaene, S. (2014). Can the meaning of multiple words be integrated unconsciously? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641), 20130212.

Gaal, S., Ridderinkhof, K. R., Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2008). Frontal cortex mediates unconsciously triggered inhibitory control. *Journal of Neuroscience*, *28*(32), 8053–8062.

Gaal, S., Scholte, H. S., Lamme, V. A., Fahrenfort, J. J., & Ridderinkhof, K. R. (2011). Pre-SMA gray-matter density predicts individual differences in action selection in the face of conscious and unconscious response conflict. *Journal of Cognitive Neuroscience*, *23*(2), 382–390.

Gilron, R., Rosenblatt, J., Koyejo, O., Poldrack, R. A., & Mukamel, R. (2017). What's in a pattern? Examining the type of signal multivariate analysis uncovers at the group level. *NeuroImage*, *146*, 113–120.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1, pp. 1969–2012). Wiley.

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*, 772–789.

Hesselmann, G., Darcy, N., Ludwig, K., & Sterzer, P. (2016). Priming in a shape task but not in a category task under continuous flash suppression. *Journal of Vision*, *16*(3), 17–17.

Hesselmann, G., Darcy, N., Sterzer, P., & Knops, A. (2015). Exploring the boundary conditions of unconscious numerical priming effects with continuous flash suppression. *Consciousness and Cognition*, *31*, 60–72.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences*, *9*(1), 1–23.

Hurme, M., Koivisto, M., Henriksson, L., & Railo, H. (2020). Neuronavigated TMS of early visual cortex eliminates unconscious processing of chromatic stimuli. *Neuropsychologia*, *136*, 107266.

Ince, R. A., Kay, J. W., & Schyns, P. G. (2022). Within-participant statistics for cognitive science. In *Trends in cognitive sciences*.

Ince, R. A., Paton, A. T., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of population prevalence. *Elife*, *10*, 62461.

Jeffreys, H. (1998). *The theory of probability*. OuP Oxford.

Jiang, Y., Costello, P., & He, S. (2007). Processing of invisible stimuli: Advantage of upright faces and recognizable words in overcoming interocular suppression. *Psychological Science*, *18*(4), 349–355.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: A critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 857–875.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412.

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57.

Lamme, V. A. (2020). Visual functions generating conscious seeing. *Frontiers in Psychology*, *11*, 83.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373.

Machado, L., Devine, A., & Wyatt, N. (2009). Distractibility with advancing age and parkinson's disease. *Neuropsychologia*, *47*(7), 1756–1764.

Machado, L., Wyatt, N., Devine, A., & Knight, B. (2007). Action planning in the presence of distracting stimuli: An investigation into the time course of distractor effects. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(5).

Maksimov, M., Vaht, M., Harro, J., & Bachmann, T. (2013). Can common functional gene variants affect visual discrimination in metacontrast masking? *PloS One*, *8*(1), 55287.

Martens, S., Munneke, J., Smid, H., & Johnson, A. (2006). Quick minds don't blink: Electrophysiological correlates of individual differences in attentional selection. *Journal of Cognitive Neuroscience*, *18*(9), 1423–1438.

Meuwese, J. D., Loon, A. M., Lamme, V. A., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: Comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics*, *76*, 1057–1068.

Meyen, S., Zerweck, I. A., Amado, C., Luxburg, U., & Franz, V. H. (2022). Advancing research on unconscious priming: When can scientists claim an indirect task advantage? *Journal of Experimental Psychology: General*, *151*(1), 65.

Miller, J., & Schwarz, W. (2018). Implications of individual differences in on-average null effects. *Journal of Experimental Psychology: General*, *147*(3), 377.

Moors, P., Boelens, D., Overwalle, J., & Wagemans, J. (2016). Scene integration without awareness: No conclusive evidence for processing scene congruency during continuous flash suppression. *Psychological Science*, *27*(7), 945–956.

Moors, P., & Hesselmann, G. (2018). A critical reexamination of doing arithmetic nonconsciously. *Psychonomic Bulletin & Review*, *25*, 472–481.

Moors, P., & Hesselmann, G. (2019). Unconscious arithmetic: Assessing the robustness of the results reported by karpinski, briggs, and yale (2018. *Consciousness and Cognition*, *68*, 97–106.

Moors, P., Hesselmann, G., Wagemans, J., & Ee, R. (2017). Continuous flash suppression: Stimulus fractionation rather than integration. *Trends in Cognitive Sciences*, *21*(10), 719–721.

