Enhancing Statistical Power While Maintaining Small Sample Sizes in Behavioral Neuroscience Experiments Evaluating Success Rates

Theo Desachy[1], Marc Thevenet[1], Samuel Garcia[1], Anistasha Lightning[1,] Anne Didier[1,2], Nathalie Mandairon[1], Nicola Kuczewski[1]

1-Université Claude Bernard Lyon 1, CNRS, INSERM, Centre de Recherche en Neurosciences de Lyon CRNL U1028 UMR5292, NEUROPOP, F-69500, Bron, France 2- Institut Universitaire de France (IUF)

Corresponding autor : Ncola Kuczewski

Correspondence : nicola.kuczewski@univ-lyon1.fr

Abstract

Studies with low statistical power reduce the probability of detecting true effects and often lead to overestimated effect sizes, undermining the reproducibility of scientific results. While several free statistical software tools are available for calculating statistical power, they often do not account for the specialized aspects of experimental designs in behavioral studies that evaluate success rates. To address this gap, we developed "SuccessRatePower" a free and user-friendly power calculator based on Monte Carlo simulations that takes into account the particular parameters of these experimental designs. Using "SuccessRatePower", we demonstrated that statistical power can be increased by modifying the experimental protocol in three ways:
1) reducing the probability of succeeding by chance (chance level), 2) increasing the number of trials used to calculate subject success rates, and 3) employing statistical analyses suited for discrete values. These adjustments enable even studies with small sample sizes to achieve high statistical power. Finally, we performed an associative behavioral task in mice, confirming the simulated statistical advantages of reducing chance levels and increasing the number of trials in such studies

Introduction

Statistical power is the likelihood of obtaining a statistically significant result from a sample when the effect in the entire population equals or exceeds the minimum effect targeted by the study. In other words, it represents the probability that the p-value of a statistical test will be lower than the α-level when the alternative hypothesis is true (p-value < α/HA) (Neyman & Pearson, 1933; Cohen, 1962). Experimental studies with low statistical power have several negative implications. Firstly, they risk overlooking a genuine effect, leading to a false negative conclusion. Secondly, they compromise the precision of estimating population parameters from the experimental sample, potentially resulting in an overestimation of the detected effect when the statistical test is deemed significant (Button et al., 2013; Colquhoun, 2014; Curran-Everett, 2017). Thirdly, they diminish the positive predictive value, which refers to the proportion of published results that are true positives (Ioannidis, 2005; Button et al., 2013; Colquhoun, 2014). Consequently, studies with low statistical power contribute to the ongoing reproducibility crisis in scientific research (Munafò et al., 2017; Ellis, 2022).

Statistical power is influenced by both the effect size (ES) in the studied populations and the sample size (Markel, 1991, Cohen, 1992; Cohen, 2002). While increasing the sample size is the most common method to enhance statistical power, it incurs additional costs in terms of money and time. Furthermore, larger sample sizes can also raise ethical concerns, particularly in invasive human and animal research studies (Sneddon et al., 2017; Andrade, 2020). However, adjustments to parameters other than sample size can also affect statistical power, depending on the study design (Brandmaier et al., 2015). For example, utilizing a repeated measurement design, where the modification of a variable within the same subject is followed overtime, can bolster statistical power (Vickers, 2003; Guo et al., 2013). Mathematical considerations suggest that studies replicating the measurement of the parameter of interest, with each subject being measured multiple times under each condition, may also enhance statistical power. However, this increase in power is contingent upon the absence of perfect correlations between the measurements, meaning that the data collected are not identical for all replications (Goulet and Cousineau, 2019).

Determining statistical power often demands a level of statistical expertise that may exceed the capabilities of many researchers (Fernandes-Taylor et al., 2011; Sullivan et al., 2016). Fortunately, several freely available software packages have been developed to facilitate power calculations (Faul et al., 2007; Lakens & Caldwell, 2021), with G*Power 3 being arguably the most comprehensive and widely utilized by the scientific community. However, these software tools can sometime oeverlook the peculiarities of specific research fields. One such scenario arises in behavioral studies, where differences in cognitive abilities (e.g., memory capabilities) are assessed by evaluating the subject success rate on a specific associative task (see, for example: Ravel et al., 1994; Devore & Linster, 2012; Alonso et al., 2012; Meier et al., 2016; Bratch et al., 2016; Mandairon et al., 2018; Grelat et al., 2018; Zhang et al., 2021).

A classical design for these types of studies is depicted in Figure 1. Here, each subject in the sample must perform several trials (replications) of a task, in which they must select the correct answer from multiple choices. Here the probability of succeeding at the task by chance, defined as

the chance level, is determined by the ratio of correct choices to incorrect ones. In this type of protocol, the experimenter can vary three different parameters: 1) the number ‘n’ of subjects performing the task, 2) the number ‘t’ of trials each subject performs, and 3) the chance level. Furthermore, two different analytical procedures can be applied to the results. The first involves calculating a global population success rate by dividing the total number of successes by the total number of trials. Alternatively, one can first calculate the success rate of each subject and then determine the average success rate of the population. In the first case, a statistical test for discrete values (e.g., a proportion test) is used, while in the second case, a statistical test for continuous values is required (e.g., a Student's t-test). The high flexibility inherent in this design type calls for a tool that enables researchers to compute and compare statistical power relative to the experimental strategies employed. As far as we are aware, such a tool has not yet been developed. Therefore, we present "SuccessRatePower" a freely accessible and user-friendly power calculator based on Monte Carlo simulation, tailored for these experimental protocols. We utilized SuccessRatePower to generate and compare statistical power across various experimental scenarios, identifying conditions that enhance statistical power. Subsequently, we implemented a behavioral protocol in mice to evaluate whether the animals' performance was influenced by the application of experimental parameters anticipated to yield higher-powered designs.

