**Supplementary Materials**

Below we describe details of our methods and results not reported in the main text. Unless otherwise noted below, all simulations followed the methods described in the text.

**Methods**

*Building two populations of “subjects”*

Across all simulated subjects, we set d’ *~ N(μd’, σd’)* and *c1 ~ N(μc1, σc1)*, with parameters defined by the mean and standard deviation of subjects’ behavior in Rounis et al.’s (2010) “pre-real TMS” condition (*μd’* = 1.669, *σd’* = 0.550; *μc1* = -.195, *σc1* = .276). Similarly, we set *c2*~ *N(μc2, σc2)*, with estimated values for the pre-real-TMS condition (*μ-c2* = -1.011, *σ-c2* = 0.536, *μc2* = 0.736, *σc2* = 0.599) and post-real-TMS condition (*μ-c2* = -1.135, *σ-c2* = 0.652, *μc2* = 0.976, *σc2* = 0.613) based on Rounis et al. To estimate these *c2* values, for every human subject in Rounis et al. (2010) we simulated 300 trials of the spatial 2AFC task (described in the main text) for both the pre-real- and post-real-TMS conditions; for each of these two simulated conditions we swept through possible ±*c2* criteria values from 0 to 5 in steps of 0.001 until the proportion of “clear” versus “unclear” responses of each type (◼◆ or ◆◼) matched that given by the human subject. Across all simulated subjects *σTMS ~ N(μTMS, ΣTMS)*, with this parameter based on the distribution of reductions in metacognitive sensitivity in Rounis et al.’s “post-real TMS” condition compared with their “pre-real TMS” condition (*μTMS* = 0.6325, *ΣTMS* = 0.15).

*Simulating populations of subjects for three within-subjects simulations*

Due to the differing structures of the between-subjects and double-repeat experiments of Bor et al. (2017) and the more conventional within-subject design of Rounis et al. (2010), it was necessary to simulate separate populations of subjects for each experiment type. We therefore simulated an effect-present and effect-absent population separately for the double-repeat and within-subject designs, with 5,000 subjects completing all experimental conditions (Pre-Real TMS, Post-Real TMS, Pre-Sham TMS, Post-Sham TMS) in each population. The within-subject population was used to conduct two separate simulations using the sample sizes found in Rounis et al. (n=20) and Bor et al. (n=27). To ensure that the effect on metacognitive sensitivity reported by Rounis et al. (2010) was successfully recreated in the simulated effect-present populations but not in the effect-absent populations, for each within-subjects populations we compared the means for each condition and also ran a 2 x 2 within-subjects ANOVA on the metacognitive sensitivity values, with factors TMS type (Real TMS, Sham TMS) and time (Pre-TMS, Post-TMS).

Within-subjects simulation 1: double-repeat design (Bor et al., 2017, Experiment 2)

In Bor et al.’s (2017) double-repeat within-subjects experiment (their Experiment 2), 27 human subjects completed up to four days of experimental manipulations. Subjects first went through one session each of the Pre-Real- and Post-Real-TMS conditions (Day 1); if the absolute value of the difference of their meta d’ – d’ values for the two conditions was greater than .4, then they advanced to Day 2. Otherwise, they were excluded from the sample. On Day 2, subjects went through the Pre-Sham- and Post-Sham-TMS conditions; if the absolute value of the difference of their meta d’ – d’ values for the two conditions was less than .2, then they advanced to Day 3. Otherwise, they were excluded from the sample. Day 3 was identical to Day 1 and Day 4 was identical to Day 2.

Based on this design, we executed 1,000 simulated experiments of 27 subjects each. Any subjects making it past at least Day 2 were included in subsequent statistical analyses; otherwise they were not. For each simulated experiment in this and the other two within-subjects simulations (see below), we used the Shapiro-Wilk test (Shapiro & Wilk, 1965) to check for normality, first before and again after exclusion. For each normally distributed sample, we ran the standard (parametric) 2 x 2 within-subjects ANOVA on metacognitive sensitivity values, with factors TMS type (Real TMS, Sham TMS) and time (Pre-TMS, Post-TMS). For each sample that violated normality assumptions, we ran a permutation test to assess statistical significance, such that we permuted the sample 1,000 times by randomly switching the Real and Sham labels for each simulated subject independently and ran the aforementioned mixed-design ANOVA at each permutation. As described in the main text, this yielded a null distribution of F-values against which the F-value in the original data could be compared to assess statistical significance. If the F statistic for the interaction term for the non-permuted sample was greater than or equal to 95% of the of F statistics for the interaction term for the 1,000 permuted samples, then the TMS x time interaction for that sample was taken to be significant. We also calculated power and false positive rates for normal versus non-normal samples, for both exclusion and non-exclusion separately.