Mudrik, L., Breska, A., Lamy, D., & Deouell, L. Y. (2011). Integration without awareness: Expanding the limits of unconscious processing. *Psychological Science*, *22*(6), 764–770.

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage*, *25*(3), 653–660.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, *10*(5), 1003588.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 4716.

Peirce, C. S., & Jastrow, J. (1884). *On small differences in sensation*.

Peters, M. A., Kentridge, R. W., Phillips, I., Block, N., Z., & Lamy, D. (2017). Does unconscious perception really exist? Continuing the ASSC20 debate. *Neuroscience of Consciousness*, *3*(1), 969.

Rabagliati, H., Robertson, A., & Carmel, D. (2018). The importance of awareness for understanding language. *Journal of Experimental Psychology: General*, *147*(2), 190.

Rahnev, D. (2021). Response bias reflects individual differences in sensory encoding. *Psychological Science*, *32*(7), 1157–1168.

Rahnev, D., Desender, K., Lee, A. L., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., & Zylberberg, A. (2020). The confidence database. *Nature Human Behaviour*, *4*(3), 317–325.

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*(6), 563–575.

Rothkirch, M., & Hesselmann, G. (2017). What we talk about when we talk about unconscious processing–a plea for best practices. *Frontiers in Psychology*, *8*, 835.

Rouder, J. N., & Haaf, J. M. (2020). Beyond means: Are there stable qualitative individual differences in cognition. *Journal of Cognition*.

Rouder, J. N., & Haaf, J. M. (2021). Are there reliable qualitative individual difference in cognition? *Journal of Cognition*, *4*(1).

Schlaghecken, F., & Eimer, M. (2004). Masked prime stimuli can bias "free" choices between response alternatives. *Psychonomic Bulletin & Review*, *11*(3), 463–468.

Schnuerch, M., Nadarevic, L., & Rouder, J. N. (2021). The truth revisited: Bayesian analysis of individual differences in the truth effect. *Psychonomic Bulletin & Review*, *28*(3), 750–765.

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *24*, 752–775.

Sklar, A. Y., Levy, N., Goldstein, A., Mandel, R., Maril, A., & Hassin, R. R. (2012). Reading and doing arithmetic nonconsciously. *Proceedings of the National Academy of Sciences*, *109*(48), 19614–19619.

Skora, L. I., Yeomans, M. R., Crombag, H. S., & Scott, R. B. (2021). Evidence that instrumental conditioning requires conscious awareness in humans. *Cognition*, *208*, 104546.

Stein, T., & Peelen, M. V. (2021). Dissociating conscious and unconscious influences on visual detection effects. *Nature Human Behaviour*, *5*(5), 612–624.

Stein, T., Utz, V., & Opstal, F. (2020). Unconscious semantic priming from pictures under backward masking and continuous flash suppression. *Consciousness and Cognition*, *78*, 102864.

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *Neuroimage*, *65*, 69–82.

Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., & Schwarzbach, J. (2003). Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences*, *100*(10), 6275–6280.

Zerweck, I. A., Kao, C. S., Meyen, S., Amado, C., Eltz, M., Klimm, M., & Franz, V. H. (2021). Number processing outside awareness? Systematically testing sensitivities of direct and

