Published in final edited form as:
Evid Based Preclin Med. ; 3(1): . doi:10.1002/ebm2.16.

Protocol for a Systematic Review of Effect Sizes and Statistical Power in the Rodent Fear Conditioning Literature

Thiago C. Moulin[1], **Clarissa F. D. Carneiro**[1], **Malcolm R. Macleod**[2], and **Olavo B. Amaral**[1]

[1]Institute of Medical Biochemistry Leopoldo de Meis, Federal University of Rio de Janeiro, Brazil
[2]Division of Clinical Neurosciences, University of Edinburgh, Scotland

Abstract

The concepts of effect size and statistical power are often disregarded in basic neuroscience, and most articles in the field draw their conclusions solely based on the arbitrary significance thresholds of statistical inference tests. Moreover, studies are often underpowered, making conclusions from significance tests less reliable. With this in mind, we present the protocol of a systematic review to study the distribution of effect sizes and statistical power in the rodent fear conditioning literature, and to analyze how these factors influence the description and publication of results. To do this we will conduct a search in PubMed for “fear conditioning” AND (“mouse” OR “mice” OR “rat” OR “rats”) and obtain all articles published online in 2013. Experiments will be included if they: (a) describe the effect(s) of a single intervention on fear conditioning acquisition or consolidation; (b) have a control group to which the experimental group is compared; (c) use freezing as a measure of conditioned fear; and (d) have available data on mean freezing, standard deviation and sample size of each group and on the statistical significance of the comparison. We will use the extracted data to calculate the distribution of effect sizes in these experiments, as well as the distribution of statistical power curves for detecting a range of differences at a threshold of $\alpha=0.05$. We will assess correlations between these variables and (a) the chances of a result being statistically significant, (b) the way the result is described in the article text, (c) measures to reduce risk of bias in the article and (d) the impact factor of the journal and the number of citations of the article. We will also perform analyses to see whether effect sizes vary systematically across species, gender, conditioning protocols or intervention types.

Introduction

Background

Basic research in biology over the last decades has been heavily influenced by the concept of statistical significance – i.e. the likelihood that a given effect size would occur by chance under the null hypothesis. Based on arbitrary thresholds set for the results of statistical tests (usually at $p < 0.05$), most articles will classify results as “significant” or “non-significant”,

Corresponding Author: Olavo B. Amaral, M.D., PhD., Instituto de Bioquímica Médica Leopoldo de Meis, Av. Carlos Chagas Filho 373, E-38, Cidade Universitária, Rio de Janeiro, RJ, Brazil, CEP 21941-902, Phone: +55-21-39386762, olavo@bioqmed.ufrj.br.

Conflicts of Interest:
The authors report no conflicts of interest

usually with no regard to the limitations of this approach.[1] Among these, two of the most striking are (a) that p values do not measure the magnitude of an effect, and thus cannot be used to assess its biological significance[2] and (b) that results of significance tests are heavily influenced by the statistical power of experiments, which affect both the chance of finding a significant result for a given effect size and the positive predictive value of a given p value.[3]

A quick inspection of the literature, however, shows that effect size and statistical power rarely receive much consideration in basic research. Discussion of effect sizes is usually scarce, and sample size/power calculations are seldom performed in the preclinical literature.[4,5] The potential impact of these omissions is large, as reliance on the results of significance tests without consideration of statistical power can lead to major decreases in the reliability of study conclusions when studies are underpowered.[6] Moreover, the biological significance of a finding can only be assessed when effect size is considered, as statistical significance by itself is dependent on statistical power, leading even small effects to yield low p values if sample size is sufficiently high. Thus, without taking effect sizes into account, researchers cannot adequately evaluate the potential usefulness of a treatment (in the case of preclinical research), or the importance of the physiological pathway affected by an intervention (in the case of basic science).