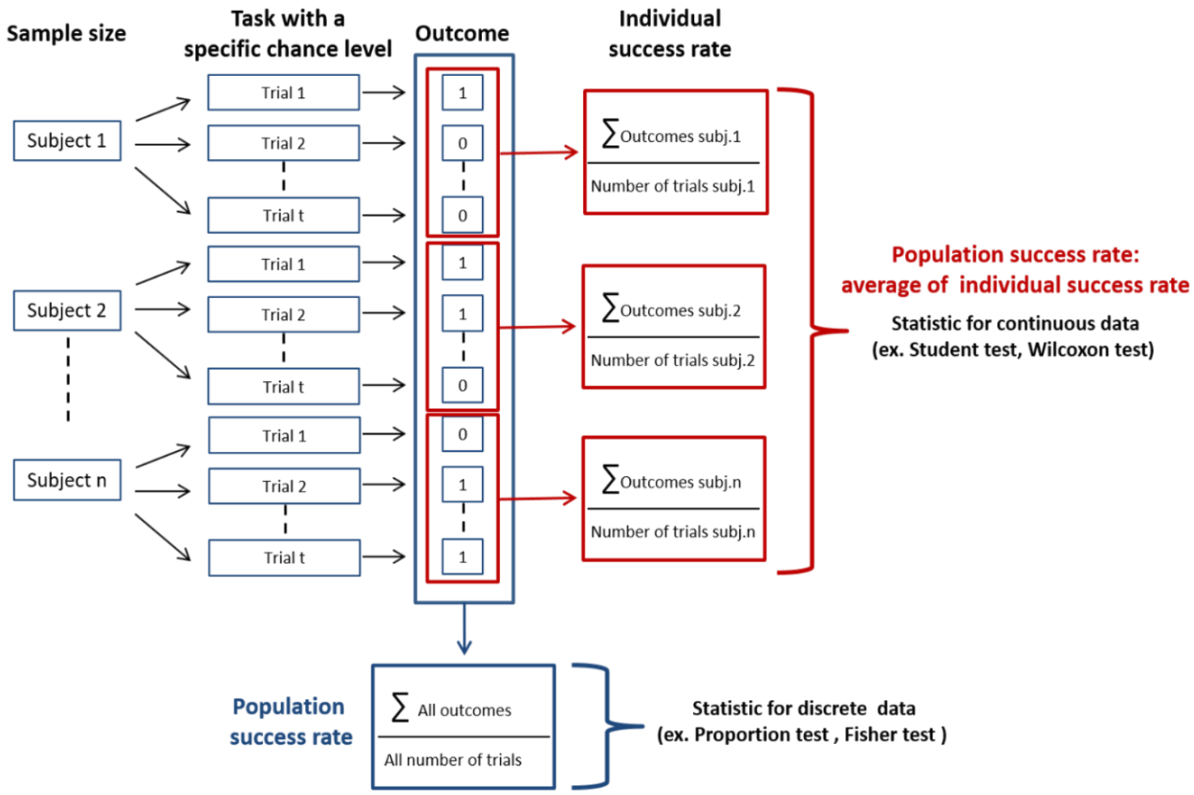


Figure 1. Typical design of experimental protocols aimed to determine the success rate for a behavioral task performed on n subjects and t trials (replication) per subject. The chance level (probability of making a correct choice) is equal to the number of correct choices divided by the total available choices. Outcome 0 represents failure, and outcome 1 represents success. Two statistical analytical procedures can be used to determine whether the success rate is different from chance: 1-Statistical tests on the individual success rates (shown in red). 2-Statistical tests on the population success rate (shown in blue). In both cases, the outcome for trained groups is compared to the chance level or to the success rate of a control group (not shown). Other than the modification of the sample size, the statistical power can potentially be affected by the modification of the chance level, the number of trials per subject, and the statistical analysis procedure

Material and methods

Statistical simulation

The statistical simulation was conducted using the Python script available at the following link: (Script, executable SuccessRatePower). The flow chart describing the script is illustrated in Supplementary Figure 1. The input parameters of the script allow users to define the following:

Mode: *Determine if "One group" is compared to a chance level, or "Two groups" are compared to each other.*

Test method: *Test that will be used for statistical analysis*

Number of trials: *Number of trials performed by each subject*

Success rate Gr1: *Expected average success rate of group 1 (%)*

Variability Gr1: *Variability in performance levels among subjects in group (STD %)*

Sample size Gr1: *Sample size of group 1*

Chance level: *Probability of subjects achieving success by chance (when comparing one group to a chance level)*

Success rate Gr2: *Expected average success rate of group 2 (%)*

Variability Gr2: *Variability in performance levels among subjects in group 2 (%)*

Sample size Gr2 : *Sample size of group 2*

Alpha risk: *α risk*

Number of iterations: *The number of iterations in the Monte Carlo simulation. A higher number provides greater precision in the power estimation*

Repetitions: *Number of Monte-Carlo simulation repetitions. Determines the average power estimation and 95% confidence interval.*

Tolerance: *Specifies the tolerance error on the average and standard deviation of the sample generated for the simulation.*

The script first generates the average success probability for each subject in each group to achieve the required average success probability and standard deviation for the entire group. For each subject, the outcome of each trial of the behavioral task (success=1, failure=0) is generated using the NumPy function random.binomial(1, panimal, size=1), where the parameter panimal represents the average success rate of the subject.

The simulation uses unilateral tests to assess the hypothesis that the trained group (Gr1) perform better compared to either the control group or the chance level. When the average probability of the trained group is set higher than the chance level, the script considers the result as a false negative (type II error) every time the test produces a p-value greater than the α risk. In a case where the average probability of the two groups is the same, the script instead calculates a false positive (type I error). Using this logic, 5000 random samples, were generated and the power calculated as follow:

$$Statistical\ power = \frac{\sum False\ negatives}{5000} * 100$$

For the results presented later in this article, the average success probability of the trained group was set at 70% with 20% variability of performance between subjects. These values are compatible with those reported in the literature, where the success rate of healthy training groups falls between 70% and 90% (see Mandairon et al., 2018). For the analytical method based on continuous values, the success rate for each subject is calculated as the ratio of the number of successes to the number of trials. In this condition, we employed the SciPy function stats.ttest_1samp to compare the success rate of the trained group with the chance level. For comparing the sample success rate of the trained group with the success rate of the control group, we utilized the SciPy function stats.ttest_ind. For the analytical method based on discrete values, the script calculated the total number of success for the sample. In this condition, "statsmodels.stats.proportion.proportions_ztest" is used to compare the success rate of the trained group with both the chance level and success rate of the control group.