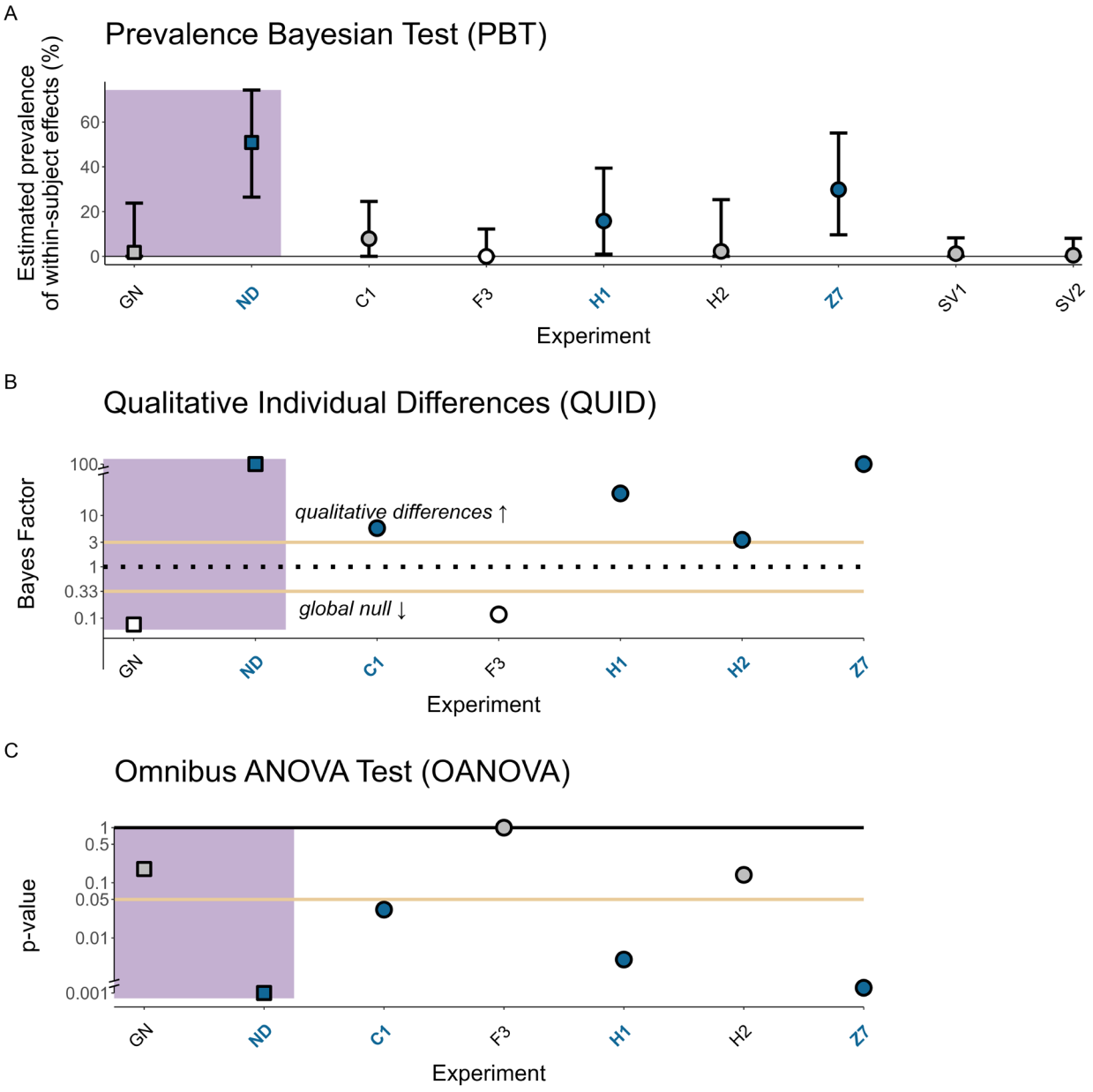
indirect measures of consciousness. *Attention, Perception, & Psychophysics*, *83*(6), 2510–2529.