Basic research on the neurobiology of memory provides an interesting example of this phenomenon. Advances in pharmacology and molecular biology have shown that hundreds of molecules can influence various forms of memory in rodents, as well as its synaptic correlates such as long-term potentiation.[7] Nevertheless, as effect sizes are rarely considered and the reproducibility of findings is unknown, it is difficult to dissect essential mechanisms in memory formation from modulatory influences affecting behavior (or even from false positive findings). Thus, the wealth of findings in the literature translates poorly into a better comprehension of the underlying phenomena, and the excess of statistically significant findings with small effect sizes and low positive predictive values can actually harm rather than help the field. Moreover, current efforts to minimize sample sizes for ethical reasons can actually make this problem worse, as underpowered studies will lead to unreliable results (and thus waste animal lives in uninformative studies).[6]

To provide an unbiased assessment of the distribution of effect sizes and statistical power in the memory literature, we will perform a systematic review of articles studying interventions that affect acquisition of fear conditioning in rodents. This task is appropriate for this kind of study, as the vast majority of articles use the same measure to evaluate memory (i.e. percentage of time spent in freezing behavior during a test session), thus allowing one to compare effects across different studies. As we will analyze studies dealing with different interventions, we will not be interested in reaching an effect estimate, but rather in describing the distribution of effect sizes and statistical power across these interventions.

Based upon these findings, we will evaluate how effect size and power are correlated among themselves, as well as with the outcome of significance tests. We will also test whether some aspects of experimental design (e.g. type of intervention, type of conditioning, species used) as well as some measures to control bias (e.g. randomization, blinding) are associated with differences in reported effect sizes or variances. Finally, we will investigate whether effect size and power correlate with the way experimental findings are discussed and published in the literature. These analyses will be mostly correlative, and no causal link should be inferred between specific variables. Nevertheless, they should provide interesting hypotheses that can later be tested formally in experimental settings. The description of the protocol will follow the standardized format proposed by de Vries et al.8

Objectives

Specify the disease/health problem of interest

The problem of interest is to assess how common practices in data analysis can affect the conclusions reached by studies in a specific area of basic science: in this case, the neurobiology of memory in rodents. Our hypothesis is that insufficient consideration given to effect sizes and statistical power has a major impact on the field’s reliability; thus, we will evaluate (a) the distribution of these variables and (b) how much they influence the interpretation and publication of results in the field, based on a representative sample of recent papers.

Specify the population/species studied

We will focus on rodent fear conditioning, which is probably the most widely used model of a simple associative learning task in animals,9 both for studying basic memory processes and for preclinical research (for example, to study cognitive impairment in Alzheimer’s disease models). It provides a simple assessment of aversive memory, and although protocols can vary (for example, by pairing the aversive stimulus with a visual/auditory cue or with a specific context,10 the vast majority of studies use the same measure of assessment (i.e. the percentage of time spent freezing in a test session in which the animal is re-exposed to the conditioning cue).

Specify the intervention/exposure

Since we are interested in investigating the distribution of effect sizes and statistical power across the fear conditioning literature in general, we will not focus on a single intervention, but rather on any intervention (i.e. pharmacological, genetic, surgical or behavioral) tested for its effect on acquisition or consolidation of a fear conditioning memory. We will not include interventions targeted at disrupting established conditioned memories, or at modulating retrieval, extinction or reconsolidation of fear memories. Moreover, we will use only individual (i.e. non-combined) interventions, in which a clear control group is available for comparison.

Specify the outcome measures

We will only include studies that use the percentage of time spent freezing in a test session (undertaken after acquisition of the task) as a measure of conditioned memory. In case multiple test sessions are performed, we will include only the first one to be performed.

State your research question

What is the distribution of effect sizes and statistical power in the fear conditioning literature, and how do these two variables affect the outcome of significance tests, the interpretation of findings and the publication of results?

Methods

Search and study identification

Identify literature databases to search—We will base our literature search on PubMed, including all articles published online in the year of 2013, in order to provide a relevant sample of the contemporary fear conditioning literature.

Define electronic search strategy—We will conduct an electronic search in PubMed for “fear conditioning” AND (“learning” OR “consolidation” OR “acquisition”) AND (“mouse” OR “mice” OR “rat” OR “rats”) to obtain all articles published between January 1st and December 31st 2013.

Identify other sources for study identification—Since our systematic review does not aim to be exhaustive (i.e. it is meant to provide a time-restricted representative sample of fear conditioning articles, not the full literature on the subject), we will not pursue other sources for study identification.

