eLife

Reviewed Preprint
v1 • July 10, 2026
Not revised

Corresponding author: Olavo B. Amaral. *olavo@bioqmed.ufrj.br*

Competing interests: Kleber Neves and Clarissa Carneiro were hired by the study's funder (the Serrapilheira Institute, a philantropic organization with no financial stake at the results) midway through the project and were thus employed by it during data analysis and manuscript preparation.

Funding: See page 38

Reviewing editor: Peter Rodgers, eLife, United Kingdom

Estimating the replicability of Brazilian biomedical science

The Brazilian Reproducibility Initiative

eLife Assessment

This **important** study assessed the replicability of a selection of lab-based biomedical experiments in papers published by authors based in Brazil. The study adds a unique perspective to the literature on replication, and provides rich data on the approach taken, the outcomes, and the challenges involved in conducting large-scale crowd-sourced research. The evidence supporting the claims is **convincing**, but there is scope for clarifying the presentation of the results and extending the discussion section.

https://doi.org/10.7554/eLife.111001.1.sa4

Abstract

Concerns over the replicability and reproducibility of published research have grown in many research fields, but empirical data to inform policies are still scarce. Biomedical research in Brazil expanded rapidly over the last three decades, with no systematic assessment of the replicability of its findings. With this in mind, we set up the Brazilian Reproducibility Initiative, a multicenter replication of published experiments from Brazilian science using three common experimental methods: the MTT assay, the reverse transcription polymerase chain reaction (RT-PCR) and the elevated plus maze (EPM). A total of 56 laboratories performed 143 replications of 56 experiments; of these, 90 replications of 45 experiments were considered valid by an independent committee. Replication rates for these experiments varied between 20 and 44% according to five predefined criteria. In median terms, ratios between group means were 58% lower in replications than in original experiments, while coefficients of variation were 82% higher. Effect size decrease was smaller for MTT experiments, original results with less variability and those considered more challenging to replicate, while t values for replications were positively correlated with researcher predictions about replicability, and negatively correlated with the rate of publications by the original article's last author. Deviations from preregistered protocols were very common in replications, most frequently due to reasons inherent to the experimental model or related to infrastructure and logistics. Our results highlight factors that limit the replicability of results published by researchers in Brazil and suggest ways by which this scenario can be improved.

Introduction

Concerns over the replicability and reproducibility of published results have increased in many areas of science over the past decade. Basic biomedical science has not been immune to these concerns: reports from the pharmaceutical industry (Begley & Ellis, 2012; Prinz et al., 2011) and prospective assessments of replicability in specific fields, such as cancer biology (Errington, Mathur, et al., 2021), spinal cord injury (Steward et al., 2012) and amyotrophic lateral sclerosis (Scott et al., 2008), have reported success rates ranging from 0 to 40%. Nevertheless, these projects have focused on selected subsets of papers in specific areas of science, and the dimension of the problem in the biomedical literature at large remains unknown.

Although concerns over reproducibility (defined here as reaching the same results when analyzing a set of data) and replicability (defined as finding similar results with new data) (National Academies of Science, 2019; Nosek et al., 2022), many of the proposed solutions to address them are local in scope, as institutions and funders are instrumental in setting up policies

and incentives that can foster reproducible research (Munafò et al., 2017). Actionable data on the replicability and reproducibility of research produced within specific institutions, regions or countries, however, is extremely scarce.

Academic research in Brazil experienced rapid growth between the late 20th and early 21st century, driven by the expansion of higher education (Leta, 2012). Concurrently, the establishment of a national evaluation system for graduate education programs instituted new expectations for productivity, with a strong emphasis on journal-based metrics that has only recently begun to be reconsidered (Barata, 2019). More recently, a series of budget cuts in the late 2010s left a large community of researchers struggling to maintain publication output under a competitive funding scenario (Quintans-Júnior et al., 2020). All of these issues raise concerns about the replicability of published results – something that has not been evaluated systematically in the country.

With this in mind, we set up the Brazilian Reproducibility Initiative, a multicenter assessment of the replicability of experiments published by Brazilian institutions using common laboratory methods (Amaral et al., 2019; Neves et al., 2020). Our assessment of a representative sample of biomedical experiments published by national researchers aims to provide systematic data to raise awareness of the issue and inform scientific policy. In addition to generating data on replicability, the project's first-person, naturalistic approach can identify challenges faced by the laboratories conducting replications, which are drawn from institutions similar to those that produced the original results. Consequently, our findings can offer insights into problems and potential solutions that are relevant to the country's reality, as well as guide future replication efforts in lab biology.

Methods

Lab recruitment

The general rationale for the project has been described previously (Amaral et al., 2019). We began by systematically reviewing a random sample of life sciences articles from Brazil to identify common experimental models and methods (see *https://osf.io/f2a6y*). Based on these findings, we chose 10 experimental methods, using rodents or cell lines as models, to select studies for replication (see *https://osf.io/qhyae*). We then opened a public call for Brazilian labs that could replicate experiments using these methods and models, advertised by email, social media and lectures in conferences and institutions, to which 73 labs initially responded. Based on the expertise of respondents and a feasibility analysis by the coordinating team, we selected 3 outcome assessment methods for replication: the MTT (3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide) reduction assay, the reverse transcription polymerase chain reaction (RT-PCR) and the elevated plus maze test of anxiety (EPM; see *https://osf.io/qxdjt* for details). Three further calls for labs were opened later to fill in specific gaps, leading to another 24 responses.

Experiment selection

We searched the Web of Science for publications from 1998 to 2017 with affiliations in Brazil, using R code to identify the methods listed above (see *https://osf.io/57f8s* for details and *https://osf.io/4sy2g* for code). We then manually screened for articles that met the following criteria:

(a) had at least 50% of authors (including the corresponding one) affiliated with a Brazilian institution;

(b) had at least one quantitative result using one of the selected methods and models that (c) was statistically significant, (d) was mentioned in the title or abstract and (e) used only commercially available materials (for details, see *https://osf.io/u5zdq*). A data extraction step collected information on the biological models, experimental procedures and results for the first comparison in the article that presented a statistically significant difference between two groups. Manuals used for extraction were based on existing guidelines for methods reporting (Bustin et al.,

2009; Kilkenny et al., 2010; NPQIP Collaborative group, 2019), and are available at *https://osf.io/udjr7*, *https://osf.io/tr6xa*, *https://osf.io/7uhb6* and *https://osf.io/rkvtm*. After this step, we excluded experiments that did not report standard deviation, standard error of mean or confidence interval, or had an estimated cost greater than BRL 5,000 (around USD 1,300 at the time) per replication.

We then shared summaries of the eligible experiments' methods with collaborating laboratories, according to their expertise, to confirm their capability to replicate them. Our replication sample comprised the first 20 experiments with each method that could be assigned to three independent labs. If labs withdrew from the project before completing their replications, those experiments were reassigned to other labs within our network or to new labs recruited by additional calls. Further information about (a) the selected experiments, (b) the original article, journal, authors and respective institutions, and (c) the teams performing replications were collected as outlined in *https://osf.io/enjxy*.

Protocol development

For each experiment, the coordinating team transcribed the methodological information available in the original article into a protocol, with missing details left as gaps to be filled. Each lab assigned to replicate an experiment received this protocol, without information about the original publication or results (for details, see *https://osf.io/gsvy2*). They were instructed to fill in protocol gaps and elaborate on procedures when necessary, keeping as close to the original methods as possible (see instructions at *https://osf.io/29vh7*). Adaptations were allowed in cases of necessity (e.g. different equipment or unavailable reagents), or when the replicating lab considered the original protocol inadequate to measure the desired outcome, either due to methodological errors (e.g. incorrect primer sequences) or to significant risk of bias (e.g. lack of a vehicle group). Guidelines used to decide about adaptations are available at *https://osf.io/e7zs9*.

Protocols then received at least two external revisions, usually by another lab in the network that was not performing the same experiment and by a member of the coordinating team, both of whom had access to the original publication (see *https://osf.io/g6ph5*). Instructions to reviewers are available at *https://osf.io/k97r4*. For some protocols, a second coordinating team member replaced the independent lab as a reviewer. Reviewer comments were incorporated into the protocol as additional questions for revision and completion by the replicating labs. This process was iterated until the lab and the coordinating team deemed the protocol complete. Experiments involving animals were submitted to local institutional animal ethics review boards at this stage, and protocols were adjusted if required. After approval, the final version was reviewed once more by an independent member of the coordinating team, who raised additional questions if necessary, and then preregistered at the OSF (see *https://osf.io/vzam6/* for the complete version history of registered protocols). A flowchart of the protocol development process is presented in Figure S1.

Sample size calculations were performed to achieve 95% statistical power to detect the original standardized mean difference in each individual replication (see details at *https://osf.io/wxzr7*). Code and data for the calculations are available at *https://osf.io/vs5rp* and *https://osf.io/atkd7*, respectively.

Authors of the original articles were contacted for missing information and raw data (see contact protocol at *https://osf.io/3dn4x*), but these were intended for future analyses and were not used in developing the replication protocols.

Researcher predictions

Before the start of experiments, an open call was issued to experimental researchers willing to make predictions regarding the success of individual replications. Seventy participants selected one of the three methods and received a survey containing a summary of the 20 experiments associated with that method, including the original result and associated statistics. They were asked to predict (a) the probability of replication success, (b) the expected effect size in the replication, and (c) how technically challenging the replication would be (see details at

https://osf.io/tm76h and full survey at *https://osf.io/29mq5*). Fifty-seven of these researchers also participated in prediction markets, in which they received USD 50 in vouchers to bet on the replication success of individual experiments. An analysis plan for this part of the project, which will be explored in future publications, was preregistered at *https://osf.io/pjhgd*, and anonymized survey data is deposited at *https://doi.org/10.7910/DVN/2RLSMG*.

Experimental procedures

For each experiment, a custom data collection spreadsheet was built by the coordinating team based on the protocol. This was meant to document not only the experimental results but also the execution of each protocol step to make the experiment auditable. Labs received these spreadsheets along with a detailed manual on how to execute and document the replications (*https://osf.io/ackyd*) and on how to deal with protocol adaptations when necessary (*https://osf.io/e7zs9*).

Experimental materials were acquired by the coordinating team and delivered to the lab performing the replication, unless they were already available there. Labs were expected to only use materials within their expiration date; however, due to the frequent postponing of experiments because of COVID-19 restrictions or difficulties with suppliers, the use of reagents with expired validity was allowed if their activity could be clearly demonstrated – either by the method demonstrably working as expected (e.g. PCR kits, molecular weight scales) or by a separate test of biological activity with a prespecified expected result (e.g. a test of cell growth with a particular culture medium). These cases were documented in data collection forms, compiled by the coordinating team and evaluated by the validation committee (see below).

Labs then performed replications and sent the data collection sheets to the coordinating team when experiments were completed. If any difficulties arose, labs were free to contact the coordinating team for orientations. If results suggested a failure to implement the method adequately (Neves & Amaral, 2020), labs were allowed to repeat experiments (see Figure S1), and general guidelines used to deal with these situations were later compiled in *https://osf.io/m35s6*.

Post-experiment debriefing

After an experiment was finished, data collection spreadsheets were sent to the coordinating team, who reviewed them and created a document with questions concerning (a) experimental steps deviating from the preregistered protocol and (b) unclear or missing information. These were sent to the replicating lab, which was asked to confirm the observations on protocol deviations, answer the coordinating team's questions and rate (on a scale of 1 to 5) how much the executed protocol deviated from the preregistered one (data available at *https://osf.io/g9jsp*, *https://osf.io/64whp* and *https://osf.io/dfp3h*).

After undergoing this step and answering further questions by the coordinating team if needed, labs received access to the original article from which the experiment was drawn. They were asked to fill in a form answering whether they felt the original result was successfully replicated, with a justification for their answer. They also rated how relevant the differences between the original protocol and the replication were (on a scale of 1 to 5). Subjective assessment decisions were reviewed by e-mail with the labs if (a) justifications suggested answers concerned the reproducibility of methods rather than results or (b) errors or inconsistencies in replication results were detected and corrected in subsequent steps. Data is available at *https://osf.io/dqwr4*, and more details on each debriefing step are presented at *https://osf.io/xgth2*.

Based on the responses to the first debriefing form, a final notes document on each experiment was compiled by the coordinating team, containing (a) a list of deviations from the original protocol, (b) a list of additional observations on the experiment, (c) a list of reagents used after the expiration date (if any) and (d) clarifications about the data collection spreadsheet.

Validation of experiments

To adjudicate whether the performed experiments constituted valid replications and should be included in the main analysis, a validation committee was formed for each method, including the coordinating team and members of participating labs that used that method. Each experiment was evaluated by three members of this committee, who received the registered protocol, the original article, the final notes document, and (in some cases) the data collection spreadsheets and/or additional material, but with no access to the replication results. Each member of the validation committee received instructions (*https://osf.io/rnemd*) on how to evaluate these documents and was asked to answer how much the executed experiment deviated from the registered protocol (on a scale of 1 to 5) and whether they considered it a valid replication, providing justifications for their answers.

If any of the three evaluators or the replicating lab considered the replication invalid or attributed a deviation rating greater than 3, or if the sum of the 4 evaluation scores was greater than 10, the experiment was discussed by the validation committee to decide whether it should be included in the primary analysis. The committee could also suggest the exclusion of individual experimental units, or the use of an experimental unit different from the one suggested by the lab. Committee members did not participate in the discussion of their own replications. After all experiments were discussed, decisions were reviewed for consistency, and inconsistent decisions were rediscussed to align criteria. Moreover, if any additional issues were raised during the analysis or qualitative assessment steps (see below), experiments were rediscussed by e-mail with the validation committee and decisions were changed if necessary. Figure S2 provides an overview of the debriefing and validation process; for more details, see *https://osf.io/e3fjg*. Decisions are available at *https://osf.io/9ta45*.

Lab self-assessment, project evaluation and qualitative assessment of replications

After the conclusion of the validation process, replicating labs received the final notes documents, validation committee scores and validation decisions. They were then asked to fill in a final debriefing form (see *https://osf.io/xgth2*) where they should answer about their agreement with these scores and decisions. They were also asked to elaborate on reasons for protocol deviations, whether those could have been prevented, and what they would change in the replication protocol if they were to start over. Data for this step is available at *https://osf.io/dp54y*.

In addition to these lab responses, individual lab members were asked to optionally fill in an anonymous project evaluation form, in which they could evaluate the project and their participation, and elaborate on challenges, difficulties, and learning opportunities. The 121 responses to this form (*https://osf.io/gdvam*) were complemented by 18 semi-structured interviews with project participants. Responses to open questions about protocol deviations and project difficulties were categorized by the coordinating team following a taxonomy available at *https://osf.io/5gjb7*. After this, the coordinating team built its own list of difficulties faced by the project. For more on the project evaluation process, see *https://osf.io/nfr6y*.

Finally, each experiment's results were sent to the labs that replicated it, along with the original article, information provided by the original authors (when available), replication protocols and validation decisions. Labs were asked to provide an assessment of the aggregate outcome of the replication, and their responses were discussed in a live meeting involving participants from the involved labs and the replication team. Based on these discussions, the coordinating team prepared qualitative assessments of each replication, which will be analyzed in future publications. More on this step can be found at *https://osf.io/w5z9a*.

Data cleaning and checking

Primary data from the experiments recorded in data collection spreadsheets (optical density measurements for MTT and conventional PCR, Ct values for quantitative PCR (RT-qPCR), and behavioral outcomes for EPM) were manually compiled by the coordinating team into a single

spreadsheet for each method, along with experiment-level metadata (available at *https://osf.io/pwfze*). When inconsistencies in calculations or discrepancies between subsections of the data collection spreadsheets were detected, they were cleared up by written contact with the replicating lab. After this, results were summarized by code for each experiment into spreadsheets and scatter plots and shared with the labs, who were asked to check for discrepancies with their calculations. If issues were found, the lab and coordinating team reviewed them via email until the sources of errors were identified and corrected (see *https://osf.io/58vsx* for details).

Nevertheless, during online discussions for qualitative assessment of replications – which occurred after publication of the initial version of the preprint – we found that several errors in data compilation were not detected by this checking process. The coordinating team then proceeded to check each experiments' data collection sheet individually against the compiled results and used the qualitative assessment meetings to review inconsistencies with the labs. This led to revisions of data in 26 replications (18% of the total), of validation decisions in 7 replications (5%) and of subjective assessments of results in 6 replications (4%), as well as adjustments in documentation for individual experiments. A complete list of changes is available at *https://osf.io/4tu2v*.

Data analysis

The protocol for analyzing replication rates and predictors was initially registered at *https://osf.io/9rnuj*, before any experimental results had been reviewed by the coordinating team. It was updated at a time point when experimental results were available, but no analyses had been performed, to account for steps added after the experiments (such as the validation process). The final dataset used for analysis is deposited at *https://doi.org/10.7910/DVN/ZJRDIV*. Analysis code in R was developed based on this protocol and is available both in the data repository and at *https://github.com/BrazilianReproducibilityInitiative/bri-analysis*. A list of deviations, additions and specifications added to the protocol after the analysis had started is available at *https://osf.io/9hj7t*.

In summary, results from the available replications of each experiment were aggregated by meta-analysis using the R package metafor (Viechtbauer, 2010), with the log-transformed ratio of means as the effect measure for MTT and EPM. For RT-qPCR experiments, we used the mean difference in ΔCt values (which is already in log scale), while for conventional PCR we used the log-transformed relative band density between the gene of interest and the reference gene. The results of these meta-analyses were used for comparison with the original result (also transformed into the natural logarithm of the ratio of means). Paired normalization between units in the experimental and control groups was implemented for all MTT experiments, as absolute optical density values were not considered commensurable across experimental units measured in different days. For PCR experiments, pairing was used only when this was the case in the original study, as described in the methods or inferred from the lack of error bars in the control group (see *https://osf.io/wxzr7*).