Behavioral experiments (pre-registered protocols at osf)

Animals

Six adult male C57BL/6J mice (Janvier-Labs, France), aged 8 weeks at the beginning of the experiments, were housed in groups of six in standard laboratory cages with water ad libitum and maintained on a 12-hour light/dark cycle at a constant temperature of 22°C. Experiments were conducted following procedures in accordance with the European Community Council Directive of 22nd September 2010 (2010/63/UE) and adhered to the standards of the National Ethics Committee (Agreement APAFIS#29948-2021021812246456 v3). Animals were handled according to the 'Mouse Handling: Tutorial' provided by NC3Rs.

Learning and Experimental Design

The mice were trained to associate a food reward (a small amount of sweetened cereal from Kellogg's, Battle Creek, MI, USA) with the presence of a particular odorant (limonene+), within a

radial maze containing either two or four arms (Figure 2). The odorant was introduced at the end of one of the maze arms through a custom-made olfactometer. Training sessions commenced after a shaping period lasting two days. During the shaping period, the animals were allowed to freely navigate the maze without any scent present and with the food reward randomly placed at the end of one of the arms. To incentivize the animals to complete the task, access to food in their home cage was limited to 2g of food pellets per mouse, resulting in an approximately 20% reduction in daily consumption and a consequent 10% approximate reduction in body weight (Mandairon et al., 2018). The shaping period consist of two 5-minute exploration periods each day.

During the subsequent training phase, mice were initially placed in a closed starting box for approximately 10 seconds. After this period, the door leading to the maze was opened, allowing the mice access. A food reward was randomly positioned at the end of one of the maze arms. This rewarded arm was scented through an odorant port that remained open throughout the trial. Mice remained in the maze until they located the reward, with a “successful” trial occurring when the arm containing the reward was the first one visited by the animal.

Two groups of mice were used. Group 1 underwent training in a two-arm maze (with a chance level of ½), while Group 2 trained in a four-arm maze (with a chance level of ¼). Each animal underwent one training session (block) consisting of 4 trials per day. For each group, training continued until the average success rate reached a plateau (no modification of success rate > 5% over previous sessions for three consecutive sessions). After the third session at the plateau, a last session consisting of 8 trials was performed.

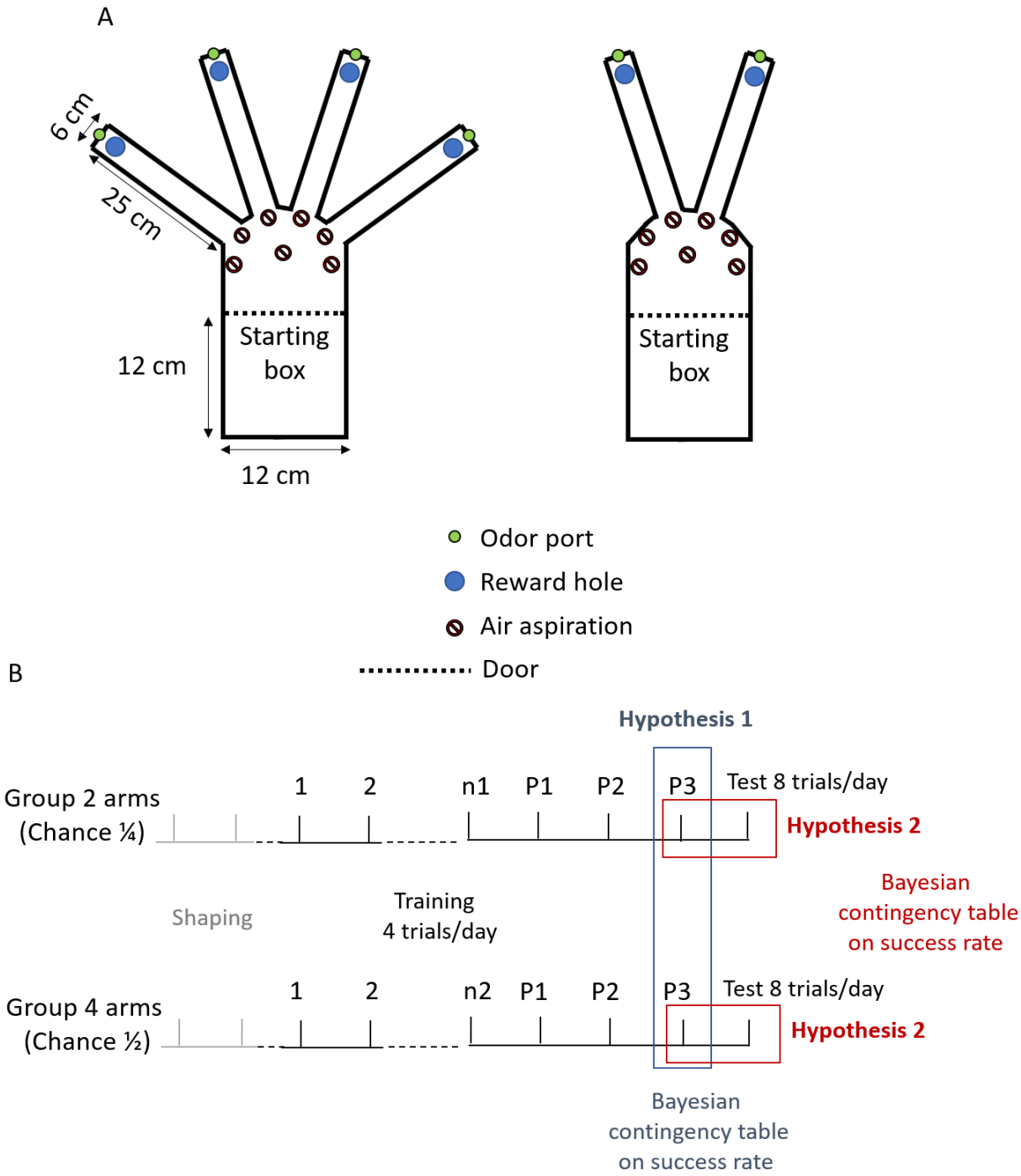


Figure 2. Experimental protocol. (**A**), Radial mazes for the behavioral experiments. Left, chance level of ¼. Right, chance level of ½. (**B**), Behavioral planning involved statistical tests to verify the hypothesis. Tests were conducted once the success rate reached a plateau, defined as no modification of the success rate exceeding 5% over previous sessions for three consecutive sessions (P1-P3). A supplementary session with 8 trials was performed thereafter to test whether the animals' success rate was modified by the increase in the number of trials. n1 and n2, last training day before plateau for group 1 and 2 respectively. P1-P3, training when success rate reaches the plateau.