Supplementary Table 1

Unconscious processing effects metadata

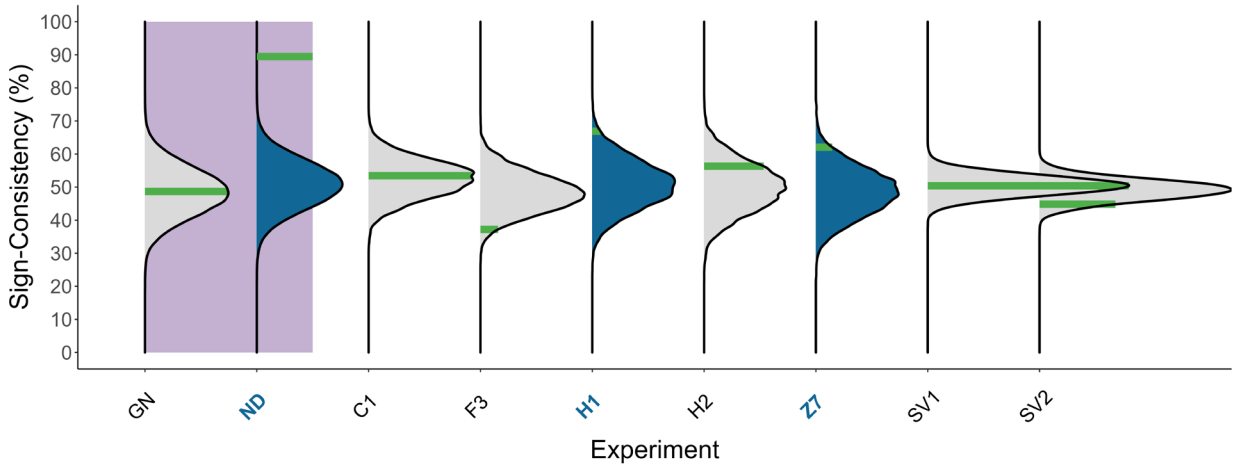
Study	Labels	Topic	Paradigm	DV	Notes
Biderman & Mudrik, 2018	BM1-3	Scene congruency	Masking	RT	Replication study. For all experiments, log(RT) was used in the original analysis
Faivre et al., 2014	F1-8	Multisensory integration	Masking	RT	Four experiments, with two effects in each experiment (identical/different targets). For all experiments, log(RT) was used in the original analysis
Stein & Peelen, 2021	SVP1-5	Location effects + PAS (detection)	CFS	d’	Two experiments (3&4 in the paper), measuring effects in different prime-mask SOAs
Zerweck et al., 2021	Z1-7	Numerical Priming	Masking	RT	Two experiments (2 and 3 in the original paper), measuring effects in different SOA / Contrast conditions
Benthien & Hesselmann, 2021	BH1	Numerical Priming	CFS	RT	Interaction effect - prime congruency X location certainty
Hurme et al., 2020	H1-4	Colours	TMS + Metacontrast Masking	RT	Redundant target effect (TMS / Masking X Blue / Red)
Skora et al., 2021	S1-2	Instrumental Learning	Masking	d’	Regression to the mean as a confound according to authors
Chien et al., 2022	C1-3	Semantic priming	CFS	RT	Word, Picture, and trait discrimination tasks

Supplementary Figure 1 - PBT, QUID and OANOVA results for datasets showing a directional effect



Supplementary Figure 1. The results of applying the PBT (A), QUID (B) and OANOVA (C) tests to effects that produced significant results in a non-parametric directional test. Same conventions as Fig. 2.

Supplementary Figure 2 – Sign Consistency test results for datasets showing a directional effect



Supplementary Figure 2. The results of applying the sign consistency test to significant directional effects (N = 7). The x-axis lists effect labels. Same conventions as Fig. 3.

Appendix A. Violating the equal within-individuals variance assumption

We used the simulation scheme described in the main text (see section 'Simulating non-directional unconscious effects'), to test the consequences of violating the equal within-individuals variance assumption for both QUID and the OANOVA test. We compared the distribution of Bayes factors and p-values obtained by applying QUID and the OANOVA test to generated data meeting and violating the equal within-participants variance assumption. In the first, equal-variance case, the within-individual standard deviation was low ($\sigma_w = 1$) for all participants. In the second, unequal-variance case, the within-individual standard deviation was low ($\sigma_w = 1$) for all participants except one, for whom it was set to a high value ($\sigma_w = 1000$). As in the main simulation, the effect sizes of each participant were sampled from a normal distribution centred at zero ($e_i \sim \mathcal{N}(0, \sigma_b)$; where e_i denotes the effect size of the i^{th} participant). Within this framework, examined two scenarios: a *non-directional differences* scenario where participants are differentially affected by the experimental manipulation ($N_p = 30$, $\sigma_b = 15$), and a *global null* condition where all participants are unaffected by the experimental manipulation ($N_p = 100$, $\sigma_b = 0$). In both scenarios, we simulated random data in 100 iterations, and used the same number of trials per condition (the total number of trials, $N_t = 200$).