Replication rates for the sample were calculated on the basis of 5 dichotomous criteria: (a) whether the original estimate was within the 95% prediction interval of a random-effects meta-analysis of the replication, (b) whether the estimate of a random-effects meta-analysis of the replications was within the 95% confidence interval of the original results, (c) whether the point estimate of a fixed-effects meta-analysis of the replication was statistically significant at $p < 0.05$ and had the same sign as the original result, (d) whether at least half of individual replications were statistically significant in the same direction as the original result and (e) whether at least half of labs considered the original result successfully replicated in their subjective assessment. Prediction and confidence intervals were based on t distributions with degrees of freedom based on the number of experimental units, combined between groups within replications and between replications within meta-analyses using the Welch-Satterthwaite equation (Welch, 1947). Sensitivity analyses use (a) z distributions, (b) the Knapp-Hartung approach to calculate degrees of freedom (Knapp & Hartung, 2003) or (c) a single mean of the experimental units from all

replications rather than meta-analysis (with data for each experimental unit normalized by their respective control unit in paired experiments or by the control mean of the respective replication in unpaired ones). Agreement between replication criteria was calculated by Cohen's kappa for pairs and Fleiss' kappa for the aggregate of the 5 criteria.

These rates were initially calculated for the primary analysis set, which included only the replications endorsed by the validation committee. Rates were also calculated for other sets of replications, namely (a) all replications, (b) all replications considered valid by the lab, (c) only experiments with at least 2 valid replications, (d) only experiments with at least 3 valid replications, (e) only replications achieving 80% post-hoc power on aggregate to detect the original relative difference in means, considering the standard deviation, sample size and correlation between pairs found in the replication (see *https://osf.io/tbkvz* for experiments within each analysis set and results of post-hoc power calculations). The primary analysis used the experimental unit as defined by the validation committee, while sets (a) and (b) above considered the lab's definition. A multiverse analysis using all possible replication criteria, sets of experiments and analysis decisions is presented as a specification curve (Simonsohn et al., 2020).

The correlation between effect sizes from the replication and original experiments was evaluated using Pearson's r, with sensitivity analyses removing a prominent outlier or using Spearman's ρ. Coefficients of variation from the original study were compared to the mean coefficient of variation of its replications using Wilcoxon's signed rank test. The mean absolute difference between the effect sizes of the original study and its replications (both expressed as log ratios of means) was compared to that between individual replications of the same experiment, also using Wilcoxon's signed rank test. These analyses were added after data collection and should thus be considered exploratory. Predictors of replication success at the level of original experiments and at the level of replications were evaluated using Spearman's correlation coefficient as planned in the protocol. A complete list of tested predictors is available in the analysis protocol and list of deviations.

Results

Lab recruitment and selection of experiments

Throughout the project, 97 labs across Brazil applied to replicate experiments in response to 4 public calls. Of these, 75 joined the Initiative at some point and 56 contributed data. The geographic distribution of these labs is displayed in Figure S3 and Table S1, with information on team members available in Table S2.

The selection of experiments for replication and their execution are summarized in Figure 1. A total of 317 experiments using the 3 selected methods (MTT, PCR and EPM) were identified by full-text screening and had protocol data extracted. 173 protocol summaries (*https://osf.io/7kagj*) were shared with participating labs until we could assign 20 experiments with each method to at least 3 labs. Information on selected experiments and their articles of origin can be found at *https://osf.io/cyjsz* and is summarized in Table S3, while replication protocols developed by labs are available at *https://osf.io/vzam6/*.

Concluded experiments

Out of the 180 planned replications, 143 were carried out by labs, covering a total of 56 experiments, of which 34 had three replications, 19 had two replications and 3 had only one (Table S4). 136 of these had enough data (i.e. more than one valid data point in each group) to be used in meta-analyses. Canceled replications were due to inaccessible materials or to labs leaving the project or not finishing replications by the project's deadline.

Experiments submitted by labs then underwent validation by our designated committee. Out of 143 replications, 53 were excluded in this step due to protocol deviations (18), using an inadequate experimental unit (12), insufficient sample size (9), insufficient documentation (6), non-interpretable results (2) or a combination of these reasons (6) (see Figure 1). Results of the

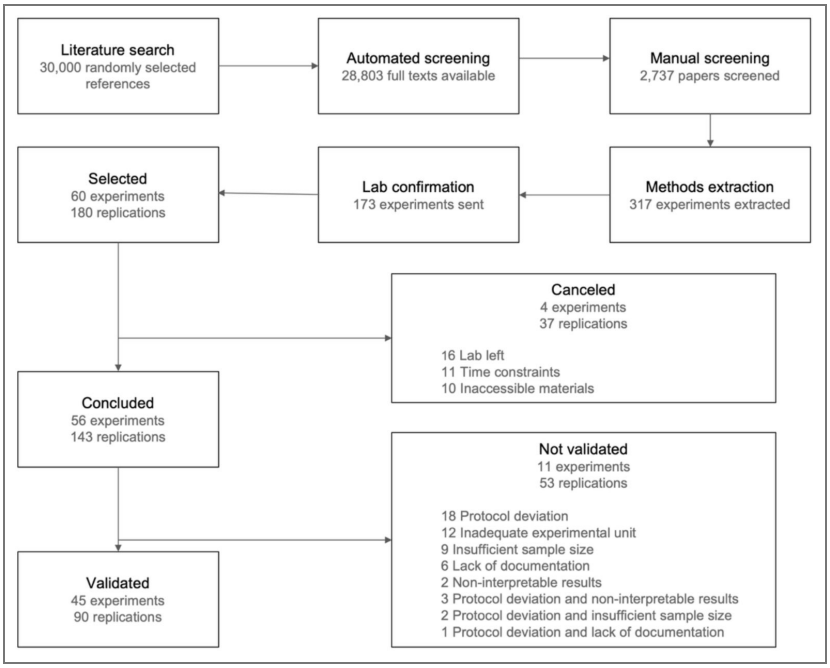

Figure 1. Flowchart describing the selection and execution of replication experiments.

Top row shows the number of articles that underwent automated and manual screening. The methods of selected experiments were extracted, and their summaries were sent to labs to determine the final replication sample. Numbers of concluded and validated experiments (i.e. those with at least one replication) and replications are shown in the bottom rows, with reasons for lack of completion or invalidation shown on the right.

validation process are available at *https://osf.io/9ta45/* and a description of reasons for invalidation can be found in Table S5. Validation decisions for different interpretations of the experimental unit in cell culture experiments are detailed in Table S6. Labs agreed with the committee's decisions on all validated experiments and on 77% of invalidated ones.

This left us with 90 valid independent replications covering 45 different experiments in our primary analysis. Five replications with insufficient sample size and n > 1/group were included in meta-analyses but not analyzed independently. Only 9 experiments remained with three valid replications, while 27 had two and 9 had only one. A list of materials concerning each replication is compiled at *https://osf.io/6av7k/wiki/Individual_Replications/*, while aggregated data for each analysis set is available at *https://osf.io/6av7k/wiki/Aggregated_Replications/*. For the primary analysis, unit-level data are available as tables and scatter plots at *https://osf.io/uzm97/files/*, and forest plots are available at *https://osf.io/sx9gv*. Qualitative assessments of experiments are available at *https://osf.io/p7b9z*.

Replication rates

Replication rates for the 45 experiments included in the primary analysis are shown in Table 1. In 17 out of 39 experiments with multiple replications (44%), the original estimate was within the 95% prediction interval of the replication meta-analysis, while 13 out of 45 aggregate replication estimates (29%) and 23 out of 90 individual replications (26%) were within the original experiment's 95% confidence interval. Nine (20%) aggregate estimates and 17 (19%) individual replications showed a statistically significant effect in the same direction as the original, with half or more of the available replications significant in 10 experiments (22%). Twenty-nine individual replications (32%) were considered successful by the replicating lab, with at least half of labs making this judgment in 19 experiments (42%). That said, many labs used loose criteria for success, such as an effect in the same direction or of comparable size, regardless of statistical analyses (see Table S7). Twelve (27%) experiments were replicated by at least half of the applicable criteria, and agreement between different criteria is shown in Figure S4, with Fleiss' kappa coefficients of 0.54 for the 5 experiment-level criteria and 0.38 for the 3 replication-level criteria.

A major caveat for the interpretation of these replication rates is the low statistical power of some replications. Although we calculated sample sizes for each individual replication to have 95% power to detect the original standardized difference (with the primary analysis including only those achieving at least 80% power for this purpose), coefficients of variation were on average much higher among replications, particularly for PCR experiments (Figure S5). This led statistical power to detect the original relative difference to be lower than planned for many experiments (Figure S6), biasing replication rates based on statistical significance downwards and those based on prediction intervals upwards (as also occurs when there is low agreement among replication results). To address this, we conducted an exploratory *a posteriori* power analysis considering only experiments with at least 80% statistical power for the aggregate of replications (n=35), using the original relative difference and the variability achieved in replications. This leads to an increase in replication rates based on statistical significance and a decrease in those based on prediction intervals, with 38% of original effects within the replications' 95% prediction interval, 31% of replication estimates within the original 95% confidence interval and 26% of experiments statistically significant in the same direction as the original (Table S8).

Table 2 displays replication rates for these and other sets of experiments, including all experiments irrespectively of the validation results (n=56), all experiments considered valid by labs (n=56), and only experiments with at least two (n=36) or three (n=9) valid replications (broken down by method in Tables S9-S12). Disregarding the validation process increases sample size but has little impact on replication rates, whereas analyzing only experiments with multiple replications leads to higher rates by some criteria, but with a lower sample size. Replacing t by z distributions in meta-analyses (Tables S13-S14) decreases replication rates based on confidence intervals, which become narrower, but increases significance rates. Basing degrees of freedom for

By experiment	All	MTT	PCR	EPM
Original in replication's 95% PI	**17/39 (44%)**	8/15 (53%)	8/13 (62%)	1/11 (9%)
Replication in original 95% CI	**13/45 (29%)**	5/17 (29%)	5/14 (36%)	3/14 (21%)
Same-sign significance ($p<0.05$)	**9/45 (20%)**	4/17 (24%)	3/14 (21%)	2/14 (14%)
≥50% replications significant	**10/45 (22%)**	5/17 (29%)	3/14 (21%)	2/14 (14%)
≥50% subjectively replicated	**19/45 (42%)**	8/17 (47%)	7/14 (50%)	4/14 (29%)
By replication	**All**	**MTT**	**PCR**	**EPM**
Replication in original 95% CI	**23/90 (26%)**	7/33 (21%)	6/28 (21%)	10/29 (34%)
Same-sign significance ($p<0.05$)	**17/90 (19%)**	10/33 (30%)	5/28 (18%)	2/29 (7%)
Subjectively replicated	**29/90 (32%)**	12/33 (36%)	11/28 (39%)	6/29 (21%)

Table 1. Replication rates in the primary analysis.

Replication rates for the primary analysis using multiple criteria. Effect size comparisons are based on random-effects meta-analyses, while same-sign significance is based on fixed-effects meta-analysis estimates. The 95% prediction interval criterion only uses experiments with more than one replication and thus has a different sample size. All statistical tests use t distributions based on the number of experimental units. PI, prediction interval; CI, confidence interval; MTT, 3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide assay; PCR, reverse transcription polymerase chain reaction; EPM, elevated plus maze. For more information on replication criteria, see *https://osf.io/9rnuj*.

meta-analyses on the Knapp-Hartung approach (Tables S15-S16) increases replication rates based on prediction intervals but drastically lowers those based on meta-analytical significance, as estimates incorporate more uncertainty. Finally, synthesizing replications by a mean of all experimental units rather than by meta-analysis (Tables S17-18) lowers replication rates, as the precision of aggregate estimates is decreased by individual replications with high variability.

We also evaluated the impact of analysis decisions made for specific methods after the protocol was registered. Pairing MTT experiments only when the original was presumably paired markedly increases coefficients of variation, decreasing statistical power and replication rates based on statistical significance (Table S19). Analyzing relative expression in a linear scale for PCR experiments (Table S20) also leads to increased variability, lower rates of statistical significance and wider prediction intervals. A multiverse analysis for all combinations of experimental sets, replication criteria and analysis options is presented as a specification curve in Figure 2 and shows a median replication rate of 25% (interquartile range = 19-33%).

Effect size and variability comparison

A comparison between the effect sizes from each original experiment and those of replications in the primary analysis is shown in Figure 3A, while the correlation between the aggregate effect size of replications and the original one is shown in Figure 3B. The high linear correlation ($r = 0.82$, $p = 6.4\times10^{-12}$) is due to one prominent outlier with very large effect sizes in both the original and replication; removal of this experiment leaves a weak correlation with $r = 0.22$ ($p = 0.16$; Figure 3C). Using a non-parametric approach for the whole sample also leads to a lower correlation coefficient (Spearman's $\rho = 0.35$, $p = 0.02$).

41 out of 45 effect sizes (91%) were smaller in the replication aggregate than in the original, with a median ratio between the original and replication relative effect sizes (both expressed as ratios of means) of 1.71 (1.51 for MTT, 1.75 for PCR and 2.01 for EPM) (Table 3). This expresses a ratio between ratios, which is used for mathematical purposes to include effects in the opposite direction: median relative differences (80% in the original vs. 8% in the replication when considering effect direction) are 90% smaller in the replications. While this may reflect exaggeration of original effect sizes or publication bias (Ioannidis, 2008), some degree of inflation is expected even in its absence, as our selection process filtered for original experiments with significant differences. In an exploratory analysis, we found that coefficients of variation were lower in the original than in the average of replications in 33 out of 45 experiments (73%), with a median ratio of 0.55 (Wilcoxon's signed rank test, $p = 1.3\times10^{-5}$). This difference was greater in PCR experiments (ratio of 0.28, $p = 2.4\times10^{-4}$) than in MTT and EPM ones (ratios of 0.85 and 0.74, $p = 0.15$ and 0.05, respectively). Sign errors (i.e. significant effects in the opposite direction of the original) were infrequent, occurring in 2 out of 45 experiments (4%) and accounting for 18% of significant replication results.

Our multicenter design also allowed us to compare the variation between individual replications to that between replications and original studies in an exploratory analysis. Relative differences (expressed as ratios between the higher and lower mean) between original effects and individual replications were larger than those between replications of the same experiment, with a median ratio of 1.21 (Wilcoxon signed-rank test between pairs, $p = 0.01$). Differences were larger and more consistent among EPM studies (median ratio = 1.55, $p = 0.007$) than among PCR (ratio = 1.36, $p = 0.19$) or MTT experiments (ratio = 1.12, $p = 0.64$) (Table 3). This suggests that part of the irreplicability observed in our study is due to factors specific to published experiments that were not present in replications; nevertheless, a sizable amount of variation was still observed among replications, as can be seen in Figure 3A.

Predictors of replication success

We analyzed whether multiple factors at the level of the original experiment and replication were correlated with replication success, as measured by the strength of evidence for an effect in the same direction (the t value for the replication or aggregate of replications) and by effect size exaggeration (the difference between the effect sizes of the original experiment and replication,

By experiment	Primary	Lab's choice	All Exps	≥ 2 copies	3 copies	≥80% power
Original in replication's 95% PI	**17/39 (44%)**	22/43 (51%)	25/52 (48%)	16/36 (44%)	4/9 (44%)	13/34 (38%)
Replication in original 95% CI	**13/45 (29%)**	15/56 (27%)	14/56 (25%)	11/36 (31%)	4/9 (44%)	11/35 (31%)
Same-sign significance (p<0.05)	**9/45 (20%)**	13/56 (23%)	12/56 (21%)	9/36 (25%)	2/9 (22%)	9/35 (26%)
≥50% replications significant	**10/45 (22%)**	11/56 (20%)	10/56 (18%)	10/36 (28%)	1/9 (11%)	9/35 (26%)
≥50% subjectively replicated	**19/45 (42%)**	19/56 (34%)	19/56 (34%)	15/36 (42%)	3/9 (33%)	12/35 (34%)
By replication	**Primary**	**Lab's choice**	**All Exps**	**≥ 2 copies**	**3 copies**	**≥80% power**
Replication in original 95% CI	**23/90 (26%)**	31/116 (27%)	35/136 (26%)	21/81 (26%)	11/27 (41%)	21/76 (28%)
Same-sign significance (p<0.05)	**17/90 (19%)**	22/116 (19%)	23/136 (17%)	17/81 (21%)	4/27 (15%)	16/76 (21%)
Subjectively replicated	**29/90 (32%)**	35/116 (30%)	41/136 (30%)	25/81 (31%)	9/27 (33%)	21/76 (28%)

Table 2. Replication rates for different analysis sets.

Replication rates for the primary and secondary analyses. Effect size comparisons are based on random-effects meta-analysis, while same-sign significance is based on a fixed meta-analysis estimate. The 95% prediction interval criterion only uses experiments with more than one replication and thus has a different sample size. Subsets for secondary analyses include all experiments judged valid by the replicating lab (Lab's Choice), all concluded experiments (All Exps) – both of which use the experimental unit as defined by the lab rather than by the validation committee –, only experiments with at least 2 (≥ 2 Reps) and 3 replications (3 Reps), and only experiments with ≥ 80% *a posteriori* power using the original relative difference and the variability achieved in replications. All statistical tests use t distributions based on the number of experimental units. PI, prediction interval; CI, confidence interval. For more information on replication criteria, see *https://osf.io/9rnuj*.

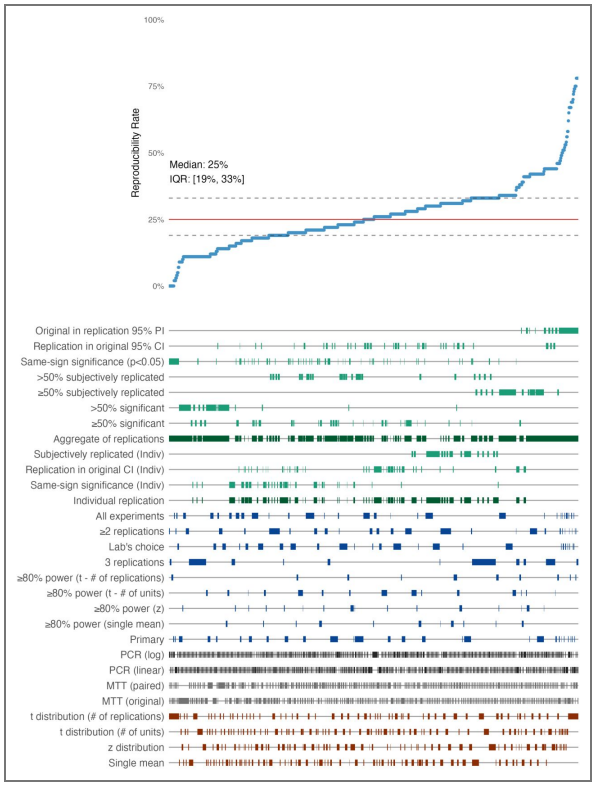

Figure 2. Specification curve analysis of replication rates.