Tested hypothesis and statistic

Two experimental hypotheses were tested:

- Hypothesis 1: The reduction of the chance level does not reduce the animal success rate. The alternative hypothesis, “success rate of group 1 > success rate of group 2”, was used to test the null hypothesis, “success rate of group 2 ≥ success rate group 1”, by using Bayesian contingency table (JASP Team, 2022). Bayes Factor (BF)01 >3 was the threshold to accept the null hypothesis (Keysers et al., 2020).
- Hypothesis 2: The increase in the number of trials does not reduce the animal success rate.The alternative hypothesis," success rate with 4 trials > success rate with 8 trials", was used to test the null hypothesis, "success rate with 8 trials ≥ success with 4 trials", by using Bayesian contingency table (JASP Team, 2022). BF_{01} >3 was the threshold to accept the null hypothesis.

Analytical procedure

For each session, the success rate of each animal is calculated as the sum of successes divided by the total number of trials. Descriptive statistics, including the average success rate of the population and its corresponding 95% confidence interval (calculated as the standard error of the mean multiplied by 1.96), are then presented for the different training sections. Statistical differences from the chance level were assessed using a one-tailed Student's t-test.

Results

Increased statistical power is achieved by reducing the chance level, increasing the number of trials, and employing statistical analysis for discrete values

In order to determine the statistical power for studies that aim to evaluate success rates, we have developed, a power calculator based on Monte Carlo simulation freely available here SuccessRatePower. Here a simple, notated interface allows users to define the different parameters of the experimental plan (Figure. 3A).

We used SuccessRatePower to assess the effect produced by the modification of the different parameters shown in figure 1A on statistical power. We ran these simulations under two different conditions: one comparing the performance of the simulated experimental group directly to chance level, and the other comparing the performance of the experimental group to a simulated control group performing at chance level. If not stated otherwise, the success rate and variability (standard deviation) of the experimental group were set at 70% and 20%, respectively, which are values commonly observed in animal research

Figure 3B shows the increase in statistical power resulting from the decrease in the chance level. Additionally, it demonstrates that statistical power can be further enhanced by employing an analysis for discrete values (proportion z-test) and by increasing the number of trials conducted per subject. As expected, the statistical power was higher when directly comparing the success rate of

the experimental group to chance level than when comparing it to the success rate of a control group performing at chance level. This is because, in the second condition, the sample variability comes from both groups. Figure 3C depicts this, along with the evolution of statistical power with varying sample sizes across different experimental conditions.

These results highlight a significant advantage in reducing the number of subjects required by adopting an experimental procedure with a low chance level. For instance, when comparing two groups, only 5-10 subjects per group are required to achieve a statistical power above 90% with a chance level at ¼, whereas 40 subjects per group are needed to achieve the same power with a chance level at ½ (refer to Figure 3C, right). This represents a reduction of 75-87% in the number of subjects needed to achieve equivalent statistical power in these behavioral studies. As depicted in Figure 3D, the impact of modifications to the experimental design on statistical power becomes negligible when the success rate of the subjects approaches and exceeds 90%. Finally, Figure 3E illustrates that statistical power decreases as the variability in performance among subjects increases. Notably, this reduction in power occurs only with statistical tests designed for continuous data analysis and not with tests for discrete values (z-testing).