Define search strategies for these sources—Not applicable.

Study selection

Define screening phases—Titles and abstracts will be scanned for articles written in English and describing original results from studies using fear conditioning in mice or rats. Experiments from these papers will undergo full-text screening and will be included in the review if they (a) describe the effects of a single intervention on fear conditioning acquisition or consolidation, (b) have a proper control group to which the experimental group is compared, (c) use freezing behavior in a test session as a measure of conditioned fear and (d) have available data on mean freezing, standard deviation and sample size of each experimental group and on the statistical significance of the comparison

Specify number of observers per screening phase—One of two independent reviewers (T.C.M. and C.F.D.C.) will scan titles and abstracts to select papers for further scrutiny that: (i) are written in English; (ii) present original results; and (iii) describe experimental procedures involving fear conditioning in mice or rats. If these criteria are met, the full text of the article will be obtained and analyzed for inclusion. Articles screened for data extraction by one investigator will be analyzed by the other (thus providing an opportunity to cross-check criteria for all included articles), and random samples of 10% of articles will be checked by both investigators to verify agreement levels on data inclusion through kappa coefficients. Any disagreements will be solved via discussion and consensus, with the participation of a third investigator (O.B.A.) when necessary.

Inclusion and exclusion criteria

Type of study—Inclusion: Original articles including fear conditioning experiments.

Exclusion: Reviews; conference proceedings; original articles not involving fear conditioning.

Type of animals/population—Inclusion: Mice and rats of all strains, including transgenic animals.

Exclusion: All other animal species.

Type of intervention (e.g. dosage, timing, frequency)—Inclusion: Any individual intervention undertaken prior or up to 6 h after fear conditioning, which could thus affect acquisition or consolidation of the task,11 in which the experimental group is compared to a control group in a test session.

Exclusion: Combined interventions; interventions undertaken more than 6 h after fear conditioning (in order to affect retrieval, reconsolidation, extinction or systems consolidation of the task); interventions without a control group.

Outcome measures—Inclusion: Percentage of time spent freezing in a test session undertaken at any time after training (when more than one test session is performed, the first one will be used).

Exclusion: All other measures of conditioned fear (e.g.. fear-potentiated startle, latency in inhibitory avoidance protocols). Test sessions in which the total percentage of time spent freezing in the test session is not recorded or not compared between groups.

Language restrictions—Inclusion: Articles with the full text written in English.

Exclusion: Articles in all other languages.

Publication date restrictions—Inclusion: Articles with online publishing dates in PubMed between January 1st, 2013 and December 31st, 2013, including those with print publication in 2014 or later.

Exclusion: Articles with other online publishing dates, including those with print publication in 2013 but published online in 2012 or earlier.

Other—Inclusion: Articles describing the mean and standard deviation (or standard error of mean) of freezing percentages and sample size for both the intervention and control groups, either in text or graph format, as well as the statistical significance of the comparison between both groups. Articles not describing sample sizes for individual groups (e.g. when pooled sample sizes or ranges are described) will be used for effect size calculations, but not for statistical power calculations, as these cannot be accurately performed in this case.

Exclusion: Articles in which these values cannot be obtained for both the intervention and control groups.

Sort and prioritize your exclusion criteria per selection phase—Screening phase (title/abstract):

1. Not an original article
2. Not in English
3. Not using fear conditioning
4. Not in rats or mice
5. Online publishing date not in 2013.

Selection phase:

1. Full-text of the article not available.
2. No intervention targeted at acquisition or consolidation of fear conditioning.
3. Combined interventions only.
4. Lack of a control group.
5. Mean, standard deviation, sample size or statistical significance data unavailable.

Study characteristics to be extracted

Study ID—First author, title, journal, impact factor (as per the 2013 Journal Citation Reports), number of citations (at the end of the study period), country of origin (defined by the corresponding author's affiliation).

Study design characteristics—Number of experiments using fear conditioning, experimental and control groups in each experiment, sample size for each group, statistical test used to compare groups.

Animal model characteristics—Species (rats vs. mice), type of fear conditioning protocol (contextual vs. cue), gender (male vs. female vs. both).