Within-subjects simulation 2: standard within-subjects design (Rounis et al., 2010), n = 27

In Rounis et al.’s (2010) experiment, 20 human subjects went through four conditions of the behavioral task in a within-subjects design: Pre-Real-TMS, Post-Real-TMS, Pre-Sham-TMS, and Post-Sham-TMS. Thus, Pre- and Post- measures were taken on the same day, before and after application of Real or Sham TMS. Mirroring this design, we performed 1,000 simulated experiments of 27 subjects each.

Within-subjects simulation 3: standard within-subjects design (Rounis et al., 2010), n = 20

Our final simulation was identical to the previous simulation, with the exception that the sample size for each simulated experiment was 20 instead of 27 to match the original sample size used by Rounis et al. (2010).

Applying subject exclusion to Rounis et al.’s (2010) original dataset:

We re-ran analyses on Rounis et al.’s (2010) original data set while adhering to Bor et al.’s exclusion criteria. We also applied the Shapiro-Wilk test to the original data set to test for a violation of the normality assumption. In the case of such a violation, we furthermore elected to include the full sample from Rounis et al., but replace the 2 x 2 within-subjects ANOVA with a permutation test to check for statistical significance.

**Results**

Assessment of effect-present and effect-absent within-subject populations

For each within-subjects population comparisons of condition means and a 2 x 2 within-subjects ANOVA, with factors TMS type (Real-TMS, Sham-TMS) and time (Pre-TMS, Post-TMS), confirmed the detrimental effect of TMS on metacognitive sensitivity (i.e., meta d’ - d’) but the expected null effect on basic task performance with significant TMS type x time interactions. For the population used for the double-repeat design, F(1,4999) = 409.39, p < .001, and condition means were as follows: Pre-Real-TMS d’=1.683 and meta d’=1.554, Post-Real-TMS d’= 1.690 and meta d’ = 1.150, Pre-Sham-TMS d’= 1.690 and meta d’ = 1.577, Post-Sham-TMS d’= 1.692 and meta d’ = 1.556. For the population used for the two within-subjects designs based on Rounis et. al. (2010); F(1,4999) = 374.81, p < .001, and condition means were: Pre-Real-TMS d’=1.689 and meta d’=1.568, Post-Real-TMS d’= 1.705 and meta d’ = 1.115, Pre-Sham-TMS d’= 1.703 and meta d’ = 1.587, Post-Sham-TMS d’= 1.703 and meta d’ = 1.600. Conversely, comparisons of condition means and 2 x 2 within-subjects ANOVAs for effect-absent within-subjects populations showed virtually no effect of TMS on metacognitive sensitivity. For the double-repeat design F(1,4999) = 5.72, p = .017, and condition means were: Pre-Real-TMS d’=1.681 and meta d’=1.541, Post-Real-TMS d’= 1.691 and meta d’ = 1.531, Pre-Sham-TMS d’= 1.689 and meta d’ = 1.535, Post-Sham-TMS d’= 1.688 and meta d’ = 1.546. For the population used for the other two within-subjects designs, For the double-repeat design F(1,4999) = 5.95, p = .015, and condition means were: Pre-Real-TMS d’=1.686 and meta d’=1.549, Post-Real-TMS d’= 1.701 and meta d’ = 1.548, Pre-Sham-TMS d’= 1.702 and meta d’ = 1.547, Post-Sham-TMS d’= 1.699 and meta d’ = 1.570. It is worth noting that while the ANOVAs for the two within-subjects effect-absent populations reached *statistical* significance, by looking at the means one can see that the statistical significance is only due to the large degrees of freedom and not due to meaningful differences between the groups or conditions (in comparison, there were meaningful differences between the groups/conditions in the effect-present populations).