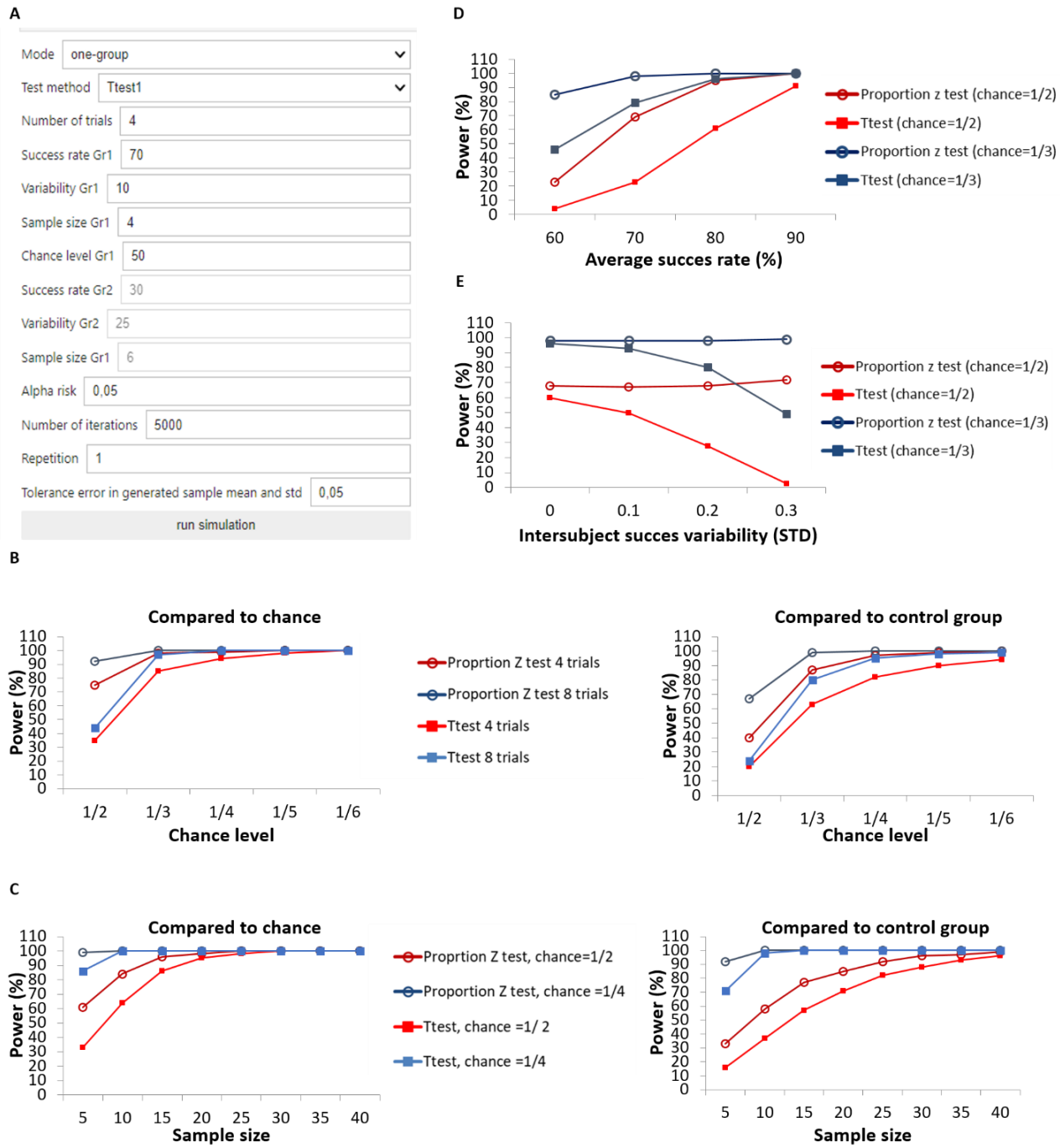


Figure 3: Simulation depicting the modulation of statistical power relative to factors such as the chance level, number of trials, sample size, and the type of statistical test utilized. **(A)** SuccessRatePower interface outlining the parameters employed in the simulation to compute statistical power (see method for more details). **(B)** Statistical power exhibits an upward trend with decreasing chance levels and increasing numbers of trials per subject, particularly when employing statistical tests suited for discrete values. The comparison is made between the trained group and the chance level on the left, and between the trained group and a non-trained control group on the right. Sample size is set at 6, with an average success rate of 70% for the trained group

and a standard deviation of 0.2 for both the trained and control groups. **(C)** Relationship between statistical power and sample size, demonstrating a substantial decrease in the required number of subjects to achieve a power >90% when employing low chance levels and statistical tests tailored for discrete values. Comparison between the trained group and the chance level on the left, and between the trained group and a non-trained control group on the right. The number of trials is fixed at 4, with an average success rate of 70% for the trained group and a standard deviation of 0.2 for both groups. **(D)** Statistical power evolution as a function of the average success rate of the studied population. The standard deviation for both the trained and control groups is 0.2, with 8 trials and a sample size of 6. **(E)** Influence of inter-subject success variability on statistical power, average success rate of 70%, 8 trials, and a sample size of 6. Notably, the 95% confidence interval bars are smaller than the size of the markers. The data corresponding to these figures can be accessed at (*https://osf.io/fcdty*). The script for the simulation can be found at (*https://osf.io/vsuja/*)

Overall, these simulations demonstrate that designing behavioral studies with a minimal chance level, a high number of trials, and employing statistical analyses suited for discrete values can lead to achieving a high level of statistical power, even with a low sample size. However, this conclusion is contingent upon ensuring that these parameters do not compromise subject performance. As illustrated in Figure 3D, statistical power decreases with a reduction in subject success rate, which may occur due to a decrease in the chance level and/or an increase in the number of trials. For instance, lowering the chance level may heighten task difficulty, consequently reducing subject success rates. Additionally, increasing the number of trials may lead to decreased success rates due to factors such as fatigue or diminished motivation (e.g., in animal studies utilizing food rewards where satiety can affect subject engagement). Therefore, to experimentally test these eventualities, we conducted an associative task in mice by varying these two parameters.

Animal performance is not reduced by decreasing the chance level or increasing the number of trials

To assess the impact of modifying the chance level on animal performance, two groups of adult mice (six animals per group) participated in an associative olfactory task within a radial labyrinth, where they encountered either two arms (chance ½) or four arms (chance ¼). The mice underwent training until performance plateaued. The absence of differences in success rates between the two groups was evaluated using a Bayesian statistical test, calculating the probability of the null hypothesis over the alternative (BF_{01}). The training blocks consisted of four trials, except for the last block where eight trials were used in order to evaluate the impact of altering the number of trials on success rates, (for further details, refer to the methods section and Figure 2).

Both the behavioral protocol and analytical plan were pre-registered (here, OSF). Figure 4A illustrates that both groups displayed similar learning curves, with the ¼ chance level group reaching statistically significant effects earlier than the ½ chance level group. Additionally, both groups exhibited comparable performance at the plateau (87.5± 14% for ¼ chance level group, 87.5 ± 20% for ½ chance level group at training block 15; BF_{01}=4.2, n=6). Moreover, the success rate remained unaffected by doubling the number of trials performed by the mice during the eight-trial block (For two arms group: 87.5± 20% for 4 trials, 91.7 ± 10% for 8 trials; BF_{01}=8.7. For four arms group: 87.5± 14% for 4 trials 87.5± 14% for 8 trials BF_{01}=5.3, n=6). The implementation of SuccessRatePower enabled the utilization of experimental data to estimate statistical power across the blocks. As depicted in Figure 4B, the estimated power is clearly larger for the four-arms design.

This superiority is particularly evident in blocks with lower average success rates. Additionally, we utilized SuccessRatePower to estimate the necessary number of subjects for achieving 95% power, and to assess the ratio of animal requirements between the four-arms and two-arms conditions (Figure 4C). This analysis shows that, on average, five times more animals would be necessary when the task is performed with chance level is ½ compared to the condition where task is performed with chance ¼.