Curve shows replication rates arising from the combination of 10 replication criteria (green; 5 criteria for the aggregate of experiments, with 2 different ways to resolve ties for those involving voting, and 3 criteria for individual replications), 8 sets of experiments with different criteria for inclusion (blue), 2 ways of handling PCR (black) and MTT (grey) data, and 4 different approaches for statistical analysis (red). Replication rates range from 0 to 78%, with a median of 25% and an interquartile range (IQR) of 19 to 33%.

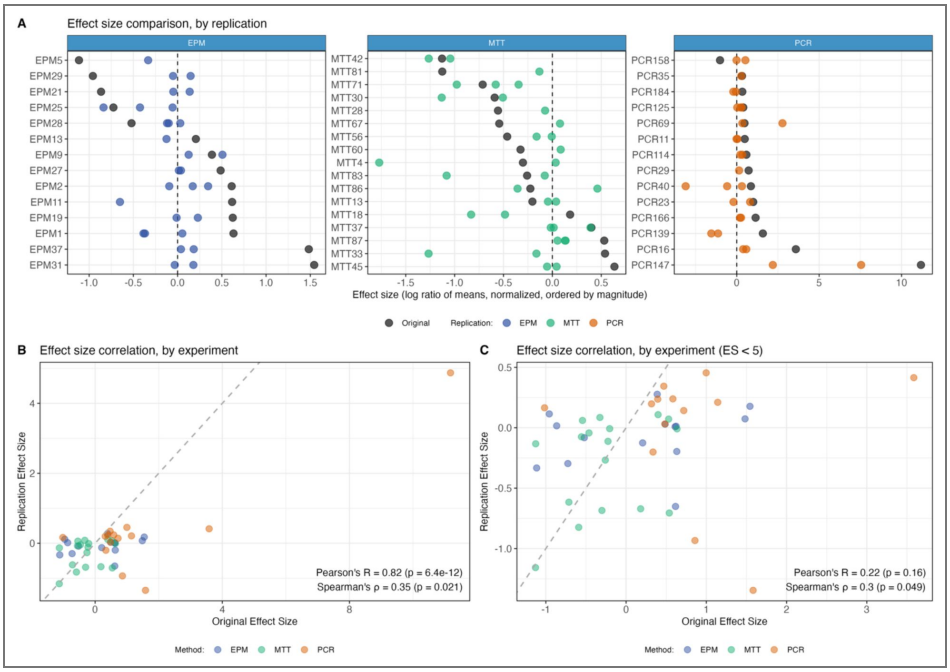

Figure 3. Effect size comparisons and correlations.

(A) Effect size comparison between original experiments (dark) and individual replications (light) for EPM (blue), MTT (green) and PCR (orange) experiments, ordered by the original effect size. X axes represent effect sizes as the natural logarithm of the ratio between the means of the experimental and control groups in behavioral outcomes (EPM), optical density (MTT) or relative gene expression (PCR), with 0 indicating no difference between groups. **(B)** Correlation between the effect sizes of original experiments (x axis) and the aggregate result of their respective replications (y axis), again expressed as log ratios of means ($r = 0.82$, $p = 6.4\times10^{-12}$; $\rho = 0.35$, $p = 0.02$). Colors are the same as in A. Dashed line indicates equivalent effect sizes between original and replication. **(C)** The same analysis excluding the prominent outlier in PCR147, yielding a much weaker linear correlation ($r = 0.22$, $p = 0.16$; $\rho = 0.3$, $p = 0.05$). All results use the primary analysis set and t distributions based on the number of experimental units.

By experiment	All (n=45)	MTT (n=17)	PCR (n=14)	EPM (n=14)
Original effect size (ratio of means)	**1.8** (1.2 - 7x10^5)	1.71 (1.2 – 3.1)	2.21 (1.4 - 7x10^5)	1.87 (1.2 - 4.7)
Replication effect size (ratio of means)	**1.08** (0.3 - 130.6)	1.08 (0.5 - 3.2)	1.23 (0.3 - 130.6)	1.02 (0.5 - 1.4)
Effect size ratio (original/replication)	**1.71** (0.7 - 548.3)	1.51 (0.7 – 3.5)	1.75 (1.1 - 548.3)	2.01 (1.1 - 4.1)
Coefficient of variation (original)	**0.21** (0.01 - 1)	0.08 (0.03 - 0.2)	0.25 (0.01 - 0.5)	0.47 (0.1 - 1)
Coefficient of variation (replication)	**0.49** (0.04 - 1.8)	0.13 (0.04 – 0.5)	0.66 (0.4 - 1.8)	0.67 (0.2 - 1.4)
Coefficient of variation ratio (original/replication)	**0.55** (0.02 – 2.5)	0.85 (0.1 - 2.5)	0.28 (0.02 – 1.4)	0.74 (0.3 - 2.5)
Mean effect size difference (original vs. replications)	**1.67** (1.1 - 550.7)	1.52 (1.1 - 3.5)	2.5 (1.2 - 550.7)	1.67 (1.2 - 4.4)
Mean effect size difference (between replications)	**1.26** (1 - 212.4)	1.26 (1 – 6.1)	1.5 (1.1 - 212.4)	1.23 (1 - 1.7)
Ratio of mean ES differences (orig-rep/rep-rep)	**1.21** (0.3 - 18.5)	1.12 (0.4 - 1.9)	1.36 (0.3 - 18.5)	1.55 (0.8 - 3.5)
Sign error (% of total)	**2/45** (4%)	1/17 (6%)	1/14 (7%)	0/14 (0%)
Opposite sign (total)	**14/45** (31%)	5/17 (29%)	4/14 (29%)	5/14 (36%)

Table 3. Comparisons between results of replications and original studies in the primary analysis.

Continuous variables are shown as median (range), while categorical ones are shown as proportion (percentage). Original effect sizes represent the ratio between the higher and lower mean of the two groups in the original study and are thus always above 1. Replication results respect the same order as the original: therefore, effects above 1 are in the same direction, and those below 1 are in the opposite direction. Ratios between original and replication effect sizes are thus ratios between ratios and are calculated by exponentiating differences between log ratios. Coefficients of variation are calculated as the pooled mean of both groups divided by the pooled standard deviation; for PCR experiments, this is done for relative expression values in linear scale. Mean effect size differences are obtained from absolute differences between effect sizes in log ratio of means, in order to measure discrepancies between the original experiment and its replications or between individual replications. These are also exponentiated and thus represent the mean ratio between the higher and lower value. Sign errors refer to effects in the opposite direction of the original with $p < 0.05$, while opposite sign (total) includes all differences in the opposite direction (irrespective of significance). All results use the primary analysis set and t distributions based on the number of experimental units. ES, effect size; MTT, 3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide assay; PCR, reverse transcription polymerase chain reaction; EPM, elevated plus maze.

both expressed as log ratios of means). A summary of these correlations in the primary analysis is shown as a heat map in Figure 4, while individual scatter plots are available at *https://osf.io/fdpbe*.

At the original experiment level (Figure 4A), replication success as measured by t value was mainly correlated with researcher predictions about the probability of replication success (ρ = 0.46, p = 0.002) (see Table S21 for data on survey participants and Figure S7 for survey results). Predictions were also correlated with dichotomous same-sign significance (ρ = 0.44, p = 0.004), which was the outcome participants were directly asked to predict. Conversely, t values were negatively correlated with the number of publications by the original article's last author in the 5 years preceding the replicated publication (ρ =-0.31, p = 0.04). Effect size exaggeration was greater for experiments with higher coefficients of variation in the original experiment (ρ = 0.41, p = 0.005) – an expected result when original results are filtered for significance –, and lower for MTT (ρ =-0.39, p = 0.008), and cell line experiments (ρ =-0.26, p = 0.08), although these variables are heavily correlated among themselves, making confounding likely (see Figure S8 for correlations among predictors). Exaggeration was also correlated with researcher predictions on how challenging the replication would be to perform (ρ = 0.34, p = 0.03), worse reporting of the experiment's methods (ρ =-0.23, p = 0.13) and, surprisingly, with higher use of bias control measures in EPM experiments (ρ = 0.6, p = 0.02), although the latter analysis had a very low sample size (n=14). Most of these correlations are weak and, due to the multiplicity of tested predictors, could plausibly have arisen by chance.

In an exploratory analysis at the replication level (Figure 4B), t values were negatively correlated with the degree of change between the original article and replication, as assessed by the replicating lab (ρ =-0.26, p = 0.01); however, as this score was attributed after seeing the results, unsuccessful replications could have led researchers to attribute greater value to protocol changes. Effect size exaggeration was positively correlated with this score (ρ = 0.19, p = 0.07), as well as with that attributed by the validation committee, which was blind to the replication results (ρ = 0.27, p = 0.009). More experienced data collection teams had slightly higher t values in replications (ρ = 0.21, p = 0.05), as did groups in higher-ranked universities (replications (ρ = 0.18, p = 0.13), which also found lower effect size decreases between the original study and replication (ρ =-0.22, p = 0.07). Conversely, replicating teams with larger numbers of published articles found greater effect size decreases (ρ = 0.28, p = 0.008 for protocol-developing teams, ρ = 0.15, p = 0.15 for data collection teams). More experienced data collection teams were also more likely to consider the replication successful in their subjective assessments (ρ = 0.31, p = 0.003 for years since first publication and ρ = 0.24, p = 0.02 for number of publications). Once more, the multiplicity of tested predictors and non-independence between replications with overlapping teams make these correlations tentative at best.

Self-assessment by labs and coordinating team

Protocol deviations occurred frequently in our sample and were registered by the coordinating team in 95% of individual replications (88% for MTT, 98% for PCR and 100% for EPM). On a scale of 1 to 5 (with 1 being no deviation and 5 being a deviation that invalidates the experiment as a direct replication), replications in the overall sample had a mean (± SD) rating of 1.9 ± 1 when evaluated by the lab and 2.5 ± 1.1 when rated by the validation committee (excluding replications where issues were found after initial validation, in which scores were not updated). Among experiments included in the primary analysis, these ratings were 1.7 ± 0.8 and 2.1 ± 0.8, respectively. When analyzing all experiments in an exploratory manner (see Figure S8), deviation scores by the committee were higher for animal (ρ = 0.42, p = p = 7.2×10^{-7}) and EPM experiments (ρ = 0.4, p = 1.7×10^{-6}) and lower for cell line and MTT ones (ρ =-0.49 p = 2.9×10^{-9}). Protocol deviations also correlated with higher original coefficients of variation (ρ = 0.42, p = 8.1×10^{-7}) and lower original standardized effect sizes (ρ =-0.36, p = 2.4×10^{-5}) – probably due to confounding of these features with EPM experiments –, as well as with earlier publication years (ρ =-0.37, p = 1.4×10^{-5}) and degree of challenge as assessed in the prediction survey (ρ = 0.31, p = 3.5×10^{-4}).

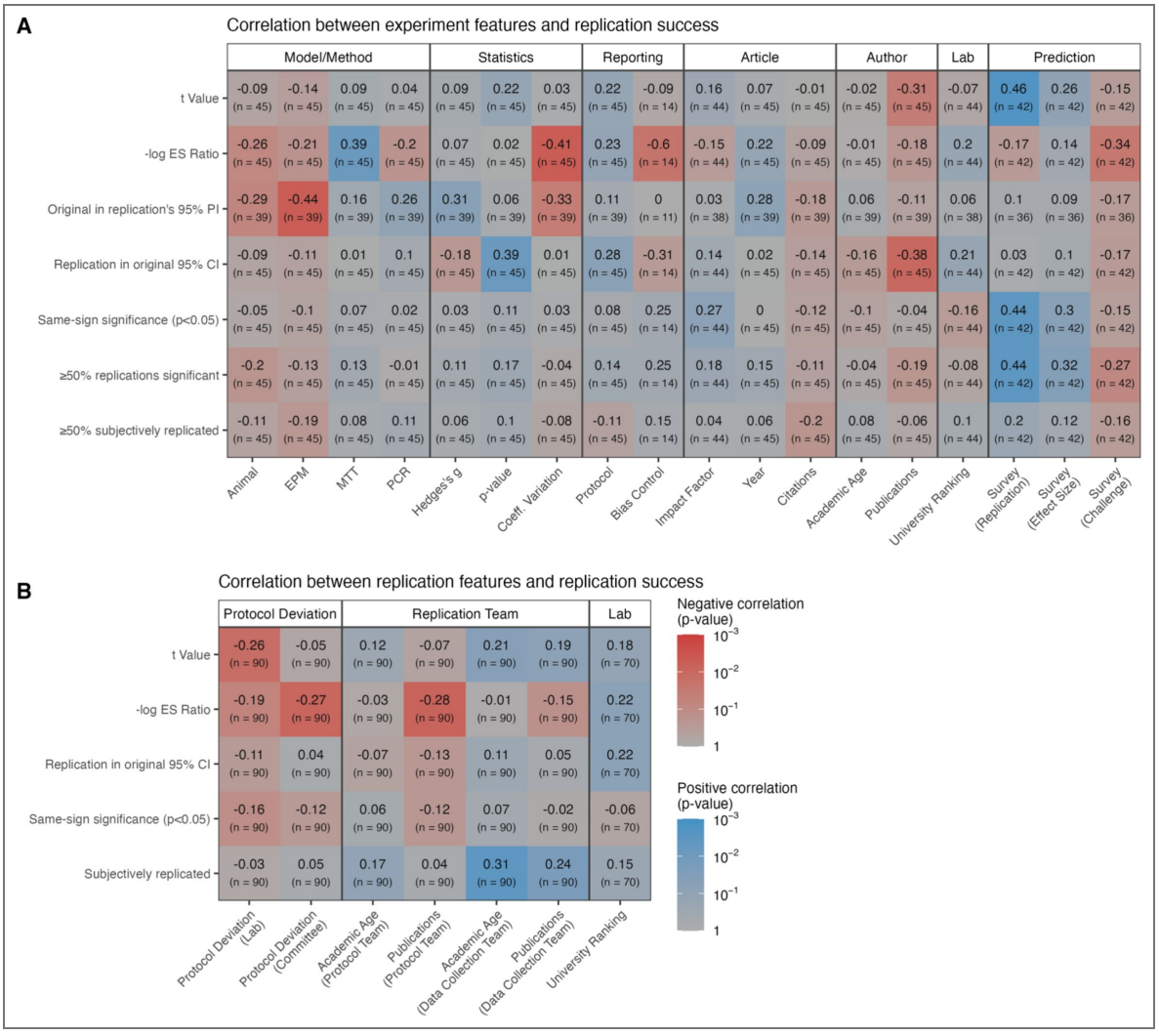

Figure 4. Predictors of replication success.

(A) Predictors of replication success at the level of original experiments (x axis) in the primary analysis. Categories include experimental model and method (animal vs. cell, EPM/MTT/PCR vs. others), statistics (original standardized effect size, p value and coefficient of variation), reporting (protocol reporting score, bias control measures), article features (journal impact factor, year and citations in the first 2 years), author features (years since first publication and number of publications at the time of original article publication), lab features (university ranking) and researcher predictions (about replication probability, expected effect size and difficulty). For details on each predictor, see *https://osf.io/9rnuj*. Y axis represents different continuous (aggregate replication t value, effect size differences in log scale) and dichotomous outcomes (same as in Table 1); effect size differences are expressed as replication minus original (i.e. with a sign opposite to that presented in the text) so that blue indicates correlations with less exaggeration/higher replicability and red with more exaggeration/lower replicability. Numbers show univariate correlations as Spearman's ρ, while color intensity represents p values. **(B)** Predictors at the level of replications, including degree of protocol deviation (judged by the lab and validation committee), features of the replication team (mean number of years since first publication and number of publications by the protocol and data collection teams) and of the lab (university ranking). Other conventions are the same as in A, but outcomes and sample size refer to individual replications. Scatter plots for individual correlations are available at *https://osf.io/fdpbe*.

Reasons for protocol deviations provided by labs are presented in Figure 5, with illustrative examples presented in Table S22. Most deviations were due to issues intrinsic to the experiment, such as the experimental model behaving differently than expected. Reasons related to the infrastructure of the lab and animal facility were also frequent, as were logistical problems involving suppliers or regulatory requirements. A smaller fraction of deviations was due to deliberate choices or errors by the lab performing the replication. An assessment of general difficulties by participating researchers (Figure S9 and Table S23) placed COVID-19 restrictions, delayed delivery of reagents, and difficulties with the experimental model as the top challenges faced by labs.

The coordinating team also assessed the obstacles faced by the project, with top-ranked difficulties shown in Table S24 and a complete list available at *https://osf.io/q76vj*. Lack of common terminologies for describing aspects of experiment design, problems with reagent delivery, poor methodological descriptions in the original articles, and experimental models not working as expected were considered the most relevant difficulties. These were grouped into five categories encompassing the key challenges faced by the project (Table 4). These challenges, along with associated opportunities for improvement, will be discussed more extensively in future publications.

Discussion

In this large multicenter replication of published biomedical experiments from Brazil, replication rates in our primary analysis varied from 20 to 44%, depending on the criteria used to define replicability. These numbers are in a similar range as previous prospective estimates from the biomedical literature (Begley & Ellis, 2012; Errington, Mathur, et al., 2021; Prinz et al., 2011; Scott et al., 2008; Steward et al., 2012). Among these, our study is unique in trying to assess a representative sample of articles by randomly selecting experiments from a wide range of publications based on methods, rather than filtering by research area, importance or citation counts. It is also the first attempt to evaluate replicability within a specific country, in order to provide data that can be of more direct use for local funders, institutions and researchers. That said, the absence of similar studies elsewhere precludes comparisons with other regions.

When compared to originally positive experiments in the Reproducibility Project: Cancer Biology (RP:CB) (Errington, Mathur, et al., 2021) – perhaps the closest counterpart to our study – replication rates in our primary analysis were lower when measured by statistical significance in the same direction (20% *vs*. 39%) or inclusion of the replication effect size in the original 95% confidence interval (29% *vs*. 43%). Correlations between effect sizes were also lower in our study (Spearman's $\rho = 0.35$ *vs*. 0.47), while effect size decreases were similar in magnitude (85% in standardized effect sizes in RP:CB, 90% in relative differences in our study). Multiple factors could account for the differences, including sampling strategy (general literature *vs*. highly cited articles), experimental selection (first significant experiment *vs.* most relevant ones), replication facilities (academic labs *vs*. commercial services), protocol development (with *vs*. without the participation of original authors), subfield (general *vs*. cancer biology) and region of origin (Brazil *vs*. worldwide).