Main results of simulated experiments

Results for all simulated experiments are shown in Table 1 below, including those not detailed in the main text. The main results listed in the table are for analyses executed without performing permutation tests in cases of violations of normality, i.e. standard ANOVAs. However, as we explain in the main text, we found similar results for all experimental designs both before and after including permutation tests in our analyses; results of the analyses which used nonparametric permutation tests to assess statistical significance are included in parentheses in the table.

**Table 1**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Design** | **FPR**  **No Excl (w/ perm.)** | **FPR Excl (w/ perm.)** | **Power**  **No Excl (w/ perm.)** | **Power Excl (w/ perm.)** | **p(data|no effect) (w/ perm.)** | **p(data|**  **effect) (w/ perm.)** | **p(effect|**  **data) (w/ perm.)** |
| **Between-**  **Subjects** | 0.046 (0.053) | 0.048 (0.053) | 0.304 (0.315) | 0.409 (0.420) | 0.029  (0.037) | 0.089 (0.088) | 0.754  (0.704) |
| **Double-**  **Repeat** | 0.121 (0.085) | 0.081 (0.026) | 0.372 (0.348) | 0.308 (0.189) | 0.075 (0.069) | 0.161 (0.203) | 0.682 (0.746) |
| **Within-**  **Subjects**  **n = 27** | 0.053 (0.069) | 0.053 (0.070) | 0.388 (0.425) | 0.567 (0.604) | 0.038 (0.049) | 0.101 (0.098) | 0.727 (0.667) |
| **Within-**  **Subjects**  **n = 20** | 0.042 (0.064) | 0.048 (0.067) | 0.311 (0.363) | 0.432 (0.453) | 0.029 (0.044) | 0.118 (0.135) | 0.803 (0.754) |

**Table 1. False positive rates, statistical power, and other related values as a function of subject exclusion and experimental design.** FPR No Excl: false positive rates with no subject exclusion. FPR Excl: false positive rates after exclusion. Power No Excl: statistical power with no exclusion. Power Excl: power after exclusion. p(data|no effect): the probability of the observed data (i.e., a positive finding that TMS impairs metacognition before but not after exclusion, as in Bor et al.’s (2017) between-subjects experiment) given no effect in the population. p(data|effect): the probability of the observed data given a true effect in the population. p(effect|data): the probability of a true effect in the population given the observed data, using a Bayesian analysis with uninformative priors. There were no salient differences in false positive rates before versus after subject exclusion for any of the experimental designs, although exclusion did slightly reduce false positive rates for the double-repeat design. We observed decreased power for exclusion in comparison with no exclusion for the double-repeat design, while for all other designs we found increased power for exclusion. Moreover, power and false positive rates varied considerably as a function of experimental design. For all experimental designs, the probability of the observed data was at least 2-3 times higher when the effect was present in the population than when it wasn’t. Therefore, for all designs we show that when finding a positive result before but not after exclusion (as in Bor et al.’s between-subjects experiment), there is likely a true effect in the population. Results from nonparametric permutation tests of significance in the case of violations of normality (shown in parentheses in Table 1) parallel the parametric test results.

Power and false positive rate for normal versus non-normal samples

We addressed the effects of exclusion on power and false positive rates separately in samples that did or did not deviate from normality (Tables 2-5). Strikingly, these results demonstrate that for several experimental designs, even exclusion of data according to the criteria used by Bor et al. (2017) does not always (or even mostly) result in an entirely normally-distributed sample.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Effect-present** | **Before exclusion** | **% non-normal samples** | | 97.1% |
| **Power** | **Normal samples** | 0.621 |
| **Non-normal samples** | 0.306 |
| **After exclusion** | **% non-normal samples** | | 66.9% |
| **Power** | **Normal samples** | 0.523 |
| **Non-normal samples** | 0.369 |
| **Effect-absent** | **Before exclusion** | **% non-normal samples** | | 98.5% |
| **FPR** | **Normal samples** | 0.000 |
| **Non-normal samples** | 0.054 |
| **After exclusion** | **% non-normal samples** | | 75.7% |
| **FPR** | **Normal samples** | 0.041 |
| **Non-normal samples** | 0.057 |