Intervention—Type of intervention (i.e. pharmacological, genetic, behavioral or surgical), intervention target (i.e. molecule or physiological mechanism affected), timing of intervention (i.e. pre-training, post-training or both), anatomical site of intervention (i.e. systemic or intracerebral)

Outcome measures—We will extract mean and standard deviation (or standard error of mean) for freezing levels (in %) for both the experimental and control groups, which will be used to calculate effect size and statistical power, based on the pooled standard deviation and sample size. We will also extract data on the statistical significance of each comparison.

Furthermore, we will also assess the effect description included in the text of the results session of the articles. For experiments in which significant differences are found, descriptions will be classified as depicting strong effects (e.g. intervention “blocks” or “abolishes” memory formation), weak effects (e.g. intervention “slightly impairs” or “partially impairs” memory formation) or effects of uncertain magnitude (e.g. intervention “decreases”, “lowers”, or “significantly decreases” memory formation). For experiments in which significant differences are not found, descriptions will be classified as depicting similarity (e.g. “similar” or “undistinguishable” levels of freezing), a trend of difference (e.g. a non-significant decrease or increase in freezing levels) or no information on the presence or absence of a trend (e.g. no significant differences were found).

Classification of the terms used to describe effects will be based on the average results of a blinded assessment of terms by a pool of at least 10 researchers with (a) experience in behavioral neuroscience and (b) good fluency in the English language. Categories will be given a score from 0 to 2 in order of magnitude (i.e. 0=weak, 1=neutral, 2=strong for significant results; 0=trend, 1=neutral, 2=similar for non-significant results), and the average results for all researchers will be used as a continuous variable for analysis.

Assessment of risk of bias (internal validity) or study quality

Number of reviewers—Study quality assessment will be performed by one investigator per study (T.C.M. or C.F.D.C.).

Study quality assessment—Scoring for study quality measures will be based on applicable criteria proposed by the CAMARADES checklist12 as well as by the ARRIVE guidelines.13 We will assess the following items: (a) randomization of animals between groups, (b) blinded and/or automated assessment of outcome, (c) presence of a sample size calculation, (d) adequate description of sample size for individual experimental groups in fear conditioning experiments (e) statement of compliance with regulatory requirements, (f) statement regarding possible conflict of interest and (g) statement of compliance with the ARRIVE guidelines for study reporting. For correlation with article-level metrics, we will also compile the individual variables into a 7-point quality score, with 1 point scored for each item. We consider this score to be semi-quantitative, as not all measures necessarily have the same value in assessing quality; nevertheless, it is useful to prevent an excessive number of secondary analyses. For cases in which one of the criteria is not applicable (e.g. randomization for transgenic animals), the score will be normalized according to the number of remaining items to allow comparison with other articles.

Collection of outcome data

Experiment-level data—Most of our data will be analyzed using the individual experiment as the observational unit. For each experiment, we will extract the following continuous variables:

- Mean, standard deviation (SD) and sample size for each group;
- Effect size of treatment (expressed as percentage variation in freezing from the control to the treated group);
- Normalized effect size (expressed as the percentage variation in freezing from the group with the highest freezing level to that with the lowest one).

- Statistical power curves, showing the power to detect a range of differences in a Student's *t* test comparison between control and intervention groups, considering the sample size and pooled standard deviation of each experiment.

We will also extract the following categorical variables:

- Statistical significance of the comparison (we will extract exact *p* values when available – however, since these are frequently not described, we will treat significance as a dichotomous variable).
- Intervention category (pharmacological, genetic, surgical or behavioral);
- Statistical test used;
- Type of conditioning (contextual vs. cued)
- Species (mice vs. rats)
- Gender (male vs. female vs. both)
- Site of intervention (systemic vs. intracerebral)
- Timing of intervention (pre-training vs. post-training vs. both)
- Effect description (from results session), as a phrase or term – each specific descriptor will later be converted to a continuous variable as described above.

Article-level data—Parts of the data will also be analyzed at an article level. For each article, we will extract the mean values obtained for all experiments or for different classes of experiments (e.g. memory-impairing vs. memory-enhancing vs. non-effective interventions), as detailed in the data analysis section.