Our study is also unique in employing a multicenter replication strategy, allowing us to assess the extent to which interlaboratory variability may account for discrepancies between the original study and its replications. Differences in effect sizes between the original study and replication were larger than those among replications, suggesting that irreplicability is partly due to factors that are intrinsic to published articles, such as publication bias, selective reporting, low statistical power and incentives for positive results (Smaldino & McElreath, 2016). That said, differences between replications were also large in many cases. As replicating labs were blind to the original result and had no obvious incentives towards particular results, heterogeneity between replications is more likely due to differences in experimental conditions or in how experiments were conducted. These may include protocol adaptations, different levels of adherence to the original protocol, or undetected errors by replicating labs. The finding of significant

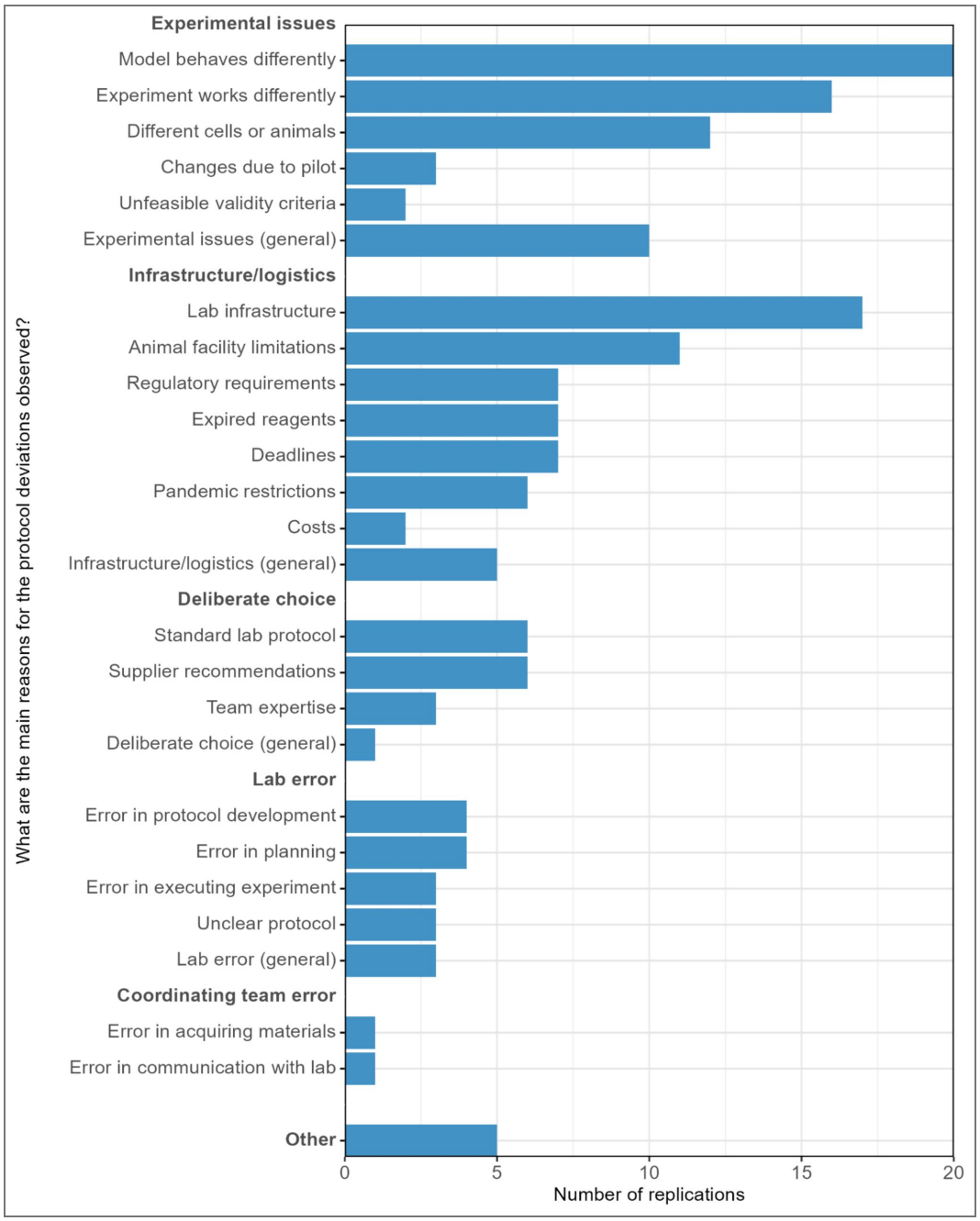

Figure 5. Reasons for protocol deviations.

Reasons were provided as open answers by the lab and categorized by the coordinating team as described at *https://osf.io/5gjb7*. They are divided into general categories (bold) and subcategories. Examples for each category are shown in Table S22.

Category	Examples
Lack of standardized terminologies for experimental design and reporting	Unstandardized/incomplete descriptions of methods and results in original articles. Discrepancies in terminology used to describe issues such as experimental units and controls by replicating labs.
Tension between replicating protocol exactly and getting methods to work	Lack of consensus between essential and adaptable steps in protocols. Experimental models behaving differently across labs. Difficulties in preregistration in the absence of pilot experiments. Difficulties in establishing criteria for methodological validity.
Multicenter project management	Communication difficulties between the coordinating team and participant labs Lack of formal training in project management by the coordinating team. Participant labs not used to external supervision or standardization of procedures. Difficulties in asynchronous review of protocols and data. Lack of data management standards among participant labs.
Logistics and infrastructure	Difficulties in obtaining licenses with regulatory agencies. Inefficiency of national suppliers in delivering reagents. Limitations in animal facilities and general laboratory infrastructure. High prevalence of junior scientists in participating teams.
General difficulties for large-scale projects	Difficulties in anticipating loss of labs and experiments during the project. Difficulties in data management and statistical analysis of complex data. Difficulties in maintaining documentation practices throughout the project.

Table 4. Challenges in performing a large-scale multicenter replication of experiments from Brazilian biomedical science.

Main categories of challenges faced by the project, based on the coordinating team's assessment of an initial list of difficulties (see Table S24 and *https://osf.io/q76vj*).

interlaboratory variability is also consistent with previous attempts to replicate identical protocols across multiple sites for animal (Crabbe et al., 1999) and cell line studies (Elliott et al., 2017; Niepel et al., 2019).

Within the biomedical literature, our study is also one of the few to assess predictors of replicability. Although the power of this analysis was limited by the small number of successful replications, MTT and cell-line experiments exhibited smaller decreases in effect sizes between the original and replication than EPM and animal experiments. Reasons for irreplicability also seemed to differ between methods. In PCR experiments, non-significant results in some experiments were likely explained by coefficients of variation being higher in replications than in original studies, leading to low statistical power. For the other methods, power was usually adequate and non-significant replication effects were closer to zero, suggesting that original results are more likely to represent false-positive or ungeneralizable results.

Although researchers' predictions overestimated replicability, they had reasonable success in forecasting the evidence strength of replications as measured by t values; still, prediction accuracy was lower than that reported by previous replication projects in psychology and the social sciences (Gordon et al., 2021). This finding might be expected, given that the replicability of findings in biomedical science is probably less intuitive than that of psychology studies, for which prediction accuracy is above chance even among laypeople (Hoogeveen et al., 2020). Assessments of how challenging the replications would be, meanwhile, correlated both with experimental complexity measured by text-based assessment of the protocols (Fanelli et al., 2025) and with greater effect size decrease between the original result and replication. Researcher justifications for predictions and analyses of whether prediction markets can improve accuracy will be explored in further work.

Other features of the original article were generally uncorrelated with replication outcome, although large rates of publications by the last author were associated with lower replicability, suggesting that incentivizing publication volume may be counterproductive for the reliability of results. Journals and university rankings were not predictive of replicability, in contrast with a recent retrospective analysis of *Drosophila* immunity studies (Lemaitre et al., 2026), although the focus on randomly selected Brazilian articles led our sample to mostly include low impact factor journals. Counterintuitively, greater use of measures to control risk of bias in animal experiments, such as randomization and blinding, was correlated with greater effect size decrease in the replication. Although this is likely to be a chance finding, as both sample size for the analysis and plausibility are low, it suggests that replicability in lab biology – as well as predictions about it – may rely more on the complexity of experiments than on traditionally used proxies for methodological rigor.

Our analyses have multiple caveats, most of which are due to difficulties faced by participating labs performing replications. Twenty-one percent of planned replications ended up not being performed, either because materials could not be acquired (5%) or because the labs to which they had been assigned left the project or could not meet its deadlines (15%). This is a much smaller fraction than the 74% unfinished replications in RP:CB (Errington, Denis, et al., 2021), likely due to our selection of simpler experiments. On the other hand, our validation process had a large impact on our sample, excluding 37% of replications from the primary analysis. Interestingly, this filtering process did not have a large impact on replication rates, as included and excluded replications had similar rates of success. This may be explained by the fact that, while some issues raised during the validation process would be expected to lower replicability (e.g. deviations from the original protocol), others could artificially raise it (e.g. insufficient biological variation among experimental units or biases in randomization).

Loss of individual replications also limited statistical power for many experiments, even though we aimed for high power in sample size calculations (95% power to detect the original effect in each individual replication) to account for the likelihood of effect size inflation (Ioannidis, 2008; van Zwet et al., 2023). Power was further constrained by the fact that coefficients of variation were higher in replications than in original studies, with a 4-fold median difference for PCR experiments. This may either indicate that replicating labs had more difficulty in controlling

technical error or that variability was underestimated in the published results, due to elimination of outliers (Gress et al., 2018; Holman et al., 2016), incorrect labeling of error bars (Cumming et al., 2007; Vaux, 2004), or insufficient biological variation among experimental units (Lazic et al., 2018; Vaux et al., 2012). Filtering the sample by achieved power (considering the original relative effect and the observed variability) suggests that higher variability in replications biased replication rates based on statistical significance downwards; nevertheless, even when considering only experiments with adequate power, replication rates remained between 26 and 38%.

For experiments with high agreement among replications, it is reasonable to assume that the reasons for irreplicability are more likely to lie in the published results. Nevertheless, as we replicated individual comparisons between groups, lack of replication success does not apply to the whole article, or even to the whole experiment from which the comparison was drawn. Due to our choice of selecting the first comparison with a significant difference in an article, it was not infrequent to select an initial experiment or a low treatment dose, meaning that an unsuccessful replication does not necessarily call the main conclusions of a study into question.

More importantly still, low agreement among replications and/or a lower number of replications than originally planned limited our ability to judge many of the original experiments. Our results should thus be thought of as naturalistic estimates of real-life academic labs trying to replicate individual experiments from the published literature. That said, due to our crowdsourced, country-based design, the population of labs that performed replications was largely similar to the one that produced the original results. Thus, whether the original estimate or the replication one is a more accurate description of the phenomenon under study does not change the conclusion that Brazilian researchers have difficulties replicating findings from other research groups within the same community.

In fact, the observation that most labs faced difficulties in following preregistered protocols and controlling experimental variability provided an opportunity to glimpse at possible causes for replication failures. While all we can say about the original results is how closely they were replicated by our labs – which provides little information about the reasons for failed replications – our data collection and review process allowed an extensive audit of how closely replication protocols were followed by participant labs. This was complemented by a self-assessment process that allowed us to investigate relevant difficulties and opportunities for improvement.

Deviations from preregistered protocols were present in most replications. They occurred most commonly due to issues that were intrinsic to the experiment – such as difficulties with growing cells in the conditions reported in the original publication, or with performing experimental procedures as stated in the protocol. In many cases, labs reported that these issues could not have been anticipated without direct experience with the experimental model. Similar limitations were observed in the RP:CB, in which two thirds of preregistered protocols required modifications after experimental work had begun (Errington, Denis, et al., 2021).

This should be considered when interpreting replication rates, as difficulties in adequately implementing a method or experimental model are likely to decrease the probability of replication success. Thus, “one-shot” replication projects such as our own, in which extensive efforts are made to define the protocol in advance, are different from “best-shot” attempts at replication that allow for tinkering with the protocol over the course of multiple replication attempts – which are arguably closer to how experiments are typically replicated by academic labs (Guttinger, 2019). While the former are likely to underestimate replicability due to the difficulty in defining methods upfront, the latter can overestimate it due to the active effort in obtaining results that are close to the original.

The challenge of prespecifying methods also poses limits to the applicability of protocol preregistration in laboratory biology – or in any other area in which data collection depends on complex methods. Although we believe that preregistration of experimental methods is worthwhile, it may be more useful when performed after pilot experiments have been carried out, in order to reduce the need for deviations. Preregistering analysis methods poses fewer difficulties

when data is relatively simple, as in our experiments; nevertheless, more data-intensive methods can involve steps of data filtering or complex analytical decisions (e.g. Carp, 2012) that may be equally hard to specify in advance.

Our attempts to establish predefined inclusion/exclusion criteria to assess the validity of individual samples or experiments (Neves & Amaral, 2020) were also less successful than we had expected. Labs had difficulties in establishing these criteria in advance, frequently setting very high standards for markers, such as RNA purity, that turned out to be difficult to achieve and led them to break their own criteria. An independent validation process performed after the experiment turned out to be more feasible for assessing methodological validity and may be more compatible with the reality of academic labs. Importantly, this process was blind to the replication results, in order to prevent replication outcomes from contaminating judgments about methodological adequacy. Nevertheless, some methodological issues went undetected by this process and were identified only when results were actively discussed with participating labs.

Other reasons for protocol deviations included infrastructural problems, such as difficulties with the delivery of materials, cells or animals on time, limitations in access to equipment, and lack of personnel to perform replications. The COVID-19 pandemic proved to be a particularly complex challenge, as many labs and institutions were closed for extended periods, and even after normal activity was resumed, changes in replicating teams and lab priorities meant that many experiments ended up being performed in non-ideal conditions. Unsurprisingly, this was the most frequently reported difficulty by participants in our project evaluation survey.

All that said, there remains a fraction of protocol deviations that had no particular reason, in which labs simply made a different methodological choice from the preregistered one. This suggests that the idea of preregistration as a commitment to a particular methodology is still poorly understood by biomedical researchers. Most of them seem to consider protocols as general guidelines, assuming – perhaps correctly – that many steps will change after the experiment begins, and sometimes preferring vague descriptions over precise specifications to allow flexibility in the process. In some cases, this also led to a lack of clarity when developing protocols, leading to misunderstandings when experiments were performed by a different researcher.

These communication issues were particularly evident when determining the appropriate experimental unit in cell culture experiments. Although this was defined by the coordinating team in the protocols as different passages of a cultured cell line, or as primary cultures from independent pools of animals, many labs opted for other ways to define biological replicates. Not all adaptations were accepted, leading many replications to be excluded from our primary analysis. Moreover, merely understanding what experimental unit was used by a particular lab could be challenging, due to the lack of standardized terminology to define this among researchers (Lazic et al., 2018; Vaux et al., 2012). To a lesser extent, this also applied to other aspects of experimental design, such as positive and negative controls, blanks and randomization, which were described using distinct terminology by individual labs. This suggests that the adoption of consensus taxonomies that systematically and unambiguously describe these concepts could potentially improve reproducibility.

A final lesson learned from this project is that organizing large-scale initiatives in laboratory biology is hampered by the lack of experience of academic labs with working in coordinated fashion. While collaboration in biomedical research is frequent, the culture of repeating the same experiment across labs and standardizing procedures is confined to specific research fields (Brancato et al., 2024) or projects (Lyden et al., 2022). As a striking example of this limitation, a recent systematic review of the biomedical literature found only sixteen examples of multicenter animal studies (Hunniford et al., 2023). Consequently, data collection and management standards by labs are idiosyncratic, and harmonizing, reviewing and analyzing data from different sources can be challenging. Moreover, managing large-scale projects is a skill for which researchers receive little training, and being externally managed or supervised is also an unusual experience that can be met with resistance (Maysami et al., 2016). If large-scale, crowdsourced research is to flourish in lab biology, cultural changes are important to capacitate the academic

workforce for such collaborative endeavors (Amaral & Neves, 2021; Coles et al., 2022; Uhlmann et al., 2019). Possible solutions in this direction will be elaborated in a future article dedicated to these issues.

As a final reflection, our project leaves contributions beyond the results reported here. Our quantitative data describing replications are fully shared and open to scrutiny. Similarly, most of our textual data – including protocols, final notes, justifications for researcher predictions and other documents – are also open and remain to be analyzed in more depth. The collaborative nature of the project has constituted an important learning experience for the many researchers involved, including many early-career scholars, and the visibility we received (e.g. de Oliveira Andrade, 2019, 2025) has cast a spotlight on replicability and reproducibility in academic debates across the country. Finally, we leave spinoffs such as the Brazilian Reproducibility Initiative in Preclinical Systematic Review and Meta-Analysis (BRISA) (*http://en.reprodutibilidade.bio.br/brisa*), an ongoing collaboration to train biomedical researchers in systematic review methods, and the Brazilian Reproducibility Network (*http://en.reprodutibilidade.org*), a multidisciplinary organization to promote reproducible research practices across the country. We expect this legacy to help place reproducibility at the forefront of Brazil's national research agenda.