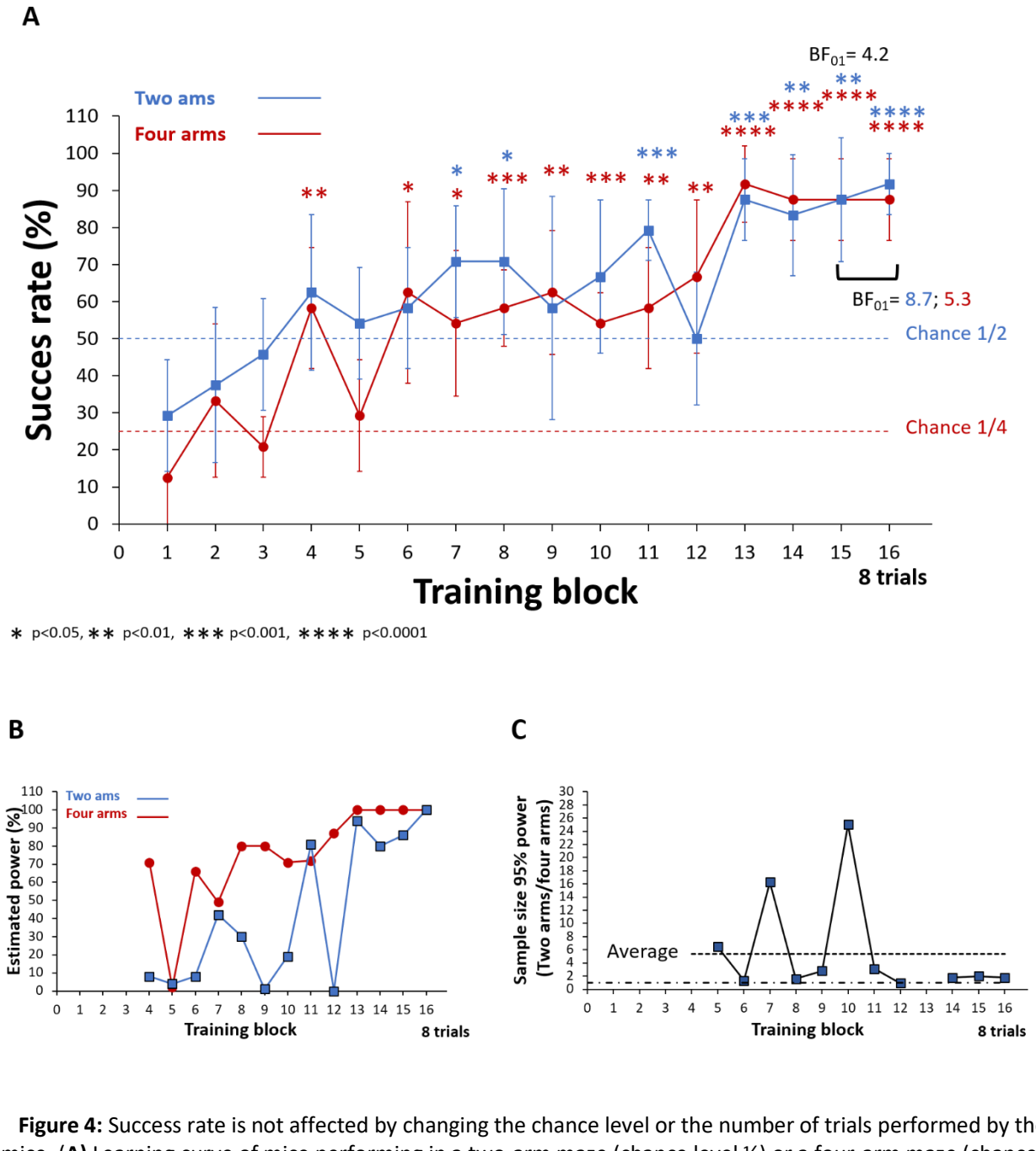


Figure 4: Success rate is not affected by changing the chance level or the number of trials performed by the mice. (**A)** Learning curve of mice performing in a two-arm maze (chance level ½) or a four-arm maze (chance level ¼). Mice underwent four trials per day until reaching block 16. Asterisks denote statistical significance

compared to chance. The Bayesian factor indicates the absence of a significant difference between the two groups at training block 15, as well as the absence of a difference between four and eight trials within each group. **(B)** Estimated statistical power calculated with the data presented in (A). **(C)** Estimation of the increase in animal requirements to reach 95% statistical power when the task is performed in a two-arm (chance level ½) versus a four-arm (chance level ¼) condition. Lower dashed line depict ratio =1. All data are available at *https://osf.io/rt7d5/*.

Discussion

In this report, we present 'SuccessRatePower' a new statistical tool designed to calculate statistical power in studies evaluating population success rates, such as those conducted in behavioral neuroscience assessing differences in cognitive abilities. This software considers various parameters of the experimental design, allowing researchers to evaluate the impact of their modifications on power and thus select the optimal experimental configuration for these studies. Using this tool, we demonstrated that power increases by modifying three parameters controllable by the experimenter: 1) decreasing the chance level, 2) increasing the number of trials performed to assess success rates, and 3) utilizing statistical analyses suitable for discrete values.

In these regards, it is interesting to note that the majority of published studies are designed with the higher chance level (½) and that the statistical analysis is performed after transforming discrete data into continuous data. The reason for the former is, in some circumstances, due to constraints associated with the scientific question. For example, when the study aims to assess the capacity of the subject to discriminate between two different stimuli only chance level ½ can be used. The latter is probably due to historical habits and practice traditions, which together may have produced an unawareness of the impact this transformation has on statistical power. The prevalence of underpowered design in the literature may also be attributable to the anecdotally common concern that decreasing the chance level or increasing the number of trials could compromise the subject's performance and reduce its success rate.

Here, we provide experimental evidence demonstrating that, at least for simple associative tasks, this is not the case. Therefore, by simply modifying a few parameters within the experimental design, we can significantly reduce the number of animals used without compromising the quality of the statistical inference. For example, in the study by Mandairon et al. (2018), the associative task was designed with 4 training trials with a chance level of ½. After 5 days of training, the success rates for the trained and control groups were 77% and 49%, with success variability at std=19 % and std=20 %, respectively. Using 20 animals per group, the statistical power achieved was 97%. However, “SuccessRatePower” demonstrated that the same power could have been reached with only 6 mice per group by performing the experiment with 8 trials at a chance level of ¼. This adjustment not only saves time, but also provides a significant ethical benefit, as minimizing the number of animals necessary for a given protocol is a cornerstone of ethical scientific practice.

It should be noted that the advantages produced by modifying the aforementioned parameters diminish as the success rate of the trained group increases, becoming negligible for values greater than 90% (Figure 3D). Moreover, the benefit of using statistical methods for discrete values

decreases as the variability in performance between subjects is reduced (Figure 3E). Finally, it is important to emphasize that statistical analyses for discrete values can only be performed if all subjects undergo the same number of trials, to avoid results being driven by the performance of over-tested subjects. Therefore, in these types of studies, the parameters that affect statistical power are multiple and interdependent. Thus, for experiments evaluating and comparing animal success rates, the "SuccessRatePower" software is a valuable tool for optimizing the experimental protocol design.

Acknowledgements

We thank Thomas Orset, Clémence Long, and Lucie Plantade who performed preliminary experiments on the behavioral task

Funding

The authors declare that they have received no specific funding for this study.

Conflict of interest disclosure

The authors declare of having no financial or non-financial conflicts of interest in relation to the content of the article.