Moreover, we will also obtain four additional article-level metrics.

- Impact factor of the journal in which the study was published, as obtained from the 2013 Journal Citation Reports.
- Number of citations of the article at the time the analysis is finished, as obtained from ISI Web of Knowledge.
- Study quality assessment to measure risk of bias of the article, as detailed above.
- Region of origin (Northern America, Latin America, Europe, Africa, Asia or Australia & Pacific), as defined by the corresponding author's affiliation.

Methods for data extraction—Numerical values will be obtained from the text or legends when available, or directly from graphs when necessary using the Gsys 2.4.6. software (Hokkaido University Nuclear Reaction Data Centre). In a preliminary analysis, we have found that values extracted by this method are very close to those given in the text, with a correlation of $r > 0.99$ between both approaches.

Number of reviewers extracting data—Each experiment will be selected for analysis by one investigator (T.C.M. or C.F.D.C.), with data analyzed by the other one. Thus, each article included in the analysis will ultimately be examined by both investigators. Any discrepancies or disagreements among them will be solved via discussion and consensus between them and a third investigator (O.B.A.)

Data analysis/synthesis

Data combination/comparison—Since we are interested in the statistical distribution of effect sizes of different interventions on fear conditioning, we do not feel that combining the data in a meta-analysis is feasible, as the idea of a summary effect estimate for diverse interventions makes little sense. However, we will hereby detail the ways in which our data will be analyzed after collection, to ensure that this will be carried out as planned *a priori*. To verify its feasibility, our analysis plan has been tested on a pilot analysis of 30 articles (around 18% of the data), which has helped us to refine our original proposal.

1. Selection of articles

A study flow diagram describing the selection of articles will be provided, detailing (a) the number of articles screened, (b) the number of articles selected at the first screening, (c) the number of articles and experiments selected for inclusion and (d) the number of articles excluded from the analysis and the reasons for exclusion.

2. Distribution of effect sizes

For each individual experiment comparing freezing levels between a treated group and a control group, we will calculate effect size as a percentage of the control value. We will then classify these experiments as memory-impairing (i.e. treatments in which freezing is significantly higher in the control group), memory-enhancing (i.e. treatments in which freezing is significantly higher in the treated group) or non-effective treatments (i.e. treatments in which a significant difference between groups is not observed in the statistical analysis used), and examine the distribution of effect sizes for the three types of experiments, providing means and 95% confidence intervals for each of them (as well as for the aggregate of all studies).

For normalization of positive and negative effect sizes (which are inherently asymmetrical, as they are defined as ratios), we will calculate a normalized effect size, expressed in terms of percentage of the group with the highest freezing levels (i.e. the control group in the case of memory-impairing interventions, or the treatment group in the case of memory-enhancing interventions). We found this approach, previously proposed by Vesterinen et al.,14 to be preferable to other forms of normalization (i.e. log-ratios) in our pilot analysis, as it led effect sizes to be more constant across different freezing levels. Since control levels are usually set at a lower baseline when memory-enhancing interventions are tested, basing calculations on control freezing levels led to higher effect sizes and unrealistic statistical power estimates for this class of studies.

Our primary analysis will use the individual experiment as an observational unit, acknowledging the limitations that (a) articles with multiple experiments will be overrepresented in the sample and that (b) experiments in which two treated groups use the same control group will lead to a small degree of data duplication, which will be quantified and reported. To address the first point, we will also provide an article-level analysis as supplementary data, using the mean effect size for each class of experiments (impairing, enhancing and non-effective) in an individual article as the observational unit.

3. Power calculations

Based on the effect size and standard deviation of each comparison, we will build power curves for each individual experiment, showing how power varies according to the difference to be detected for α=0.05, based on each experiment's variance and sample size. The distribution of statistical power curves will be presented for memory-enhancing, memory-impairing and non-effective interventions (as well as for the aggregate of all studies). As performed for effect size, we will also provide an article-level analysis as supplementary data.