Table 2. Power and false positive rates (FPR) in normal vs. non-normal samples before and after exclusion in the between-subjects design.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Effect-present** | **Before exclusion** | **% non-normal samples** | | 47.8% |
| **Power** | **Normal samples** | 0.189 |
| **Non-normal samples** | 0.515 |
| **After exclusion** | **% non-normal samples** | | 10.7% |
| **Power** | **Normal samples** | 0.132 |
| **Non-normal samples** | 0.664 |
| **Effect-absent** | **Before exclusion** | **% non-normal samples** | | 21.9% |
| **FPR** | **Normal samples** | 0.017 |
| **Non-normal samples** | 0.329 |
| **After exclusion** | **% non-normal samples** | | 4.2% |
| **FPR** | **Normal samples** | 0.006 |
| **Non-normal samples** | 0.476 |

Table 3. Power and false positive rates (FPR) in normal vs. non-normal samples before and after exclusion in the double-repeat design.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Effect-present** | **Before exclusion** | **% non-normal samples** | | 98.6% |
| **Power** | **Normal samples** | 0.786 |
| **Non-normal samples** | 0.420 |
| **After exclusion** | **% non-normal samples** | | 30.8% |
| **Power** | **Normal samples** | 0.633 |
| **Non-normal samples** | 0.539 |
| **Effect-absent** | **Before exclusion** | **% non-normal samples** | | 99.2% |
| **FPR** | **Normal samples** | 0.125 |
| **Non-normal samples** | 0.069 |
| **After exclusion** | **% non-normal samples** | | 62.4% |
| **FPR** | **Normal samples** | 0.064 |
| **Non-normal samples** | 0.074 |

Table 4. Power and false positive rates (FPR) in normal vs. non-normal samples before and after exclusion in the within-subjects design (n=27).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Effect-present** | **Before exclusion** | **% non-normal samples** | | 96.1% |
| **Power** | **Normal samples** | 0.641 |
| **Non-normal samples** | 0.352 |
| **After exclusion** | **% non-normal samples** | | 23.6% |
| **Power** | **Normal samples** | 0.466 |
| **Non-normal samples** | 0.411 |
| **Effect-absent** | **Before exclusion** | **% non-normal samples** | | 97.0% |
| **FPR** | **Normal samples** | 0.000 |
| **Non-normal samples** | 0.066 |
| **After exclusion** | **% non-normal samples** | | 52.4% |
| **FPR** | **Normal samples** | 0.059 |
| **Non-normal samples** | 0.074 |

Table 5. Power and false positive rates (FPR) in normal vs. non-normal samples before and after exclusion in the within-subjects design (n=20).

Applying subject exclusion to Rounis et al.’s (2010) original dataset:

We found that Rounis et al.’s (2010) original dataset of 20 subjects exhibited violations of normality, and so we applied Bor et al.’s (2017) exclusion criteria and ran a 2 x 2 within-subjects ANOVA on metacognitive sensitivity with factors TMS type (Real-TMS, Sham-TMS) and time (Pre-TMS, Post-TMS) after exclusion. Although this analysis did return a non-significant result in the interaction between TMS type and time (F(1,6) = .0734, p = 0.796), we note that applying Bor et al.’s exclusion criteria led to eliminating 13 out of 20 subjects -- i.e., a loss of 65% -- leaving only 7 for the analysis. This large reduction in sample size is more than enough to account for the null finding after exclusion. This interpretation is bolstered by the result obtained when we included the full 20 subject sample from Rounis et al. but replaced the (parametric) 2 x 2 within-subjects ANOVA used in the original study (for which a significant interaction was reported) with permutation tests to examine statistical significance. Even with the loss of power that can sometimes result from replacing parametric tests with nonparametric ones (although sometimes parametric tests have *greater* power when assumptions may not be met; Colquhoun, 1971), we observed an interaction between TMS type and time, suggesting again that the observed significant interaction in the original Rounis et al. (2010) study was indeed indicative of a true effect of TMS on metacognitive sensitivity.