The Brazilian Reproducibility Initiative Author list

- Olavo Bohrer Amaral
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Clarissa França Dias Carneiro
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil Serrapilheira Institute, Rio de Janeiro, Brazil
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil Serrapilheira Institute, Rio de Janeiro, Brazil
- Kleber Neves
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil Serrapilheira Institute, Rio de Janeiro, Brazil
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil Serrapilheira Institute, Rio de Janeiro, Brazil
- Ana Paula Wasilewska Sampaio
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil National Center for Research in Energy and Materials, Campinas, Brazil
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil National Center for Research in Energy and Materials, Campinas, Brazil
- Bruna Valério Gomes
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil D'Or Institute for Research and Education - Rio de Janeiro, Brazil
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil D'Or Institute for Research and Education - Rio de Janeiro, Brazil
- Mariana Boechat de Abreu
  - Carlos Chagas Filho Biophysics Institute - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Pedro Batista Tan
  - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro,

Rio de Janeiro, Brazil

- Gabriel Paz Souza Mota
  - Faculty of Medicine - Federal University of Juiz de Fora, Juiz de Fora, Brazil
- Ricardo Netto Goulart
  - Faculty of Medicine - Federal University of Pelotas, Pelotas, Brazil
- Nathalia Raquel de Souza Fernandes
  - Federal University of Amazonas, Manaus, Brazil Polytechnic Institute of Bragança - Santa Apolónia Campus, Bragança, Portugal
  - Federal University of Amazonas, Manaus, Brazil Polytechnic Institute of Bragança - Santa Apolónia Campus, Bragança, Portugal
- Jimmy Hayden Linhares
  - Federal University of Amazonas, Manaus, Brazil
- Adriana Mércia Guaratini Ibelli
  - Embrapa Southeast Livestock - São Carlos, Brazil
- Adriano Defini Andricopulo
  - Laboratory of Medicinal and Computational Chemistry - São Carlos Institute of Physics, University of São Paulo, São Carlos, Brazil
- Adriano Sebollela
  - Department of Biochemistry and Immunology - Ribeirão Preto Medical School - University of São Paulo, Ribeirão Preto, Brazil
- Adryano Augustto Valladão de Carvalho
  - Research Group on Development of Pharmaceutical Products - Center for Biological and Health Sciences - Federal University of Western Bahia, Barreiras, Brazil
- Airton Pereira e Silva
  - Institute of Biology - Fluminense Federal University - Niterói, Brazil
- Alana Silva Oliveira Souza
  - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Alessandra Catarina Chagas de Lima
  - Multidisciplinary Research Center in Biology - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Alessandra Mara de Sousa
  - René Rachou Institute - Oswaldo Cruz Foundation, Belo Horizonte, Brazil
- Alexander Birbrair
  - University of Wisconsin-Madison, Madison, United States
- Alexandre Urban Borbely
  - Federal University of Alagoas - Maceió, Brazil
- Aline Maria Machado
  - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil
- Alinne do Carmo Costa
  - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Aliny Pereira de Vasconcelos

    - Department of Biotechnology - Federal University of Paraíba, João Pessoa, Brazil
- Alvaro Henrique Bernardo de Lima Silva
    - Neuropsychopharmacology Laboratory - Department of Pharmacology - Federal University of Paraná, Curitiba, Brazil
- Amanda de Souza
    - Hospital de Clínicas de Porto Alegre - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Ana Beatriz Rezende Paula
    - Cell Signaling Laboratory - Department of Biological Sciences - Federal University of Ouro Preto, Ouro Preto, Brazil
- Ana Carolina Caetano Nunes
    - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil
    - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil
- Ana Maria de Lauro Castrucci
    - Department of Physiology - Institute of Biosciences - University of São Paulo, São Paulo, Brazil
- Ana Paula Almeida Bastos
    - Embrapa Swine and Poultry - Concórdia, Brazil
- Ana Paula Farias Waltrick
    - Neuropsychopharmacology Laboratory - Department of Pharmacology - Federal University of Paraná, Curitiba, Brazil
- Ana Paula Herrmann
    - Department of Pharmacology - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Andre Gustavo Ferreira de Macedo
    - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Andrelson Wellington Rinaldi
    - State University of Maringá, Maringá, Brazil
- André Souza Mecawi
    - Laboratory of Molecular Neuroendocrinology - Department of Biophysics - Federal University of São Paulo, São Paulo, Brazil
- Andrés Delgado Canedo
    - CIPBiotec - Federal University of Pampa, São Gabriel, Brazil
- Anna Paula Marçal
    - Pharmacology Department - Institute of Biological Science - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Anna Paula Perin Vidigal
    - Center for Health Sciences - Federal University of Espírito Santo, Vitória, Brazil
- Antonio Martins Monteiro
    - Rio de Janeiro Cell Bank - Rio de Janeiro, Brazil
- Beatriz Gonçalves Silva Rocha

    - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Bianca Cruz Pachane
    - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil
    - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil
- Brenda da Silva Andrade
    - Laboratory of Molecular Pharmacology - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Bruna Carla Casali
    - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil Brazilian Center For Research in Energy and Materials, Campinas, Brazil
    - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil Brazilian Center For Research in Energy and Materials, Campinas, Brazil
- Bruno Popik
    - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Camila Pasquini de Souza
    - Department of Pharmacology - Federal University of Paraná, Curitiba, Brazil
- Camila da Silva Dos Santos
    - D'Or Institute for Research and Education - Rio de Janeiro, Brazil
- Camilla Mendes Gonçalves
    - Federal University of Alagoas - Maceió, Brazil
- Camilla Ribeiro Barroso do Nascimento
    - Rio de Janeiro Cell Bank - Rio de Janeiro, Brazil
- Carla Pires Veríssimo
    - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Carlos Eduardo Neves Girardi
    - Laboratory of Molecular Neuroendocrinology - Department of Biophysics - Federal University of São Paulo, São Paulo, Brazil
- Carmen Penido
    - Center for Technological Development in Health - Institute of Drug Technology - Oswaldo Cruz Foundation, Rio de Janeiro, Brazil
- Carolina Batista
    - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Carolina Panis
    - Tumor Biology Laboratory - State University of Western Paraná, Francisco Beltrão, Brazil
- Carolina Saibro-Girardi
    - Department of Biochemistry - Institute of Basic Health Sciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Carolina da Silva Gouveia Pedrosa
    - D'Or Institute for Research and Education - Rio de Janeiro, Brazil

- Caroline de Carvalho Picoli
  - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Cristiane Regina Guerino Furini
  - Laboratory of Cognition and Memory Neurobiology - Brain Institute - Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil
- Daniel Pens Gelain
  - Department of Biochemistry - Institute of Basic Health Sciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Daniel Sturza Lucas Caetano
  - Experimental Research Center - Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil
- Daniela Costa Silva
  - Rio de Janeiro Cell Bank - Rio de Janeiro, Brazil
- Daniele Cristina de Aguiar
  - Pharmacology Department - Institute of Biological Science - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Demetrius Antonio Machado De Araújo
  - Department of Biotechnology - Federal University of Paraíba, João Pessoa, Brazil
- Deoclécio Alves Chianca Júnior
  - Cardiovascular Physiology Laboratory - Department of Biological Sciences - Federal University of Ouro Preto, Ouro Preto, Brazil
- Dilza Balteiro Pereira de Campos
  - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Débora Aguirre Gonçalves
  - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Débora Santos da Silva
  - Oswaldo Cruz Institute - Oswaldo Cruz Foundation, Rio de Janeiro, Brazil
- Eduarda Godfried Nachtigall
  - Laboratory of Cognition and Memory Neurobiology - Brain Institute - Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil
- Eloiza Lopes Lira Tanabe
  - Federal University of Alagoas - Maceió, Brazil
- Erika Seki Kioshima
  - State University of Maringá, Maringá, Brazil
- Fabio Rodrigues Ferreira Seiva
  - Institute of Biosciences - São Paulo State University, Botucatu, Brazil
- Fabrício de Araujo Moreira
  - Pharmacology Department - Institute of Biological Science - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Felipe Saceanu Leser
  - Laboratory of Glial Cell Biology - Biomedical Sciences Institute - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil Paris Cardiovascular Research Center - French National Institute of Health and Medical Research - Paris, France
  - Laboratory of Glial Cell Biology - Biomedical Sciences Institute - Federal University of Rio

de Janeiro, Rio de Janeiro, Brazil Paris Cardiovascular Research Center - French National Institute of Health and Medical Research - Paris, France

- Felipe Vanz
  - Department of Pharmacology - Federal University of Santa Catarina, Florianópolis, Brazil
- Fernanda Cacilda dos Santos Silva
  - Cardiovascular Physiology Laboratory - Department of Biological Sciences - Federal University of Ouro Preto, Ouro Preto, Brazil
- Fernanda Magalhães Ferrão
  - Multidisciplinary Research Center in Biology - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Fernanda Nogueira Lotz
  - Psychobiology and Neurocomputation Laboratory - Department of Biophysics - Institute of Biosciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Flavia Fonseca Bloise
  - Institute of Biophysics Carlos Chagas Filho - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Flavia Regina Souza Lima
  - Laboratory of Glial Cell Biology - Biomedical Sciences Institute - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Flávio Alves Lara
  - Oswaldo Cruz Institute - Oswaldo Cruz Foundation, Rio de Janeiro, Brazil
- Franciana Aparecida Volpato
  - Embrapa Swine and Poultry - Concórdia, Brazil
- Francieli Moro Stefanello
  - Biomarkers Laboratory - Federal University of Pelotas, Pelotas, Brazil
- Francisca Nathalia de Luna Vitorino
  - Butantan Institute, São Paulo, Brazil
- Francisco Noé Fonseca
  - Embrapa Headquarters - Brasília, Distrito Federal, Brazil
- Fábio Jorge Moreira da Silva
  - Laboratory of Glial Cell Biology - Biomedical Sciences Institute - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Fábio de Almeida Mendes
  - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Gabriel Vasata Furtado
  - Hospital de Clínicas de Porto Alegre - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Gilda Angela Neves
  - Laboratory of Molecular Pharmacology - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Giovanna Zanetti
  - Department of Physiology - Institute of Biosciences - University of São Paulo, São Paulo,

Brazil

- Giulia Scarcella Cancelliero
  - Department of Biochemistry and Immunology - Ribeirão Preto Medical School - University of São Paulo, Ribeirão Preto, Brazil
- Graziele Fernanda Deriggi Pisani
  - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil
- Guilherme Curi Aiub Casagrande
  - Experimental Research Center - Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil
- Gustavo Roberto Villas-Boas
  - Research Group on Development of Pharmaceutical Products - Center for Biological and Health Sciences - Federal University of Western Bahia, Barreiras, Brazil
- Heitor Roque Oliveira Alvez da Cruz
  - Institute of Biology - Fluminense Federal University - Niterói, Brazil
- Helena Lobo Borges
  - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Heloisa Sobreiro Selistre De Araujo
  - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil
- Helvécio Cardoso Corrêa Póvoa
  - Nova Friburgo Health Institute - Fluminense Federal University, Nova Friburgo, Brazil
- Hilana dos Santos Sena Brunel
  - BioInnova Laboratory - Biomolecular Tests and Solutions, São Paulo, Brazil
- Hugo Bayer
  - Department of Pharmacology - Federal University of Santa Catarina, Florianópolis, Brazil
- Igor Luchini Baptista
  - Faculty of Applied Sciences - University of Campinas, Limeira, Brazil
- Isabel Werle
  - Department of Pharmacology - Federal University of Santa Catarina, Florianópolis, Brazil
- Isabela Alcântara Barretto Araújo Jardim
  - Cell Signaling Laboratory - Department of Biological Sciences - Federal University of Ouro Preto, Ouro Preto, Brazil
- Isabela Aparecida Divino
  - Faculty of Applied Sciences - University of Campinas, Limeira, Brazil
- Janaina Menezes Zanoveli
  - Neuropsychopharmacology Laboratory - Department of Pharmacology - Federal University of Paraná, Curitiba, Brazil
- Jane de Oliveira Peixoto
  - Embrapa Swine and Poultry - Concórdia, Brazil
- Jaqueline de Carvalho Rinaldi
  - State University of Maringá, Maringá, Brazil
- Jeane Bachi Ferreira

    - Federal University of Santa Catarina, Florianópolis, Brazil
- Jeferson Luis Franco
    - CIPBiotec - Federal University of Pampa, São Gabriel, Brazil
- Jeronimo Marteleto Nunes Rugani
    - René Rachou Institute - Oswaldo Cruz Foundation, Belo Horizonte, Brazil
- Jociane de Carvalho Myskiw
    - Psychobiology and Neurocomputation Laboratory - Department of Biophysics - Institute of Biosciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Jorge Alberto Quillfeldt
    - Psychobiology and Neurocomputation Laboratory - Department of Biophysics - Institute of Biosciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- José Marcos Janeiro Pereira da Costa
    - Laboratory of Glial Cell Biology - Biomedical Sciences Institute - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- João Victor Roza Cruz
    - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Julia Helena Oliveira de Barros
    - Institute of Biophysics Carlos Chagas Filho - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Julia Mello Barros
    - Multidisciplinary Research Center in Biology - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Julia Pinheiro Chagas da Cunha
    - Butantan Institute, São Paulo, Brazil
- Juliana Nunes Roson
    - Butantan Institute, São Paulo, Brazil
- Júlia Dummer Rodrigues de Freitas
    - Laboratory of Cognition and Memory Neurobiology - Brain Institute - Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, Brazil
- Júlia Grigorini Mori Ayub
    - Center for Health Sciences - Federal University of Espírito Santo, Vitória, Brazil
- Karen Steponavicius Cruz Borbely
    - Federal University of Alagoas - Maceió, Brazil
- Karina Barbosa de Queiroz
    - School of Nutrition - Federal University of Ouro Preto, Ouro Preto, Brazil
- Karina Dutra Asensi
    - Institute of Biophysics Carlos Chagas Filho - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Karoline dos Santos Rodrigues
    - Experimental Research Center - Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil
- Kellyne de Castro Vieira

    - Department of Pharmacology - Federal University of Paraná, Curitiba, Brazil
- Keyla Silva Nobre Pires
    - Federal University of Alagoas - Maceió, Brazil
- Kétlyn Talise Knak Guerra
    - Psychobiology and Neurocomputation Laboratory - Department of Biophysics - Institute of Biosciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Lays Adrianne Mendonça Trajano-Silva
    - Department of Cell and Molecular Biology and Pathogenic Bioagents - Ribeirão Preto Medical School - University of São Paulo, Ribeirão Preto, Brazil
- Leandro Jose Bertoglio
    - Department of Pharmacology - Federal University of Santa Catarina, Florianópolis, Brazil
- Leonardo Vinicius Monteiro de Assis
    - Department of Physiology - Institute of Biosciences - University of São Paulo, São Paulo, Brazil
- Leticia Menezes Vasconcelos
    - Department of Physiology - Institute of Biosciences - University of São Paulo, São Paulo, Brazil
- Leticia Miranda Santos Lery
    - Oswaldo Cruz Institute - Oswaldo Cruz Foundation, Rio de Janeiro, Brazil
- Letícia de Oliveira Marinho da Silva
    - Department of Physiology - Institute of Biosciences - University of São Paulo, São Paulo, Brazil
- Lorena de O Fernandes-Siqueira
    - Leopoldo de Meis Institute of Medical Biochemistry - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Lucas De Oliveira Alvares
    - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Lucas Gabriel Vieira
    - Cardiovascular Physiology Laboratory - Department of Biological Sciences - Federal University of Ouro Preto, Ouro Preto, Brazil
- Lucas Henrique de Melo Falcione
    - Department of Cell and Molecular Biology and Pathogenic Bioagents - Ribeirão Preto Medical School - University of São Paulo, Ribeirão Preto, Brazil
- Lucas Dos Santos da Silva
    - Department of Biochemistry - Institute of Basic Health Sciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Lucianne Fragel Madeira
    - Institute of Biology - Fluminense Federal University - Niterói, Brazil
- Luis Eduardo Duarte Nunes
    - Laboratory of Molecular Pharmacology - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Luiz Gustavo de Almeida Chuffa
    - Department of Structural and Functional Biology - Institute of Biosciences - São Paulo

State University, São Paulo, Brazil

- Luiza Marques Prates Behrens
  - Department of Biochemistry - Institute of Basic Health Sciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Marcos Vinicius Esteca
  - Faculty of Applied Sciences - University of Campinas, Limeira, Brazil
- Maria Luiza Saraiva-Pereira
  - Hospital de Clínicas de Porto Alegre - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Maria Nathália Moraes
  - Department of Physiology - Institute of Biosciences - University of São Paulo, São Paulo, Brazil
- Maria Sueli Soares Felipe
  - Catholic University of Brasília - Brasília, Brazil
- Mariana Saldanha Viegas Duarte
  - Clinical and Functional Genomics Laboratory - A.C. Camargo Cancer Center, São Paulo, Brazil
- Mariana Souza da Silveira
  - Institute of Biophysics Carlos Chagas Filho - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Marimelia A Porcionatto
  - Department of Biochemistry - Paulista School of Medicine - Federal University of São Paulo, São Paulo, Brazil
- Marisa Salvi
  - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Maryana Albino Clavero
  - Department of Pharmacology - Federal University of Paraná, Curitiba, Brazil
- Mateus Lobato Ferreira
  - Pharmacology Department - Institute of Biological Science - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Matheus Alves de Moura
  - Multidisciplinary Research Center in Biology - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Matheus Gallas-Lopes
  - Department of Pharmacology - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Mauro César Isoldi
  - Cell Signaling Laboratory - Department of Biological Sciences - Federal University of Ouro Preto, Ouro Preto, Brazil
- Mayara Terra Villela Vieira Mundim
  - Department of Biochemistry - Paulista School of Medicine - Federal University of São Paulo, São Paulo, Brazil
- Michael Andrades

    - Experimental Research Center - Hospital de Clínicas de Porto Alegre, Porto Alegre, Brazil
- Micheline Freire Donato
    - Department of Biotechnology - Federal University of Paraíba, João Pessoa, Brazil
- Milena Piaia Barboza
    - CIPBiotec - Federal University of Pampa, São Gabriel, Brazil
- Miliane Martins de Andrade Fagundes
    - School of Nutrition - Federal University of Ouro Preto, Ouro Preto, Brazil
- Miriane Avelino Silva
    - Department of Physiology - Institute of Biosciences - University of São Paulo, São Paulo, Brazil
- Morgana Duarte da Silva
    - Federal University of Santa Catarina, Florianópolis, Brazil
- Munira Muhammad Abdel Baqui
    - Department of Cell and Molecular Biology and Pathogenic Bioagents - Ribeirão Preto Medical School - University of São Paulo, Ribeirão Preto, Brazil
- Mychelle Pacheco de Souza
    - Nova Friburgo Health Institute - Fluminense Federal University, Nova Friburgo, Brazil
- Mônica Corrêa Ledur
    - Embrapa Swine and Poultry - Concórdia, Brazil
- Nathalia Stark Pedra
    - Neurochemistry, Inflammation, and Cancer Laboratory - Federal University of Pelotas, Pelotas, Brazil
- Natália Iorio Lopes Pontes Póvoa
    - Nova Friburgo Health Institute - Fluminense Federal University, Nova Friburgo, Brazil
- Natália Karla Bellini
    - Butantan Institute, São Paulo, Brazil
- Nauana Somensi
    - Department of Biochemistry - Institute of Basic Health Sciences - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Nicolly Maria Payva Nunes
    - CIPBiotec - Federal University of Pampa, São Gabriel, Brazil
- Nícia Pedreira Soares
    - Pharmacology Department - Institute of Biological Science - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Nívea Ferreira Silva
    - Rio de Janeiro Cell Bank - Rio de Janeiro, Brazil
- Pablinny Moreira Galdino de Carvalho
    - Research Group on Development of Pharmaceutical Products - Center for Biological and Health Sciences - Federal University of Western Bahia, Barreiras, Brazil
- Paola Alejandra Cappelletti
    - Rio de Janeiro Cell Bank - Rio de Janeiro, Brazil
- Patricia Pestana Garcez