Although actual power for individual experiments will vary according to each intervention's effect size, we will also try to estimate power for a "typical" effect size for an effective intervention, using the mean normalized effect size for interventions with significant effects in our sample as the difference to be detected. This can be thought of as an upper-bound estimate of the average effect size in fear conditioning experiments, since the calculation is likely to exclude some interventions with real effects in which non-significant results were due to insufficient statistical power (i.e. false negatives): thus, the true average effect size for effective interventions is likely to be smaller. However, as we have no way of differentiating these interventions from those with no effect (i.e. true negatives), we consider that using only significant results will provide the best estimate of the average effect size in fear conditioning studies (which in our pilot analysis was 43%).

We acknowledge that this power calculation is only a rough estimate of the true power of each individual experiment (which will inevitably vary according to the actual effect size of the intervention being tested). Nevertheless, since this estimate will be based on the same expected difference for all experiments, it avoids the limitations associated with "post-hoc" power calculations based on observed differences for individual experiments (which are known to be largely circular).15 Thus, we believe that it is a valid approach for comparing power across experiments and/or correlating it with other variables. Moreover, reporting this power estimate along with the power curves should provide a useful example of how to translate these curves into a meaningful number, something that might be important for readers who are less familiar with statistics.

Since we will calculate power using the variances of each individual experiment (which are themselves subject to random variation), we also acknowledge that our calculations will have a degree of sampling error. However, we consider this to be a better approach than using the mean variance of all experiments for the power calculations, as different protocols and laboratories might have very different levels of variance in their experiments. Thus, experimentally calculated variances are likely to be better estimates of an individual laboratory's variance than an overall average. Nevertheless, we will use the median and interquartile boundaries of the observed variance (i.e. the 25th, 50th and 75th percentiles) across all experiments to build power curves and estimates for fear conditioning experiments with different sample sizes. These will be presented as supplementary data, as they could provide a useful rule of thumb for estimating sample sizes for fear conditioning experiments.

4. Effect size / statistical power / mean freezing correlations

To examine whether normalized effect size is related to statistical power (as estimated by the method described above), we will perform a linear correlation between both values, obtaining Pearson's coefficients for the whole sample of articles, as well as for the subgroups of experiments with statistically significant and with non-significant results. We note that a correlation is to be expected mathematically when significant and non-significant articles are analyzed separately, due to the influence of statistical power on the outcome of significant testing. However, this is not necessarily the case when all effect sizes are analyzed together, as the individual experiment's effect size should have no impact on its power (which will vary according to its sample size and variance, neither of which is mathematically expected to correlate with effect size).

We will also perform a correlation between effect size and mean sample size, in an approach that has been proposed as an indirect measurement of publication bias (which is expected to lead to a significant negative relationship).16 However, we note that, since we are analyzing experiments rather than articles, the presence or absence of a correlation in this case will refer to experiments within articles. As the negative results included might be published alongside positive ones, a lack of correlation in this case should therefore not be taken as evidence for absence of publication bias at the article level. Conversely, the presence of a correlation might reflect not only publication bias, but also selective reporting of positive experiments within articles.

Finally, we will correlate freezing levels (using the group with the highest mean as the reference, as done in the normalization of effect sizes) with effect size and statistical power, to evaluate whether the presence of high or low freezing levels might also be a source of bias in determining the chances of a particularly experiment being statistically significant.

5. Comparison of effect sizes across different conditioning protocols, species and genders

To examine whether effect sizes differ systematically across different conditioning protocols, species and genders, we will divide experiments between those using (a) cued or contextual fear conditioning, (b) mice or rats and (c) males, females or both. We will then compare the normalized effect size distribution between protocols, species and gender using Student's *t* test, to test whether systematic differences are observed. Since different protocols/species/genders could also differ in inter-individual variability (even though absolute effect sizes might be similar), we will compare the coefficient of variation (defined as pooled standard deviation/mean across groups) of individual experiments in each condition, as this is also relevant to test whether a specific protocol/species/gender might be associated with greater statistical power in fear conditioning experiments.

Although these analyses might yield interesting associations between larger effect sizes or variances and particular types of conditioning, gender or species, one should keep in mind that they should not be taken to imply a causal relationship between a specific protocol and larger or smaller effect sizes. There are multiple confusion biases that can lead to such correlations, including interventions with large/small effect sizes being preferentially tested in a given protocol, or specific research groups who tend to perform the task in a particular manner testing interventions with particularly large/small effect sizes.