Data, scripts, code, and supplementary information availability

Scripts is available here, row data and analysis here

References

Brandmaier, AM, von Oertzen, T, Ghisletta, P, Hertzog, C, & Lindenberger, U (2015). LIFESPAN: A tool for the computer-aided design of longitudinal studies. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.00272

Alonso M, Lepousez G, Wagner S, Bardy C, Gabellec M-M, Torquet N & Lledo P-M (2012). Activation of adult-born neurons facilitates learning and memory. *Nat Neurosci* **15,** 897–904.

Andrade C (2020). Sample Size and its Importance in Research. *Indian J Psychol Med* **42,** 102–103.

Bratch A, Kann S, Cain JA, Wu J-E, Rivera-Reyes N, Dalecki S, Arman D, Dunn A, Cooper S, Corbin HE, Doyle AR, Pizzo MJ, Smith AE & Crystal JD (2016). Working Memory Systems in the Rat. *Curr Biol CB* **26,** 351–355.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ & Munafò MR (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* **14,** 365–376.

Cohen J (1962). The statistical power of abnormal-social psychological research: a review. *J Abnorm Soc Psychol* **65,** 145–153.

Colquhoun D (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* **1,** 140216.

Curran-Everett D (2017). CORP: Minimizing the chances of false positives and false negatives. *J Appl Physiol Bethesda Md 1985* **122,** 91–95.

Devore S & Linster C (2012). Noradrenergic and cholinergic modulation of olfactory bulb sensory processing. *Front Behav Neurosci* **6,** 52.

Ellis RJ (2022). Questionable Research Practices, Low Statistical Power, and Other Obstacles to Replicability: Why Preclinical Neuroscience Research Would Benefit from Registered Reports. *eNeuro* **9,** ENEURO.0017-22.2022.

Faul F, Erdfelder E, Lang A-G & Buchner A (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* **39,** 175–191.

Fernandes-Taylor S, Hyun JK, Reeder RN & Harris AH (2011). Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC Res Notes* **4,** 304.

Grelat A, Benoit L, Wagner S, Moigneu C, Lledo P-M & Alonso M (2018). Adult-born neurons boost odor-reward association. *Proc Natl Acad Sci U S A* **115,** 2514–2519.

Guo Y, Logan HL, Glueck DH & Muller KE (2013). Selecting a sample size for studies with repeated measures. *BMC Med Res Methodol* **13,** 100.

Ioannidis JPA (2005). Why Most Published Research Findings Are False. *PLOS Med* **2,** e124.

Keysers C, Gazzola V & Wagenmakers E-J (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nat Neurosci* **23,** 788–799.

Lakens D & Caldwell AR (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Adv Methods Pract Psychol Sci* **4,** 2515245920951503.

Mandairon N, Kuczewski N, Kermen F, Forest J, Midroit M, Richard M, Thevenet M, Sacquet J, Linster C & Didier A (2018). Opposite regulation of inhibition by adult-born granule cells during implicit versus explicit olfactory learning. *eLife*; DOI: 10.7554/eLife.34976.

Meier C, Lea SEG & McLaren IPL (2016). Task-switching in pigeons: Associative learning or executive control? *J Exp Psychol Anim Learn Cogn* **42,** 163–176.

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ & Ioannidis JPA (2017). A manifesto for reproducible science. *Nat Hum Behav* **1,** 1–9.

Neyman J & Pearson ES (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Math Proc Camb Philos Soc* **29,** 492–510.

Ravel N, Elaagouby A & Gervais R (1994). Scopolamine injection into the olfactory bulb impairs short-term olfactory memory in rats. *Behav Neurosci* **108,** 317–324.

Schönbrodt FD & Wagenmakers E-J (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychon Bull Rev* **25,** 128–142.

Sneddon LU, Halsey LG & Bury NR (2017). Considering aspects of the 3Rs principles within experimental animal biology. *J Exp Biol* **220,** 3007–3016.

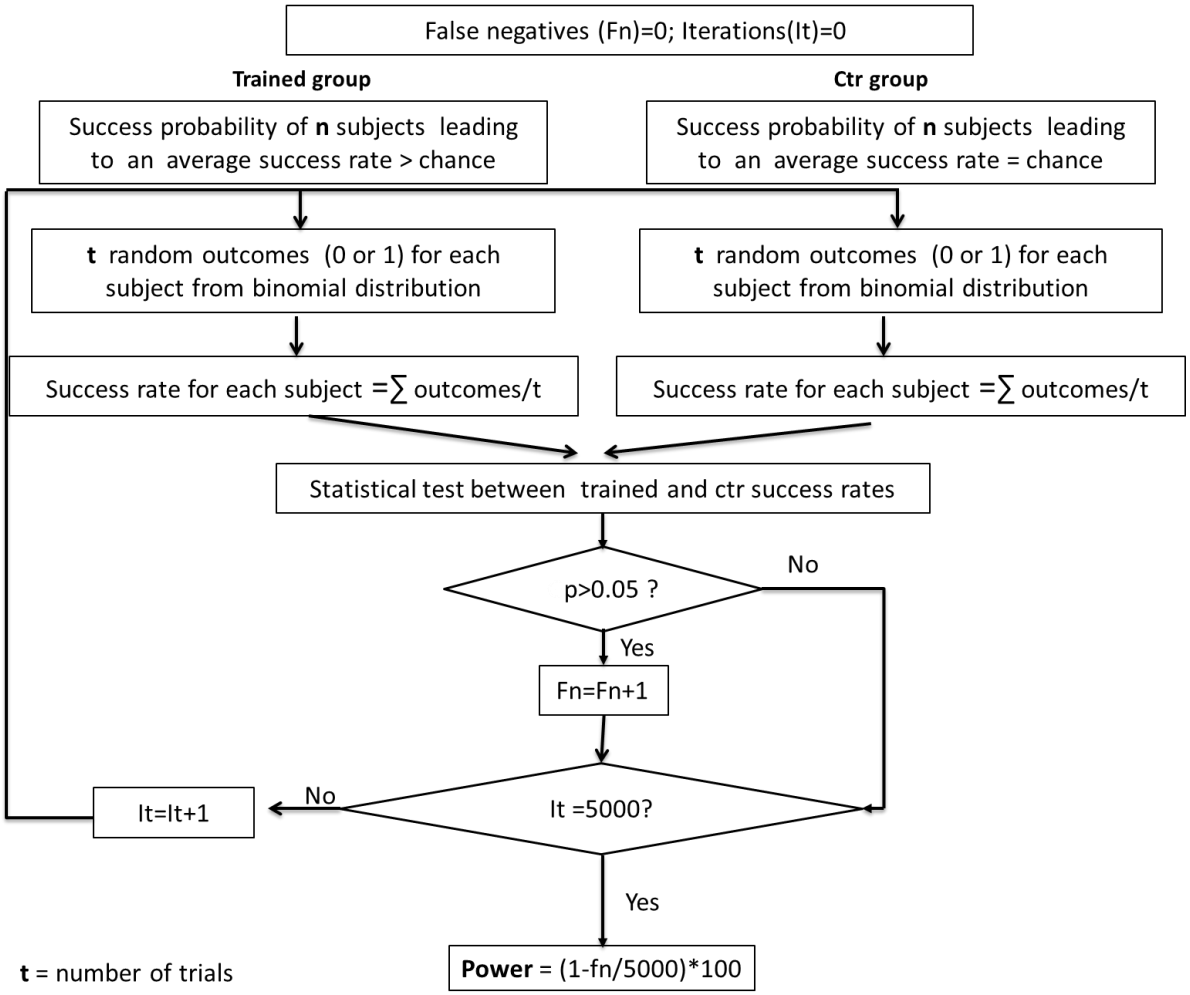
Sullivan LM, Weinberg J & Keaney JF (2016). Common Statistical Pitfalls in Basic Science Research. *J Am Heart Assoc* **5,** e004142.