    - Institute of Biomedical Sciences - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Patrícia Kellen Martins Oliveira Brito
    - Department of Cellular and Molecular Biology - Ribeirão Preto Medical School - University of São Paulo, Ribeirão Preto, Brazil
- Patrícia Rocha Martins
    - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Paulo Roberto Moreira Lauar
    - Federal University of Minas Gerais, Belo Horizonte, Brazil
- Priscila Nicolicht-Amorim
    - Department of Biochemistry - Paulista School of Medicine - Federal University of São Paulo, São Paulo, Brazil
- Rafael Chitolina
    - Department of Pharmacology - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
- Raiana Andrade Quintanilha Barbosa
    - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil National Institute of Cardiology, Rio de Janeiro, Brazil
    - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil National Institute of Cardiology, Rio de Janeiro, Brazil
- Raul Alexander Gonzales Cordova
    - Department of Cell and Molecular Biology and Pathogenic Bioagents - Ribeirão Preto Medical School - University of São Paulo, Ribeirão Preto, Brazil
- Regina Coeli Dos Santos Goldenberg
    - Institute of Biophysics Carlos Chagas Filho - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Rejane Giacomelli Tavares
    - Biomarkers Laboratory - Federal University of Pelotas, Pelotas, Brazil
- Renata Krogh
    - Laboratory of Medicinal and Computational Chemistry - São Carlos Institute of Physics, University of São Paulo, São Carlos, Brazil
- Renata de Freitas Saito
    - Center for Translational Research in Oncology - Cancer Institute of the State of São Paulo (ICESP) - Clinical Hospital of the University of São Paulo Medical School, São Paulo, Brazil Comprehensive Center for Precision Oncology, University of São Paulo, São Paulo, Brazil
    - Center for Translational Research in Oncology - Cancer Institute of the State of São Paulo (ICESP) - Clinical Hospital of the University of São Paulo Medical School, São Paulo, Brazil Comprehensive Center for Precision Oncology, University of São Paulo, São Paulo, Brazil
- Roberto Andreatini
    - Department of Pharmacology - Federal University of Paraná, Curitiba, Brazil
- Roberto Farina Almeida
    - Federal University of Pelotas - Pelotas, Brazil
- Rodrigo Cunha Alvim de Menezes
    - Cardiovascular Physiology Laboratory - Department of Biological Sciences - Federal

University of Ouro Preto, Ouro Preto, Brazil

- Rodrigo Rorato
  - Laboratory of Molecular Neuroendocrinology - Department of Biophysics - Federal University of São Paulo, São Paulo, Brazil
- Roger Chammas
  - Center for Translational Research in Oncology - Cancer Institute of the State of São Paulo (ICESP) - Clinical Hospital of the University of São Paulo Medical School, São Paulo, Brazil Comprehensive Center for Precision Oncology, University of São Paulo, São Paulo, Brazil
  - Center for Translational Research in Oncology - Cancer Institute of the State of São Paulo (ICESP) - Clinical Hospital of the University of São Paulo Medical School, São Paulo, Brazil Comprehensive Center for Precision Oncology, University of São Paulo, São Paulo, Brazil
- Rosangela Vieira de Andrade
  - Catholic University of Brasília - Brasília, Brazil
- Roselia Maria Spanevello
  - Neurochemistry, Inflammation, and Cancer Laboratory - Federal University of Pelotas, Pelotas, Brazil
- Rosiane Andrade Costa
  - Catholic University of Brasília - Brasília, Brazil
- Rubens Lima do Monte-Neto
  - René Rachou Institute - Oswaldo Cruz Foundation, Belo Horizonte, Brazil
- Sabrina Alves dos Reis
  - Oswaldo Cruz Institute - Oswaldo Cruz Foundation, Rio de Janeiro, Brazil
- Scheila Iria Kraus
  - Federal University of Santa Catarina, Florianópolis, Brazil
- Silvana Chedraoui Silva
  - Department of Biochemistry and Immunology - Ribeirão Preto Medical School - University of São Paulo, Ribeirão Preto, Brazil
- Simone Michelan-Duarte
  - Laboratory of Medicinal and Computational Chemistry - São Carlos Institute of Physics, University of São Paulo, São Carlos, Brazil
- Stevens Kastrup Rehen
  - D'Or Institute for Research and Education - Rio de Janeiro, Brazil Department of Genetics - Institute of Biology - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
  - D'Or Institute for Research and Education - Rio de Janeiro, Brazil Department of Genetics - Institute of Biology - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Sylvana Izaura Salyba Rendeiro de Noronha
  - Cardiovascular Physiology Laboratory - Department of Biological Sciences - Federal University of Ouro Preto, Ouro Preto, Brazil
- Tania Maria Ortiga-Carvalho
  - Institute of Biophysics Carlos Chagas Filho - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Tanira da Silveira Prieto
  - CIPBiotec - Federal University of Pampa, São Gabriel, Brazil
- Tatiane Renata Fagundes

- State University of Northern Paraná, Paraná, Brazil

- Thabatta Karollynne Estevam Nakamura
  - Department of Biochemistry - Paulista School of Medicine - Federal University of São Paulo, São Paulo, Brazil
- Thadeu Estevam Moreira Maramaldo Costa
  - Center for Technological Development in Health - Institute of Drug Technology - Oswaldo Cruz Foundation, Rio de Janeiro, Brazil
- Tharcísio Citrângulo Tortelli Junior
  - Center for Translational Research in Oncology - Cancer Institute of the State of São Paulo (ICESP) - Clinical Hospital of the University of São Paulo Medical School, São Paulo, Brazil Comprehensive Center for Precision Oncology, University of São Paulo, São Paulo, Brazil
  - Center for Translational Research in Oncology - Cancer Institute of the State of São Paulo (ICESP) - Clinical Hospital of the University of São Paulo Medical School, São Paulo, Brazil Comprehensive Center for Precision Oncology, University of São Paulo, São Paulo, Brazil
- Thays Barboza da Luz
  - CIPBiotec - Federal University of Pampa, São Gabriel, Brazil
- Thiago Aparecido da Silva
  - Department of Clinical Analysis - School of Pharmaceutical Sciences - São Paulo State University, Araraquara, Brazil
- Tiago Goss Dos Santos
  - Clinical and Functional Genomics Laboratory - A.C. Camargo Cancer Center, São Paulo, Brazil
- Vanessa Beijamini
  - Center for Health Sciences - Federal University of Espírito Santo, Vitória, Brazil
- Victória Regina de Siqueira Monteiro
  - Institute of Biophysics Carlos Chagas Filho - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Vinícius Alexandre Wippel
  - Federal University of Santa Catarina, Florianópolis, Brazil
- Viviane Medeiros Oliveira Valença
  - Institute of Biophysics Carlos Chagas Filho - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
- Wanessa Fernanda Altei
  - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil
  - Department of Physiological Sciences - Federal University of São Carlos, São Carlos, Brazil Molecular Oncology Research Center, Barretos Cancer Hospital, Barretos, Brazil

Acknowledgements

The coordinating team thanks Tim Errington for initial input on the project, Wolfgang Viechtbauer for input on the analysis, Felipe Dias for help in creating PCR protocol templates, Andreia Oliveira, Débora Moretti and Thaisa Santos for logistical support in obtaining supplies, and Anna Dreber, Domenico Viganola, Thomas Pfeiffer, Yiling Chen and Tiago Bortolini for participation in the prediction project. Participant labs thank Antonio Carlos Campos de Carvalho (LAB08), Daniel Caetano, Guilherme Casagrande (LAB09), Lanuza Faccioli (LAB13), Adriane Rosa, Daniela Varisco Müller, Leonardo Bastos, Sofia Becker, Thailana Stahlhofer-Buss (LAB16), Bruno Verçosa, Luísa

Ketzer, Silas Pessini (LAB19), Maria Cristina Barreira (LAB27), Ana Lívia Carvalho (LAB32), Andrea Da Poian (LAB42), Uliana Stotzer (LAB43), Marcelo Amaral (LAB50), Stephanie Megale, Bárbara Curvelano (LAB55), Leonardo Ferreira (LAB66), Denise Miranda (LAB75), Higor Santos (LAB80), Marcus Vinicius Bunscheit, Veronica Trujillo, Nathalia Ferreira (LAB82), Tarcisio Silva, Eylane Feitosa, Evellyn Tocha, Emili Santos, Hudson Campos (LAB91) for assistance with infrastructure, protocol design and/or data collection. This work was funded by a grant from Instituto Serrapilheira, with additional resources provided by FAPERJ (E-26/200.824/2021) and CNPq (310813/2021-2) to Olavo B. Amaral.

Additional information

Author contributions

A detailed list of author contributions to different experiments and project roles using the CRediT taxonomy is available at *https://osf.io/uxbtf*.

Ethical approval

Ethical approval for data collection involving human participants (researcher predictions and responses from original authors) was obtained from institutional review boards at the D'Or Institute for Research and Education (20112819.1.0000.5249) and Clementino Fraga Filho University Hospital (33532020.6.0000.5257). Ethical approval for replication of animal experiments was obtained individually by each participant laboratory, with institutions and protocol numbers available at *https://osf.io/4uzkj*.

Funding

Funder	Grant reference number	Author
Instituto Serrapilheira (Serrapilheira Institute)		Olavo B Amaral
Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ)	E-26/200.824/2021	Olavo B Amaral
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)	310813/2021-2	Olavo B Amaral

Additional files

Supplementary Figures and Tables

References

Amaral OB, Neves K (2021) Reproducibility: Expect less of the scientific paper. *Nature* **597**:329-331 https://doi.org/10.1038/d41586-021-02486-7 | PubMed

Amaral OB, Neves K, Wasilewska-Sampaio AP, Carneiro CF (2019) The Brazilian Reproducibility Initiative. *eLife* **8**:e41602 https://doi.org/10.7554/eLife.41602 | PubMed

Barata RB (2019) Necessary changes in the evaluation of graduate programs in Brazil. *Interface - Comunicação, Saúde. Educação* **23**:e180635 https://doi.org/10.1590/Interface.180635

Begley CG, Ellis LM (2012) Raise standards for preclinical cancer research. *Nature* **483**:531-533 https://doi.org/10.1038/483531a | PubMed

Brancato V, Esposito G, Coppola L, Cavaliere C, Mirabelli P, Scapicchio C, Borgheresi R, Neri E, Salvatore M, Aiello M (2024) Standardizing digital biobanks: Integrating imaging, genomic, and clinical data for precision medicine. *Journal of Translational Medicine* **22**:136 https://doi.org/10.1186/s12967-024-04891-8 | PubMed

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, *et al.* (2009) The MIQE Guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry* **55**:611-622 https://doi.org/10.1373/clinchem.2008.112797 |

PubMed

Carp J (2012) On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience* **6**:149 https://doi.org/10.3389/fnins.2012.00149 | PubMed

Coles NA, Hamlin JK, Sullivan LL, Parker TH, Altschul D (2022) Build up big-team science. *Nature* **601**:505-507 https://doi.org/10.1038/d41586-022-00150-2 | PubMed

Crabbe JC, Wahlsten D, Dudek BC (1999) Genetics of mouse behavior: Interactions with laboratory environment. *Science* **284**:1670-1672 https://doi.org/10.1126/science.284.5420.1670 | PubMed

Cumming G, Fidler F, Vaux DL (2007) Error bars in experimental biology. *Journal of Cell Biology* **177**:7-11 https://doi.org/10.1083/jcb.200611141 | PubMed

de Oliveira Andrade R (2019) Brazilian biomedical science faces reproducibility test. *Nature* **569**:318-319 https://doi.org/10.1038/d41586-019-01485-z | PubMed

de Oliveira Andrade R (2025) Huge reproducibility project fails to validate dozens of biomedical studies. *Nature* **641**:293-294 https://doi.org/10.1038/d41586-025-01266-x | PubMed

Elliott JT, Rösslein M, Song NW, Toman B, Ovaskainen AK.-, Maniratanachote R, Salit ML, Petersen EJ, Sequeira F, Romsos E, *et al.* (2017) Toward achieving harmonization in a nanocytotoxicity assay measurement through an interlaboratory comparison study. *ALTEX - Alternatives to Animal Experimentation* **34**:201-218 https://doi.org/10.14573/altex.1605021 | PubMed

Errington TM, Denis A, Perfito N, Iorns E, Nosek BA (2021) Challenges for assessing replicability in preclinical cancer biology. *eLife* **10**:e67995 https://doi.org/10.7554/eLife.67995 | PubMed

Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA (2021) Investigating the replicability of preclinical cancer biology. *eLife* **10**:e71601 https://doi.org/10.7554/eLife.71601 | PubMed

Fanelli D, Tan PB, Amaral OB, Neves K (2025) A metric of knowledge as information compression reflects reproducibility predictions for biomedical experiments. *Royal Society Open Science* **12**:241446 https://doi.org/10.1098/rsos.241446 | PubMed

Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T (2021) Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLoS One* **16**:e0248780 https://doi.org/10.1371/journal.pone.0248780 | PubMed

Gress TW, Denvir J, Shapiro JI (2018) Effect of removing outliers on statistical inference: Implications to interpretation of experimental data in medical research. *Marshall Journal of Medicine* **4**:9 https://doi.org/10.18590/mjm.2018.vol4.iss2.9 | PubMed

Guttinger S (2019) A new account of replication in the experimental life sciences. *Philosophy of Science* **86**:453-471 https://doi.org/10.1086/703555

Holman C, Piper SK, Grittner U, Diamantaras AA, Kimmelman J, Siegerink B, Dirnagl U (2016) Where have all the rodents gone? The effects of attrition in experimental research on cancer and stroke. *PLoS Biology* **14**:e1002331 https://doi.org/10.1371/journal.pbio.1002331 | PubMed

Hoogeveen S, Sarafoglou A, Wagenmakers E.-J (2020) Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science* **3**:267-285 https://doi.org/10.1177/2515245920919667

Hunniford VT, Grudniewicz A, Fergusson DA, Montroy J, Grigor E, Lansdell C, Lalu MM (2023) A systematic assessment of preclinical multilaboratory studies and a comparison to single laboratory studies. *eLife* **12**:e76300 https://doi.org/10.7554/eLife.76300 | PubMed

Ioannidis JPA (2008) Why Most Discovered True Associations Are Inflated. *Epidemiology* **19**:640-648 https://doi.org/10.1097/EDE.0b013e31818131e7 | PubMed

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG (2010) Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biology* **8**:e1000412 https://doi.org/10.1371/journal.pbio.1000412 | PubMed

Knapp G, Hartung J (2003) Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* **22**:2693-2710 https://doi.org/10.1002/sim.1482 | PubMed

Lazic SE, Clarke-Williams CJ, Munafò MR (2018) What exactly is 'N' in cell culture and animal experiments?. *PLoS Biology* **16**:e2005282 https://doi.org/10.1371/journal.pbio.2005282 | PubMed

Lemaitre J, Popelka D, Ribotta B, Westlake H, Chakrabarti S, Xiaoxue L, Hanson MA, Jiang H, Cara FD, Kurant E, *et al.* (2026) A retrospective analysis of 400 publications reveals patterns of irreproducibility across an entire life sciences research field. *eLife* **15** https://doi.org/10.7554/eLife.108403.1

Leta J (2012) Brazilian Growth in the Mainstream Science: The Role of Human Resources and National Journals – Journal of Scientometric Research. *Journal of Scientometric Research* **1**:44-52 https://doi.org/10.5530/jscires.2012.1.9

Lyden PD, Bosetti F, Diniz MA, Rogatko A, Koenig JI, Lamb J, Nagarkatti KA, Cabeen RP, Hess DC, Kamat PK, *et al.* (2022) The Stroke Preclinical Assessment Network: Rationale, design, feasibility, and stage 1 results. *Stroke* **53**:1802-1812 https://doi.org/10.1161/STROKEAHA.121.038047 | PubMed

Maysami S, Wong R, Pradillo JM, Denes A, Dhungana H, Malm T, Koistinaho J, Orset C, Rahman M, Rubio M, *et al.* (2016) A cross-laboratory preclinical study on the effectiveness of interleukin-1 receptor antagonist in stroke. *Journal of Cerebral Blood Flow & Metabolism* **36**:596-605 https://doi.org/10.1177/0271678X15606714 | PubMed

Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert Percie, Simonsohn N, Wagenmakers U, Ware E.-J, Ioannidis JPA (2017) A manifesto for reproducible science. *Nature Human Behaviour* **1**:1-9 https://doi.org/10.1038/s41562-016-0021 | PubMed

National Academies of Science (2019) Reproducibility and Replicability in Science. The National Academies Press. https://www.nationalacademies.org/our-work/reproducibility-and-replicability-in-science

Neves K, Amaral OB (2020) Addressing selective reporting of experiments through predefined exclusion criteria. *eLife* **9** https://doi.org/10.7554/eLife.56626 | PubMed

Neves K, Carneiro CF, Wasilewska-Sampaio AP, Abreu M, Valério-Gomes B, Tan PB, Amaral OB, Neves K, Carneiro CF, Wasilewska-Sampaio AP, *et al.* (2020) Two years into the Brazilian Reproducibility Initiative: Reflections on conducting a large-scale replication of Brazilian biomedical science. *Memórias Do Instituto Oswaldo Cruz* **115** https://doi.org/10.1590/0074-02760200328 | PubMed

Niepel M, Hafner M, Mills CE, Subramanian K, Williams EH, Chung M, Gaudio B, Barrette AM, Stern AD, Hu B, *et al.* (2019) A multi-center study on the reproducibility of drug-response assays in mammalian cell lines. *Cell Systems* **9**:35-48.e5. https://doi.org/10.1016/j.cels.2019.06.005 | PubMed

Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Dreber A, Fidler F, Hilgard J, Struhl MK, Nuijten MB, *et al.* (2022) Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology* **73** https://doi.org/10.1146/annurev-psych-020821-114157 | PubMed

NPQIP Collaborative group (2019) Did a change in Nature journals' editorial policy for life sciences research improve reporting?. *BMJ Open Science* **3**:e000035 https://doi.org/10.1136/bmjos-2017-000035 | PubMed

Prinz F, Schlange T, Asadullah K (2011) Believe it or not: How much can we rely on published data on potential drug targets?. *Nature Reviews Drug Discovery* **10**:712-712 https://doi.org/10.1038/nrd3439-c1 | PubMed

Quintans-Júnior LJ, Albuquerque GR, Oliveira SC, Silva RR (2020) Brazil's research budget: Endless setbacks. *EXCLI Journal* **19**:1322-1324 https://doi.org/10.17179/excli2020-2887 | PubMed

Scott S, Kranz JE, Cole J, Lincecum JM, Thompson K, Kelly N, Bostrom A, Theodoss J, Al-Nakhala BM, Vieira FG, *et al.* (2008) Design, power, and interpretation of studies in the standard murine model of ALS. *Amyotrophic Lateral Sclerosis* **9**:4-15 https://doi.org/10.1080/17482960701856300 | PubMed

Simonsohn U, Simmons JP, Nelson LD (2020) Specification curve analysis. *Nature Human Behaviour* **4**:1208-1214 https://doi.org/10.1038/s41562-020-0912-z | PubMed

Smaldino PE, McElreath R (2016) The natural selection of bad science. *Royal Society Open Science* **3**:160384 https://doi.org/10.1098/rsos.160384 | PubMed

Steward O, Popovich PG, Dietrich WD, Kleitman N (2012) Replication and reproducibility in spinal cord injury research. *Experimental Neurology* **233**:597-605 https://doi.org/10.1016/j.expneurol.2011.06.017 | PubMed

Uhlmann EL, Ebersole CR, Chartier CR, Errington TM, Kidwell MC, Lai CK, McCarthy RJ, Riegelman A, Silberzahn R, Nosek BA (2019) Scientific Utopia III: Crowdsourcing Science. *Perspectives on Psychological Science* **14**:711-733 https://doi.org/10.1177/1745691619850561 | PubMed

van Zwet E, Gelman A, Greenland S, Imbens G, Schwab S, Goodman SN (2023) A New Look at P Values for Randomized Clinical Trials. *NEJM Evidence* **3**:EVIDoa2300003 https://doi.org/10.1056/EVIDoa2300003 | PubMed

Vaux DL (2004) Error message. *Nature* **428**:799-799 https://doi.org/10.1038/428799c | PubMed

Vaux DL, Fidler F, Cumming G (2012) Replicates and repeats—What is the difference and is it significant?. *EMBO Reports* **13**:291-296 https://doi.org/10.1038/embor.2012.36 | PubMed

Viechtbauer W (2010) Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software* **36**:1-48 https://doi.org/10.18637/jss.v036.i03

Welch BL (1947) The generalization of "Student's" problem when several different population variances are involved. *Biometrika* **34**:28-35 https://doi.org/10.1093/biomet/34.1-2.28 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

This article describes a very ambitious metascience project aimed at testing the reproducibility of a corpus of publications conducted in Brazil. The strength of the approach lies in its systematic, multicenter replication design. The authors focus on three commonly used experimental paradigms in biology: the MTT assay, RT-PCR, and the elevated plus maze.

The effort is commendable and reveals a rather low rate of reproducibility, in line with findings from fields considered less reproducible in the life sciences, such as cancer biology.

Strengths:

The study is supported by a substantial dataset, incorporating multiple independent replication attempts and the use of stringent, well-defined protocols, which strengthens confidence in the overall conclusions.

Weaknesses:

(1) Being neither an expert in metascience nor in statistics, I cannot fully judge the methodological aspects of the article or its extensive supplementary material. I will therefore focus my comments on readability. I found the manuscript difficult to digest. The authors should improve readability if they wish to reach a broad audience of experimental biologists. In particular, they should simplify the description of protocols and highlight the key findings more clearly, using accessible language. See specific points below

(2) The article appears to oscillate between:

i) a description of the approach and the inherent challenges of such a multicenter replication program.

ii) an estimation of reproducibility.

These could potentially form two separate articles: one aimed at a broad audience emphasizing key results, and another focused on methodological aspects for a more specific metascience audience. The Results section currently contains redundancies and is difficult to follow for non-experts in statistics. I also find it challenging to extract the main findings.

A possible improvement would be to include an initial section clearly describing the protocol (replication of a single experiment, across several labs, for three types of assays), followed by a concise presentation of the main results regarding reproducibility in Brazilian science with subsections. Methodological details could be moved either to a Supplementary Information or to a more specific article, while being summarized in the Discussion.

(3) This study evaluates the reproducibility of a single experiment from each article, taken out of its broader context. While this provides an estimate of reproducibility, it does not directly contribute to resolving uncertainties within a specific field. This may represent a limitation compared to other reproducibility projects that attempt to replicate multiple key claims within a given study (e.g., in cancer biology or Drosophila immunity). I found that a weakness is that it does play a role in cleaning a field of wrong statements.

(4) The observation that external observers can predict which experiments are likely to be reproducible is interesting and should be more clearly emphasized.

(5) The manuscript frequently refers to future publications. It would be helpful to clarify what is included in the present article versus what is deferred to subsequent papers

https://doi.org/10.7554/eLife.111001.1.sa3

Reviewer #2 (Public review):

Summary:

This is an important contribution to science, not only because large-scale replication studies remain rare despite their value, but also because this one focuses on research that was under represented in previous large-scale efforts. The findings reveal concerningly low replicability in this field, pointing to a problem that warrants immediate attention. Particularly noteworthy is the study's sampling strategy: by randomly selecting experiments from a wide range of publications based on methods, rather than filtering by research area, importance, or citation counts, the authors have produced results that are potentially more representative of the broader literature than those of previous large-scale replication projects in this and other fields. Overall, this is a fantastic contribution that I will be recommending and using in all my open science talks, and from which I have learned a great deal. Congratulations to the team!

Strengths:

A study of this scale inevitably requires an enormous amount of work and methodological care, and this one is clearly both robust and thoughtfully designed. I want to particularly acknowledge the considerable efforts the authors have made to ensure the robustness of their findings. The use of multiple approaches to estimate replicability, combined with a substantial battery of sensitivity analyses, including a multiverse approach on top of everything else, clearly reflects the authors' genuine commitment to understanding their results and the limits of their conclusions. The transparency and sharing of all protocols, materials, and challenges and limitations encountered is also outstanding.

Weaknesses:

There were several instances during my reading of the methodology where I felt the authors relied too heavily on the external supplementary materials, at the expense of basic detail in

the main manuscript. I appreciate how overwhelming it can feel to integrate more into an already substantial paper, but without some minimum integration, the reading experience and overall comprehension are too often compromised, at times posing more questions than answers. And it is unrealistic to expect most readers to engage with the extensive supplementary materials provided. Please see the comments below for specific suggestions.

Additionally, I found the discussion rather underdeveloped. There is relatively little engagement with the broader literature, not only with replicability studies from other fields, but more generally with relevant meta-research work on publication bias, blinding, risk of bias, citation practices, etc. Some of the most novel and interesting findings in the paper also receive less attention than they deserve, and the discussion at times reads as a repetition of the results section rather than a critical engagement with them. I would encourage the authors to engage more deeply here, as the study clearly has much more to say. Doing so would further highlight why this study is important for the answers it provides and the questions it can spur. Again, please see the comments below for specific suggestions.

Specific suggestions:

Page 1, abstract: "while t values for replications were positively correlated with researcher predictions about replicability, and negatively correlated with the rate of publications by the original article's last author" - I need to address the question: why t values and not effect sizes, p values, or something else? Update after reading the study: although the authors used others, they seem to place more emphasis on t values, which is not well explained. Without a clear explanation, it just left me wonder why, given that effect sizes would, in principle, be more information.

Page 2, paragraph 2: "reproducibility (defined here as reaching the same results when analyzing a set of data)" - In my opinion, this definition is vague enough that it encompasses not only reproducibility (same data, same methods) but also robustness (same data, different methods), and I would therefore recommend providing a more precise definition. The same applies to replicability (different data, same methods), since the definition used does not highlight the importance of using the same methods, and thus also encompasses generalisability (different data, different methods). Explicitly clarifying these distinctions is particularly important as the field grows and the terms become increasingly mixed up and confusing.

Page 2, paragraph 3: "All of these issues raise concerns about the replicability of published results - something that has not been evaluated systematically in the country" - I would suggest providing more information about why those factors may lead to expected lower replicability, ideally with a couple of sentences supported by references. As it stands, less experienced readers may not follow the argumentation and may consider it speculative.

Page 3, paragraph 2: "We then opened a public call for Brazilian labs that could replicate experiments using these methods and models, advertised by email, social media and lectures in conferences and institutions, to which 73 labs initially responded" - Since recruiting is an important component of this study, I would recommend providing additional details so the reader can better assess how comprehensive and unbiased the recruitment process was. AND Page 5, paragraph 2: Please provide more information about this open call: how was it advertised, where, and when? This is needed so that the reader can assess its comprehensiveness and potential biases. Even the link provided is not specific enough to understand the process, as it only states: "Calls were open to participants > 18 years old with current or previous experience in experimental research in any field and were advertised via e-mails, lectures and social media."

Page 3, paragraph 2: "Based on the expertise of respondents and a feasibility analysis by the coordinating team, we selected 3 outcome assessment methods for replication" - Since this

choice determined what was ultimately studied and who could participate, I would like to see more information to understand it: was it based on the most common expertise among respondents? How was feasibility defined and estimated?

Page 3, paragraph 3: How was the manual screening performed? Was it done by one or more people? Was there double-screening to ensure reliability of the screening protocol? Did the authors use a specific decision tree or tool? How were conflicts between observers resolved? Were any other validation steps taken to ensure reliability? The same comments apply to the data extraction (who, how many, validation, protocol, etc.).

Page 3, paragraph 3: As a non-expert, I would need more context about the expected average cost of experiments in this field; otherwise, I cannot assess how representative this sample is or whether potential biases may exist (e.g., cheaper experiments perhaps being expected to be less replicable than more expensive ones). Could expected costs also have affected the reduction in geographical coverage eventually observed in this study (Figure S3)?

Page 6, paragraph 2: "(on a scale of 1 to 5)" - Could you clarify whether 1 means no deviations and 5 means everything deviated? Is that how it was phrased to participants? Was there a threshold used by the coordinating team to decide how many deviations were acceptable? (I would briefly clarify all scales mentioned below to allow easier interpretation throughout.)

Page 6, paragraph 4: How were long-text answers (e.g., justifications) reviewed? Was this done manually by one or more members of the coordinating team, or using any text interpretation tool? What steps were taken to ensure the interpretation of these answers was as objective as possible?

Page 8, paragraph 1: "If issues were found, the lab and coordinating team reviewed them via email until the sources of errors were identified and corrected (see *https://osf.io/58vsx* for details)." - Could you please provide information about how often these disagreements arose and briefly explain their causes? I am struggling to understand why these discrepancies occurred and how frequently. Without more detail, the error rate presented in the next paragraph is a little concerning.

Page 8, paragraph 4: Please provide the version of any package or software used throughout, and make sure to cite R appropriately (R Core Team XXX). In addition, did the authors calculate the log ratio of means (ROM/lnRR) using escalc()? If so, please report this. If not, I would recommend doing so, as escalc() implements recommended small-sample adjustments that produce slightly different values compared to a simple manual calculation of log(mean1/mean2).

Page 10, paragraph 1: "Coefficients of variation from the original study were compared to the mean coefficient of variation of its replications using Wilcoxon's signed rank test" - I wonder how these CVs were calculated - whether simply as SD/mean or using escalc() from the R package metafor, which includes a correction for small-sample size. This may affect the fairness of the comparison, particularly since CVs from original studies are expected to be slightly overestimated given their smaller sample sizes relative to the replications. I also have concerns about using the mean CV of all replications and comparing it to a single CV value, as this ignores the uncertainty around that mean. An additional check could involve calculating the log coefficient of variation ratio (lnCVR; Nakagawa et al. 2015, Methods in Ecology and Evolution; implemented in escalc()) between the original CV and each replication CV, and running a random-effects (or multilevel) meta-analysis that accounts for shared-control non-independence. I believe this would provide a more robust approach, as it does not ignore the uncertainty around the mean CV of the replications - uncertainty that, if neglected, is expected to increase the likelihood of false positive findings. This concern would also apply to the subsequent analysis on absolute means.

Page 10, paragraph 2: The change in geographical distribution shown in Figure S3 appears rather striking, with western states disappearing step by step. Should the reader be concerned about the eventual geographical representability of the sample?

Page 15, Figure 3A: I wonder whether adding 95% CIs calculated from the sampling variance of each ratio would improve interpretation and help readers appreciate the real differences between the dots (i.e., means) - along the lines of a forest plot.

Page 17, section "Predictors of replication success": It is unclear to me how the decision was made about which results from Figure 4 to present in the text. Intuitively, given that correlations were calculated for both t values and lnRR (and other metrics), I would have expected that whenever a result is highlighted in the text, the authors also report how it changes depending on the metric used - for example, the interesting result regarding the 5-year number of publications, whose correlation is notably lower when using lnRR (−0.31 vs. −0.18). Presenting this nuance in the text would reduce the risk of inadvertently giving the impression of cherry-picking.

Page 23, paragraph 1: (this comment should have come during the first % reported, but only in the discussion I realized how important this would be for comparing estimates) I wonder whether the authors should calculate 95% confidence intervals for all their percentages (and those of Errington et al.) using the Wilson method via the function binom.confint() in R, which handles extreme proportions (0% or 100%) more gracefully. This would ensure that uncertainty around these percentages is not neglected and would aid interpretation when comparisons are made. In addition, in the next sentence, the authors are comparing correlation coefficients, at least verbally, these could in principle be transformed into Pearson's r and assigned 95% confidence intervals following meta-analytic workflows, which would better allow us to assess whether these correlations are meaningfully larger or smaller, and help avoid potentially misleading arguments.

Page 24, paragraph 2: The following result is really interesting and I would love for the authors to expand on it a little. There must be other meta-research studies that, despite not studying replicability directly, have explored a similar predictor: "Other features of the original article were generally uncorrelated with replication outcome, although large rates of publications by the last author were associated with lower replicability, suggesting that incentivizing publication volume may be counterproductive for the reliability of results."

Page 25, paragraph 1: I believe the authors could explore if there is evidence for "incorrect labeling of error bars (Cumming et al., 2007; Vaux, 2004)" by plotting log(SD) vs log(mean) across all original studies, and exploring if large outliers (i.e., points largely deviating from the positive regression) exist. That should provide some insights into whether some values reported as SD in the original studies were indeed SE, which I am assuming is what the authors of the study are referring to when they say "incorrect labelling of error bars" here.

Code: I could not engage with the data and code, but I would like to highlight that the organisation and clarity of the GitHub repository is of high quality.

https://doi.org/10.7554/eLife.111001.1.sa2

Reviewer #3 (Public review):

Summary:

The authors conducted a large-scale replication effort of lab-based biomedical experiments with an emphasis on the country of origin and who conducted the replication experiments. The authors aimed to understand this context in both the outcomes produced, but also in the approach. Finally, the authors aimed to conduct multi-lab replications to provide richer data

from the replications. Overall, the authors find replication rates that are like other large-scale replication efforts in the biomedical space. The authors provide rich detail into the three experimental techniques that were the focus of this effort, potential moderators of replication success, and challenges in conducting replications and coordinating a large-scale crowd-sourced effort.

Strengths:

The paper is outstanding in being transparent and calibrated in how the results are presented. While the authors were challenged by mundane aspects (e.g., difficulty with logistics), unexpected aspects (e.g., COVID pandemic), and very insightful aspects unique to conducting replications (e.g., experimental issues). The authors also provide variation in how they present the results, including confirmatory, multiverse, and exploratory analysis. A unique strength for this study is the rich in-depth insights about the process and interpretation of conducting replications, including predicting replication success in the lab-based biomedical space.

Weaknesses:

The study has weaknesses that the authors acknowledge in their discussion, such as lower number of replications than originally planned that limited the intended effort to compare multiple experiments with multiple attempts against a single original experiment. Another weakness is the limited discussion connecting these findings to the Brazilian research ecosystem.

https://doi.org/10.7554/eLife.111001.1.sa1

Author response:

> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *This article describes a very ambitious metascience project aimed at testing the reproducibility of a corpus of publications conducted in Brazil. The strength of the approach lies in its systematic, multicenter replication design. The authors focus on three commonly used experimental paradigms in biology: the MTT assay, RT-PCR, and the elevated plus maze.*
>
> *The effort is commendable and reveals a rather low rate of reproducibility, in line with findings from fields considered less reproducible in the life sciences, such as cancer biology.*
>
> *Strengths:*
>
> *The study is supported by a substantial dataset, incorporating multiple independent replication attempts and the use of stringent, well-defined protocols, which strengthens confidence in the overall conclusions.*

We thank the reviewer for the comments.

> *Weaknesses:*
>
> *(1) Being neither an expert in metascience nor in statistics, I cannot fully judge the methodological aspects of the article or its extensive supplementary material. I will therefore focus my comments on readability. I found the manuscript difficult to digest. The authors should improve readability if they wish to reach a broad audience of experimental biologists. In particular, they should simplify the description of protocols*

> *and highlight the key findings more clearly, using accessible language. See specific points below*

We can try to simplify the description of protocols at specific points for example, by providing an overarching description of the study design in the beginning of the Methods, rather than citing our previous eLife paper (Amaral et al., 2019), as suggested below. The methods are indeed quite extensive, but the this may be inevitable in a large-scale project such as this and we note that Reviewer #2 thought that part of the supplementary material should be incorporated back in the main text, which is a suggestion in the opposite direction. It may thus be hard to strike a balance between readability and comprehensibility that can address both reviewers' opinions.

> *(2) The article appears to oscillate between:*
>
> *(i) a description of the approach and the inherent challenges of such a multicenter replication program*
>
> *(ii) an estimation of reproducibility.*
>
> *These could potentially form two separate articles: one aimed at a broad audience emphasizing key results, and another focused on methodological aspects for a more specific metascience audience. The Results section currently contains redundancies and is difficult to follow for non-experts in statistics. I also find it challenging to extract the main findings.*

There is a bit of redundancy between tables and text, but this was intentional to make both of them self-explanatory. We also think stating the results in the text can allow us to make each of the replication criteria clearer, a concern that was also mentioned by the reviewer.