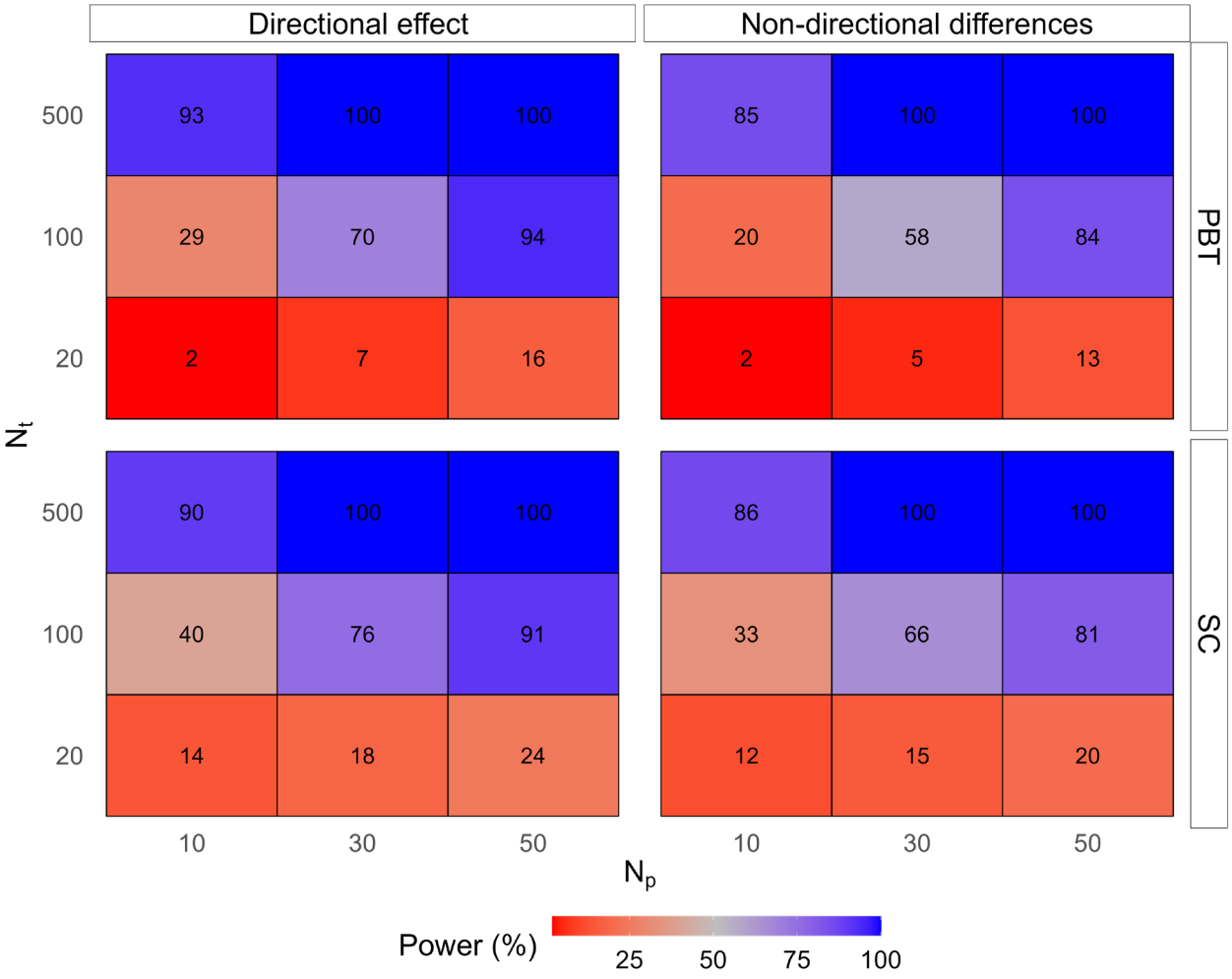
To examine the tests' specificity, we measured the proportion of iterations where evidence for an effect was erroneously found in the global null condition. In the equal-variance case, all iterations provided evidence for the lack of an effect according to QUID (all BFs $< \frac{1}{3}$). Similarly, nonsignificant results were found by the OANOVA test in 93% of the iterations (since the 95% CI of a random binomial variable with $p = 0.05$ and n = 100 is [1 10], the finding 7% false-positive results is expected due to randomness). However, in the unequal-variance case, false-

positives were obtained in 22% of the QUID Bayes Factors (BF > 3), and 2% showed inconclusive evidence. Again, the OANOVA test showed a similar pattern, detecting falsely significant effects in 27% of the iterations. Thus, we show that the specificity of these tests is compromised by violations of the equal-variance assumption.

We then analyzed the tests' outcomes in the non-directional differences scenario to examine their sensitivity. When the equal-variance assumption was met, both tests found evidence for an effect (all BFs > 3, and all p-values < 0.05). In contrast, in the unequal-variance case, only 29% of QUIDs BFs showed evidence for an effect, whilst 64% showed evidence for no effect (the remaining 7% were inconclusive). Similarly, the OANOVA test found significant effects in only 35% of the iterations. Hence, both tests missed true effects when the assumption was not met, demonstrating that their sensitivity is compromised by violations of the equal-variance assumption.