Vickers AJ (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Med Res Methodol* **3,** 1–9.

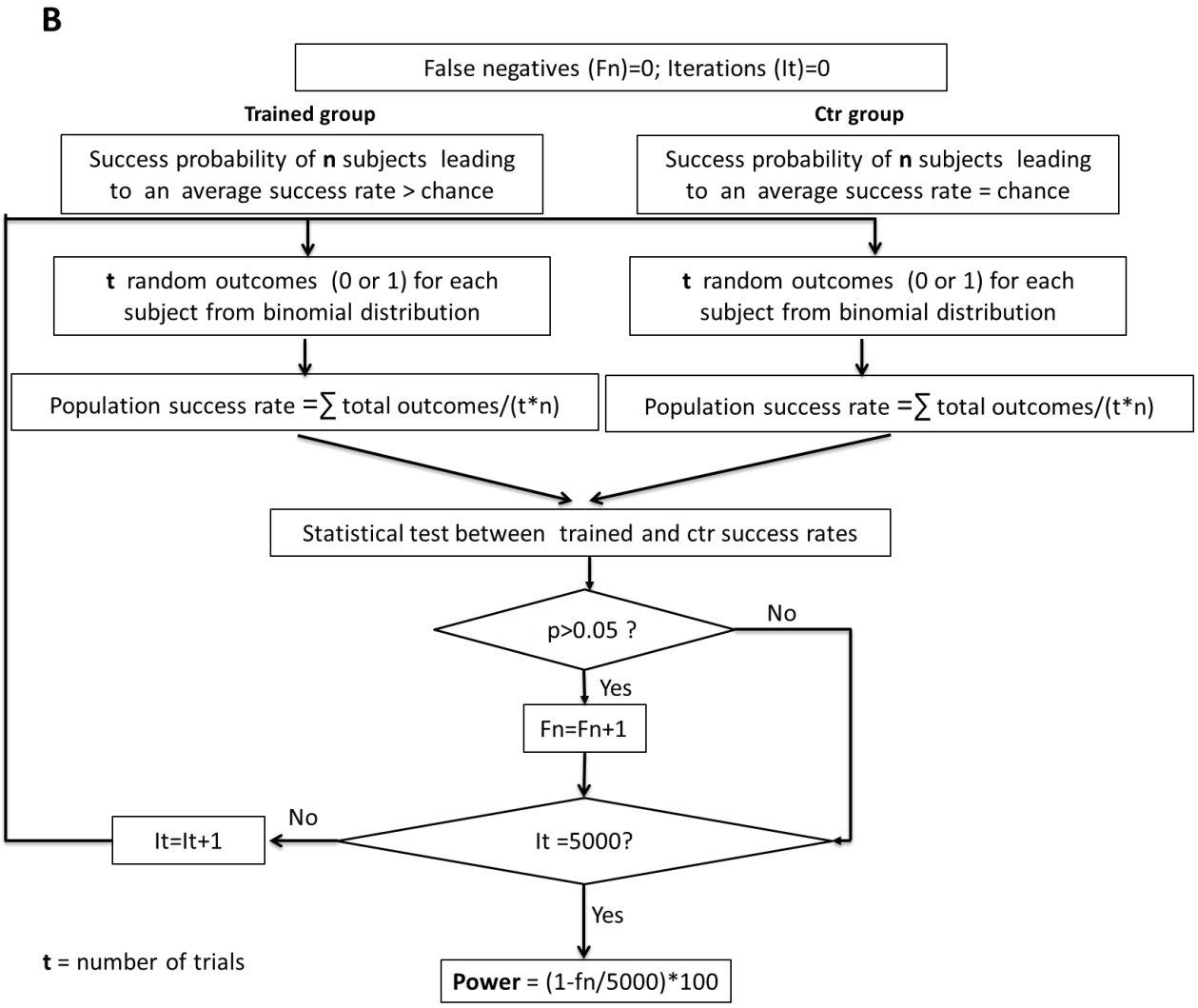
Zhang J, He Y, Liang S, Liao X, Li T, Qiao Z, Chang C, Jia H & Chen X (2021). Non-invasive, opsin-free mid-infrared modulation activates cortical neurons and accelerates associative learning. *Nat Commun* **12,** 2730.

Supplementary figures

A



t = number of trials



Supplementary figure 1: Flow charts illustrating the procedure for determining statistical power through statistical simulation: (A) Flow chart for analytical procedures designed for continuous data. (B) Flow chart for analytical procedures designed for discrete data. When comparing a trained group to chance, the control group is replaced by the chance value. As the two groups exhibit different average success rates, statistical results with a p-value > α risk are identified as false negatives. The script calculates the β risk (false negatives divided by 5000) and computes the statistical power as 1 - β. When the two groups exhibit the same success rate, the script computes the occurrence of type I error.