6. Comparison of effect sizes across different types of interventions

To examine whether effect sizes differ across different interventions, we will divide experiments between those using (a) surgical, pharmacological, genetic or behavioral interventions, (b) systemic vs. intracerebral interventions and (c) pre-training vs. post-training interventions. Again, we will compare the distribution of normalized effect sizes and coefficients of variation among experiments between different groups. Once more, care should be taken not to interpret any detected associations as necessarily causal in nature.

7. Correlation between effect size/statistical power and effect description

To examine whether the effect size and statistical power of experiments correlate with the way they are described in the articles, we will correlate each experiment with a description score based on the analysis of the text describing the finding by multiple investigators (see Outcome Measures section). This score reflects how description varies from “weak” to “strong” effects (for significant results) and from “trend” to “similar” effects (for non-significant results). The effect size and statistical power of each significant result will be correlated with its corresponding description score, and the same will be done for non-significant results.

8. Correlation between effect size/statistical power/percentage of significant results and risk of bias indicators

To study whether indicators influencing the risk of bias of a study (randomization, blinding, sample size calculations, sample size description, statement of compliance with ethical regulations, statement of conflict of interest, and statement of compliance with the ARRIVE guidelines) correlate with effect size and power, we will compare (a) the mean normalized effect size for effective interventions, (b) the percentage of experiments with significant results and (c) the mean statistical power of articles with and without each one of these measures. We will perform this analysis using articles as an experimental unit due to the fact that, unlike experiment-level variables (e.g. protocol, gender), indicators of risk of bias are obtained at the article level. Since averaging all effect sizes in an article (which may include interventions with positive as well as negative results) would make little sense, we chose to use both the mean effect size for effective interventions and the percentage of experiments with significant results as summarizers, as they describe separate dimensions (the fraction of experiments with positive results and the average effect size in these experiments) that cannot be captured in a single number. Once again, any associations between variables should be considered correlative rather than causal in nature.

9. Correlation between effect size/statistical power/study quality score and impact factor/number of citations/region of origin of articles

To examine whether effect size, statistical power and methodological issues correlate with the citation metrics of individual articles, we will correlate (a) the mean normalized effect size of effective interventions, (b) the percentage of experiments with significant results, (c) the mean statistical power of experiments and (d) the combined 7-point study quality score of each article with both its journal's impact factor (using the 2013 Journal Citation Reports) and its number of citations at the end of the review period, obtaining Pearson's coefficients for each correlation. Moreover, to assess whether metrics (a) to (d) correlate with the region of origin of the paper, we will compare the four variables among articles originating from the six geographical regions chosen. Again, the option for article-level metrics is justified by the fact that impact factor, number of citations and region are extracted for articles and not experiments.

Statistical analysis—For our primary outcomes, namely the distribution of effect sizes and statistical power across experiments (steps 2 and 3 above), we will present the whole distribution of values and/or curves across experiments in the figures, as well as the mean and 95% confidence intervals.

As for secondary outcomes, comparisons between effect sizes, statistical power or coefficients of variance among different groups of experiments (steps 5 and 6) will be performed using either Student's *t* test (when there are only two groups) or one-way ANOVA with Tukey's post-hoc (when there are more than two groups), using a 0.05 significance threshold adjusted for the total number of experiment-level comparisons performed (12 in total) using the Holm-Sidak method. For experiment-level correlations between quantitative variables (steps 4 and 7), Pearson's correlation coefficients will be obtained for each individual correlation, using a 0.05 significance threshold, also adjusted for the total number of experiment-level correlations performed (8 in total).

For article-level group comparisons, the same approach will be used, using *t* tests to compare studies with/without each quality indicator (step 8) and one-way ANOVA with Tukey's post hoc to compare studies from different regions (step 9), with the 0.05 significance threshold adjusted for the total number or article-level comparisons performed (25 in total). For article-level correlations between quantitative variables (step 9), Pearson's correlation coefficients will be obtained for each individual correlation, and the 0.05 significance threshold will be adjusted for the total number of article-level correlations performed (8 in total).