Appendix B. Comparing the power of the PBT and sign consistency test

To examine the differential sensitivity of PBT and the sign consistency tests, we conducted a power analysis, simulating two scenarios under the simulations scheme described in the main text (see section 'Simulating non-directional unconscious effects'): First, a *non-directional differences* scenario where an effect exists for each participant but it is inconsistent within participants ($e_i \sim \mathcal{N}(0, \sigma_b)$; where σ_b=2). Second, a *directional effect* scenario, with individual variation around a positive mean effect size ($e_i \sim \mathcal{N}(1, \sigma_b)$; where σ_b=2). For each scenario, we manipulated the number of simulated participants (N_p=10/30/50) and trials (N_t=20/100/500) across 1000 random iterations, with the within-participant SD (σ_w) set to 10 in both scenarios. Statistical power was defined as the proportion of significant results in the Sign Consistency case, and as the proportion of HDI_{95} intervals that do not include zero in the PBT case. While both tests were similarly sensitive when applied to well-powered datasets (e.g., when N_p=50 or N_t=500), the sign consistency test proved to be more sensitive in the remaining conditions (see Figure 7 in Ince et al., (2021), for a comparison between the power of PBT and a standard t-test for the *directional effect* scenario).



Supplementary Figure B1. Power analysis for the sign consistency (SC) and Prevalence Bayesian (PBT) tests for simulated datasets. Each cell depicts the percent of iterations where SC resulted in significant effects, and where the prevalence HDI_{95} precluded zero (upper and lower panels, respectively). Rows and columns correspond to the number of simulated participants (N_p), and the number of simulated trials per participant (N_t), respectively, Left panel: the results of the PBT and SC in the *non-directional differences* scenario. Right: the results of the PBT and SC in the *directional effect* scenario.

BOX A: Non-directional testing: best practice recommendations

- **When should we use the non-directional approach?**
  - Not all hypotheses are suitable for examination under the non-directional approach. Since the non-directional approach is targeted at detecting the presence of effects rather than their direction, it cannot be used to establish average differences between conditions at the group level (e.g., when comparing memory performance for items presented first and later in an experiment, rejecting a non-directional hypothesis does not entail evidence for an overall primacy or recency effect on recollection). In the case of unconscious processing, the theoretical question is regarding the presence or absence of a difference between the two conditions at the single-participant level, and as such, it lends itself to non-directional testing. Thus, selecting whether to use the non-directional or directional approach is directly linked to the theoretical question at stake.
- **Which test should be used?**
  - When testing whether the experimental manipulation affects the dependent variable (e.g., either main or interaction effects on reaction times, accuracy, brain activity etc.), unless normality and equal variance of within-participant variability can be assumed with high certainty, we recommend using the sign consistency test.
  - When these assumptions hold, QUID or OANOVA can be used for effects that are measured on a trial-by-trial basis (as opposed to effects measured by summarizing data from multiple trials, e.g., d' or correlation effects). Specifically, when prior data is available, we advise incorporating it into the analysis using QUID, and when examining an interaction effect, OANOVA provides an easy-to-use solution.
  - For an estimation of the true prevalence of individual-level effects, rather than the mere existence of an effect at the group level, we recommend using PBT.
- **Non-directional tests require within-participant counterbalancing of confounding variables**
  - As we discuss in the text, special care should be given to counterbalancing of confounding variables when using the non-directional approach. Specifically, unlike standard directional tests, the effects of confounders are not averaged out at the group level when counterbalanced across participants. Thus, counterbalancing should be done not only across participants but also across trials within participants.
- **How to interpret non-directional effects?**
  - In contrast to directional tests, where signal is measured relative to variability across individuals, in non-directional tests it is measured relative to variability across different trials, within an individual. Hence, a positive result of a directional test indicates that effects are consistent between participants, while a non-directional test reveals the presence of an effect on the dependent variable within

participants, regardless of the alignment of within participants effects across participants.

– A significant non-directional effect without a corresponding directional effect suggests reliable variability in effect signs across individuals. Whether this variability reflects transient or stable individual differences can be further tested by correlating individual effect scores from two experimental sessions: stable differences should result in a positive correlation. Whenever stable individual differences are observed, further research may be needed to identify the relevant personal traits that interact with the experimental manipulation.