For all statistical analyses, we will report exact *p* values for comparisons as well as 95% confidence intervals of effect size/power estimates, differences and correlation coefficients.

Power / confidence interval calculations—Based on our preliminary sample, we will be able to include around 39% of screened articles, with a mean of 3.69 experiments/article. Thus, our estimate for the 395 articles detected in our PubMed search would be to include around 153 articles and 564 experiments in our final sample, of which we expect around 160 to be memory-impairing, 61 to be memory-enhancing and 343 to be non-significant. Based on these numbers (and on the mean effect sizes and variances obtained in our preliminary analysis), we expect to estimate mean normalized effect sizes and mean statistical power for all experiments (our primary outcome) with 95% confidence intervals of ± 2%. For the mean normalized effect size of impairing, enhancing and non-significant interventions, 95% confidence intervals are expected to be ± 3%, ± 5% and ± 2%, respectively.

As for comparisons between groups of experiments, statistical power to detect 20% differences in effect size between different types of conditioning, species and sites of intervention are expected to be between 0.90 and 0.95 at α=0.05, and between 0,61 and 0.78 at α=0.004 (the most stringent threshold using our Holm-Sidak correction for the number of group comparisons), again on the basis of our preliminary data. For gender, type of intervention and timing of intervention (in which some categories – such as female animals and post-training interventions – are less common than others), power is expected to be between 0.71 and 0.85 at α=0.05 and between 0.33 and 0.52 at α=0.0043.

For experiment-level correlation analyses, we expect statistical power to detect a moderate correlation of r=±0.3 to be above 0.97 for all analyses, even after correcting for family-wise error with the Holm-Sidak approach at α=0.006. For the correlations involving article-level data, statistical power should be 0.99 for α=0.05, and 0.91 for α=0.006, as sample sizes at the level of articles will be smaller than at the level of experiments.

Funding Information

This work is supported by FAPERJ grants E-26/111.277/2014 and E-26/201.544/2014 to O.B.A., by RCUK NC3Rs grant NC/L000970/1 to M.R.M., and by CNPq scholarships to T.C.M. and C.F.D.C.

References

1. Nuzzo R. Scientific method: statistical errors. Nature. 2014; 506:150–152. [PubMed: 24522584]

2. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev. 2007; 82:591–605. [PubMed: 17944619]

3. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2:e124. [PubMed: 16060722]

4. Macleod MR, McLean AL, Kyriakopoulou A, et al. Risk of bias in reports of in vivo research: a focus for improvement. PLoS Biol. 2015; 13:e1002273. [PubMed: 26460723]

5. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS One. 2009; 4:e7824. [PubMed: 19956596]

6. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013; 14:365–376. [PubMed: 23571845]

7. Sanes JR, Lichtman JW. Can molecules explain long-term potentiation? Nat Neurosci. 1999; 2:597–604. [PubMed: 10404178]

8. de Vries R, Hooijmans CR, Langendam MW, et al. A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies. Evid Based Preclin Med. 2015; 2:1–9.

9. Maren S. Neurobiology of pavlovian fear conditioning. Annu Rev Neurosci. 2001; 24:897–931. [PubMed: 11520922]

10. Phillips R, LeDoux J. Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. Behav Neurosci. 1992; 106:274. [PubMed: 1590953]

11. Johansen JP, Cain CK, Ostroff LE, LeDoux J. Molecular mechanisms of fear learning and memory. Cell. 2011; 147:509–524. [PubMed: 22036561]

12. Sena E, van der Worp HB, Howells D, Macleod M. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci. 2007; 30:433–439. [PubMed: 17765332]

13. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 2010; 8:e1000412. [PubMed: 20613859]

14. Vesterinen HM, Sena ES, Egan KJ, et al. Meta-analysis of data from animal studies: a practical guide. J Neurosci Methods. 2014; 221:92–102. [PubMed: 24099992]

15. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med. 1994; 121:200–206. [PubMed: 8017747]

16. Kühberger A, Fritz A, Scherndl T. Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. PLoS One. 2014; 9:e105825. [PubMed: 25192357]