As for requiring particular expertise in statistics for understanding, we mostly disagree. The main results (Tables 1 and 2, Figure 2) are expressed as percentages, and the only statistical concepts needed for interpreting these results are understanding prediction and confidence intervals. For this, we could provide a bit more guidance on their interpretation in the Methods section. Beyond that, most of the secondary results (e.g. Figure 3 and Figure 4) involve linear correlations, which is about as simple as statistical analysis gets.

Of the results presented in the main manuscript, only Table 3 contains anything beyond percentages and correlations. We do agree that the meaning of each ratio in this table could be more clearly described, but there are essentially no expert-level statistics involved in their calculations.

Other than that, the main statistical issues are the ideal way to aggregate the results from different replications for which we use different strategies for robustness purposes. However, all of these results are already in the supplementary material, so we don't feel they interfere to much with the readability of the main manuscript.

> *A possible improvement would be to include an initial section clearly describing the protocol (replication of a single experiment, across several labs, for three types of assays), followed by a concise presentation of the main results regarding reproducibility in Brazilian science with subsections.*

This is indeed a good idea, and we plan to include an initial overarching description of the project in the Methods section of the revised manuscript.

> *Methodological details could be moved either to a Supplementary Information or to a more specific article, while being summarized in the Discussion.*

Again, this is the opposite of what was suggested by Reviewer #2, so we would rather keep the Methods section more or less at its current level of detail.

> *(3) This study evaluates the reproducibility of a single experiment from each article, taken out of its broader context. While this provides an estimate of reproducibility, it does not directly contribute to resolving uncertainties within a specific field. This may represent a limitation compared to other reproducibility projects that attempt to replicate multiple key claims within a given study (e.g., in cancer biology or Drosophila immunity). I found that a weakness is that it does play a role in cleaning a field of wrong statements.*

The reviewer is correct in his interpretation. Evaluating the main findings of articles or cleaning a field of wrong statements was never a goal of our study (and we were clear about this from the start). Our aim with the project was metascientific (i.e. evaluate the reproducibility of biomedical experiments with a set of common methods) rather than driven by a particular interest in the findings themselves. This is reflected by our choice of selecting experiments from a random sample of articles from multiple fields, rather than filtering by area of interest or importance. It also underlies our choice to evaluate experiments rather than claims, as this was more statistically tractable and potentially more objective as a meta-research goal.

To be clear, we don't feel this approach is inherently better or worse than evaluating claims in the literature, as in the *Drosophila* immunity article case (i.e. Westlake et al., 2026), which is also an important goal. They are merely approaches that answer different questions. Ultimately, we probably made our choice based on (a) our expertise/interest in meta-research rather than in the fields the replications stemmed from and (b) an attempt to engage Brazilian researchers in the project in a way that was non-confrontational and minimized backlash from their peers. We feel this was valuable for many of the lessons learned, although it also meant learning less about the research findings in question.

Even though this was not a goal of the study, there is some knowledge obtained about the findings that is indeed largely absent from the current manuscript. We do not feel the current format allows for much discussion of 45 different findings, but we do have plans to address these in future articles (as outlined in our response to point 5). In the meantime, qualitative descriptions of each experiment can be found at *https://osf.io/w5z9a*. This is already mentioned in the Methods but could be reiterated in the results as well.

> *(4) The observation that external observers can predict which experiments are likely to be reproducible is interesting and should be more clearly emphasized.*

We did not go too deep into that finding because we are publishing a separate article focused on the prediction project, which should look into factors that correlate with prediction accuracy, both at the level of predictors (e.g. research field, career level) and of individual predictions (e.g. information taken into account for each answer). We also feel that, given the multiplicity of predictors in the prediction analyses, these findings are a bit tentative, as the strongest predictors may be subject to effect size inflation from the "winner's curse" effect (as outlined by Reviewer #2). We can try to emphasize it a little more in the discussion (although it already merits a whole paragraph on pages 23-24), but we feel we would be able to discuss it more critically in a follow-up article.

> *(5) The manuscript frequently refers to future publications. It would be helpful to clarify what is included in the present article versus what is deferred to subsequent papers.*

Indeed, some of our results did not fit this overarching analysis and were left for future publications. One of them is already available as a preprint, while the others are currently in preparation. Specifically, other results from the project should be spread about across five different articles.

(a) A narrative article focused on challenges and lessons learned with the project, already published as a preprint at *https://osf.io/preprints/metaarxiv/8y3tg_v1* (Amaral et al., 2026).

(b) An article analyzing the prediction survey and markets results in detail (following the pre-analysis plan detailed in *https://osf.io/6av7k/files/pjhgd* and adding some exploratory analyses on prediction rationales).

(c) Three articles describing the results of specific experiments with each experimental method (MTT, PCR, elevated plus maze) along with a discussion of aspects inherent to the method that seem to influence reproducibility.

We can add this information more explicitly to the Methods section, including the links to the papers that have already been published at the time the manuscript is revised.

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *This is an important contribution to science, not only because large-scale replication studies remain rare despite their value, but also because this one focuses on research that was underrepresented in previous large-scale efforts. The findings reveal concerningly low replicability in this field, pointing to a problem that warrants immediate attention. Particularly noteworthy is the study's sampling strategy: by randomly selecting experiments from a wide range of publications based on methods, rather than filtering by research area, importance, or citation counts, the authors have produced results that are potentially more representative of the broader literature than those of previous large-scale replication projects in this and other fields. Overall, this is a fantastic contribution that I will be recommending and using in all my open science talks, and from which I have learned a great deal. Congratulations to the team!*

Thanks!

> *Strengths:*
>
> *A study of this scale inevitably requires an enormous amount of work and methodological care, and this one is clearly both robust and thoughtfully designed. I want to particularly acknowledge the considerable efforts the authors have made to ensure the robustness of their findings. The use of multiple approaches to estimate replicability, combined with a substantial battery of sensitivity analyses, including a multiverse approach on top of everything else, clearly reflects the authors' genuine commitment to understanding their results and the limits of their conclusions. The transparency and sharing of all protocols, materials, and challenges and limitations encountered is also outstanding.*

We once more thank the reviewer for the compliments.

> *Weaknesses:*
>
> *There were several instances during my reading of the methodology where I felt the authors relied too heavily on the external supplementary materials, at the expense of basic detail in the main manuscript. I appreciate how overwhelming it can feel to integrate more into an already substantial paper, but without some minimum integration, the reading experience and overall comprehension are too often compromised, at times posing more questions than answers. And it is unrealistic to expect most readers to engage with the extensive supplementary materials provided. Please see the comments below for specific suggestions.*

We do acknowledge that the article currently includes a lot of supplementary material. This includes both supplementary figures/tables relating to the paper and many supplementary methods files (mostly hosted at the Open Science Framework). However, we also note that this is already a rather long paper as it stands and that Reviewer #1 has made the opposite suggestion of simplifying it. Thus, it may be hard to strike a balance that will suit all preferences, and we feel that maybe our attempt has landed somewhere in the middle of both reviewers' ideal versions of the paper.

> *Additionally, I found the discussion rather underdeveloped. There is relatively little engagement with the broader literature, not only with replicability studies from other fields, but more generally with relevant meta-research work on publication bias, blinding, risk of bias, citation practices, etc. Some of the most novel and interesting findings in the paper also receive less attention than they deserve, and the discussion at times reads as a repetition of the results section rather than a critical engagement with them. I would encourage the authors to engage more deeply here, as the study clearly has much more to say. Doing so would further highlight why this study is important for the answers it provides and the questions it can spur. Again, please see the comments below for specific suggestions.*

We can try to engage with some of the above-mentioned literature in more depth in particular replication studies from other fields (some of which have appeared after our preprint (e.g. Tyner et al., 2026) and with the risk of bias and transparency literature (e.g. Serghiou et al., 2021). That said, we note once more that the article (and the Discussion section) are already quite long, and that analyzing each of these articles in depth is likely to be unfeasible.

> *Specific suggestions:*
>
> *Page 1, abstract: "while t values for replications were positively correlated with researcher predictions about replicability, and negatively correlated with the rate of publications by the original article's last author" - I need to address the question: why t values and not effect sizes, p values, or something else? Update after reading the study: although the authors used others, they seem to place more emphasis on t values, which is not well explained. Without a clear explanation, it just left me wonder why, given that effect sizes would, in principle, be more information.*

Our original plan was to use p values as a predictor (see protocol at *https://osf.io/9rnuj*), but we later realized this was inadequate as it did not account for effect direction (i.e. significant effects in the opposite direction as the original may yield low p values, but this should not count as replication success). We thus switched to t values to be able to assign positive and negative signs depending on effect size direction. We note that, as we are using non-parametric Spearman coefficients (in which the module of t correlates negatively with the p value), the two approaches are effectively equivalent when original and replication effects have the same direction. This change was accounted for and justified in our list of protocol deviations at *https://osf.io/9hj7t*.

Effect size (in relative terms) is already being used in the second predictor in the analysis (i.e. effect size decrease), as our idea was to use one significance-based predictor and one effect size-based predictor, to match what was done for the replication rates). We feel that using relative effects (e.g. response ratios) by themselves may not be as adequate, as for experimental methods with large coefficients of variation and/or low sample sizes (especially PCR ones), one can find large relative effects that are nevertheless far from statistical significance. This also makes relative effects not very commensurable between methods.

We do believe there is a fair argument, however, to use standardized effect sizes as an alternative to t values (i.e. difference measured in standard errors of the mean) to measure

significance/evidence strength. As some replications ended up underpowered, low t values may sometimes be due to insufficient statistical power/low sample size rather than replication failures. Using standardized effect sizes is not devoid of pitfalls (e.g. they can be quite variable when sample size is low), but it is worth doing as a robustness analysis.

That said, there are a few statistical issues to be decided on how to calculate this (e.g. whether studies should be meta-analyzed using standardized mean differences rather than relative ones for this purpose, or whether an analog of the standardized effect size should be calculated for the log ratio of means). We would have to look more carefully into the multiple possibilities to decide on the best approach (and we do accept suggestions!).

In the meantime, we note that running the prediction analysis using only experiments with ≥80% power yields a slightly higher correlation of t scores with researcher predictions (ρ = 0.49, p = 0.005), so we do not think that these underpowered experiments affect the trend too much. If anything, they could be masking a higher correlation between researcher predictions and replicability.

> *Page 2, paragraph 2: "reproducibility (defined here as reaching the same results when analyzing a set of data)" - In my opinion, this definition is vague enough that it encompasses not only reproducibility (same data, same methods) but also robustness (same data, different methods), and I would therefore recommend providing a more precise definition. The same applies to replicability (different data, same methods), since the definition used does not highlight the importance of using the same methods, and thus also encompasses generalisability (different data, different methods). Explicitly clarifying these distinctions is particularly important as the field grows and the terms become increasingly mixed up and confusing.*

We agree that we should make the description more precise (e.g. “reaching the same results when analyzing a set of data in the same way” for reproducibility and “finding similar results with new data collected under similar conditions” for replicability). We will update these definitions in the revised manuscript.

> *Page 2, paragraph 3: "All of these issues raise concerns about the replicability of published results - something that has not been evaluated systematically in the country" - I would suggest providing more information about why those factors may lead to expected lower replicability, ideally with a couple of sentences supported by references. As it stands, less experienced readers may not follow the argumentation and may consider it speculative.*

We would argue that the reader would be correct in this case: the argument is a bit speculative. It does go in the direction of what is generally accepted within the field (i.e. that publication pressure can lead to lower reproducibility for a range of factors), but we’re not sure this connection has been demonstrated empirically, except for indirect evidence (such as the lower reproducibility in papers stemming from top institutions and “trophy journals” in, the higher frequency of positive results in US states with more researchers in Fanelli, 2010, or the higher number of problematic images for highly productive researchers in some countries in Fanelli et al., 2022. We could cite this evidence in the introduction and make the speculated connection more explicit, perhaps adding modeling work as well (e.g. Ioannidis, 2005; Smaldino & McElreath, 2016) to explain why this could be the case. But essentially, our opinion is that the connection remains a speculation.

> *Page 3, paragraph 2: "We then opened a public call for Brazilian labs that could replicate experiments using these methods and models, advertised by email, social media and lectures in conferences and institutions, to which 73 labs initially responded" - Since recruiting is an important component of this study, I would recommend providing additional details so the reader can better assess how comprehensive and unbiased the*

> *recruitment process was. AND Page 5, paragraph 2: Please provide more information about this open call: how was it advertised, where, and when? This is needed so that the reader can assess its comprehensiveness and potential biases. Even the link provided is not specific enough to understand the process, as it only states: "Calls were open to participants > 18 years old with current or previous experience in experimental research in any field and were advertised via e-mails, lectures and social media."*

We can offer a more detailed description of the recruitment process (e.g. number and distribution of lectures, social media strategy used, etc.), although we would rather do this in a supplementary document so as not to make the Methods section even lengthier. We note, however, that we never aimed to recruit a "representative sample" of labs from the country: we were busy enough trying to get enough labs for the project to happen, and aware that the call would be inevitably biased by our own communication capabilities and personal networks.

That said, the response rates for different regions of Brazil do generally match the distribution of research labs and graduate programs within the country (with some distortions likely caused by our personal networks, such as the large number of labs in Rio de Janeiro state), and seem to indicate a rather wide dissemination of the call. One way to visualize this would be to present the distribution of corresponding articles from the original studies selected for the replication (or even from the whole sample of articles obtained for experimental selection) along with the distribution of labs at different stages of the project in Figure S3, which generally show similar patterns. This would actually lend support to our statement that "the population of labs that performed replications was largely similar to the one that produced the original results" in the discussion.

> *Page 3, paragraph 2: "Based on the expertise of respondents and a feasibility analysis by the coordinating team, we selected 3 outcome assessment methods for replication" - Since this choice determined what was ultimately studied and who could participate, I would like to see more information to understand it: was it based on the most common expertise among respondents? How was feasibility defined and estimated?*

We tried to find the combination of methods that would maximize the number of labs that would be included in the project. This is explicitly stated in our Methods Selection document at *https://osf.io/qxdjt*, but could be stated more explicitly in the paper as well.

> *Page 3, paragraph 3: How was the manual screening performed? Was it done by one or more people? Was there double-screening to ensure reliability of the screening protocol? Did the authors use a specific decision tree or tool? How were conflicts between observers resolved? Were any other validation steps taken to ensure reliability? The same comments apply to the data extraction (who, how many, validation, protocol, etc.).*

We initially used single screening by three different reviewers (see *https://osf.io/6av7k/files/u5zdq* for criteria), as we were merely looking for a sample of experiments; thus, comprehensive inclusion of all eligible studies was not a priority. After this initial screening step, inclusions were confirmed in a consensus meeting with the three reviewers involved.

Data extraction was also done by a single individual, but the resulting data led to a protocol that was later checked by two reviewers who had access to the paper and were explicitly oriented to judge whether the protocol consisted in a valid replication. Thus, discrepancies between what was in the paper and what was included in the protocol could potentially be flagged at these stages (as they were in many cases). We do note, however, that this is likely not as effective to prevent errors as having data extracted independently, as reviewers may overlook mistakes more easily when comparing two documents rather than extracting data anew. We did find that some errors in extraction slipped by, such as an MTT experiment

where treatment concentration was inadvertently changed from mM to µM in a particular protocol step; this was picked up and corrected by 2 out of the 3 labs, but not by the third one, leading the latter replication to be invalidated.

> *Page 3, paragraph 3: As a non-expert, I would need more context about the expected average cost of experiments in this field; otherwise, I cannot assess how representative this sample is or whether potential biases may exist (e.g., cheaper experiments perhaps being expected to be less replicable than more expensive ones). Could expected costs also have affected the reduction in geographical coverage eventually observed in this study (Figure S3)?*

As stated in the manuscript, we initially capped experiments at a predicted cost of R$ 5.000 (around USD 1336 at that time), considering reagent cost alone (as equipment and labor was provided by labs), as mentioned in the manuscript. Exclusion rates for that reason were 12/74 (16%) for MTT experiments, 36/132 (27%) for PCR ones and 4/40 (10%) for EPM ones. This is stated at

This turned out to be an underestimation in many cases, especially as it did not account for pilot experiments, need for repetition, etc; thus, many experiments ended up costing considerably more than that ceiling. As we had included a contingency fund for those cases which we expected would occur , we avoided removing experiments from the sample for this reason as much as possible. Nevertheless, one elevated plus maze experiment ended up not being replicated for cost reasons, as the necessary rat strain was provided by a single facility in the country, meaning that a large number of rats would have to be acquired and transported to all labs at a cost that we were not able to cover.

As these costs were covered by the coordinating team, we do not feel that this is likely to underlie the reduction in geographical coverage. Other reasons related to lab structure could have led to labs in less well-resourced regions to leave the project, but they probably has nothing to do with the experiments selected.

That said, the cost cap does mean that the selection of experiments is not completely representative of the literature, but is enriched in relatively cheap and simple experiments which were able to perform (which was our next step for selecting the final sample of experiments. Exclusion rates due to lack of lab expertise and/or infrastructure to perform the experiment were 21/56 (37%) for MTT experiments, 67/89 (75%) for PCR ones and 7/34 (21%) for EPM experiments.

We will try adding some of this information to the flowchart in Figure 1, as we agree it provides more context on the representativeness of the selected experiments.

> *Page 6, paragraph 2: "(on a scale of 1 to 5)" - Could you clarify whether 1 means no deviations and 5 means everything deviated? Is that how it was phrased to participants? Was there a threshold used by the coordinating team to decide how many deviations were acceptable? (I would briefly clarify all scales mentioned below to allow easier interpretation throughout.)*

The scale ranged from 1 (No relevant differences) to 5 (Very relevant differences that prevent considering the study as a direct replication). This scale was used for both the lab and the validation committee scores, and is described at *https://osf.io/xgth2* (debriefing protocol) and *https://osf.io/e3fjg* (validation protocol).

For the validation committee, we did use a threshold (any score of 4 or a sum of scores of 10 or more among 3 evaluators) to decide what had to be discussed to decide on inclusion, as mentioned on Page 7 of the Methods. For the labs, we used no threshold labs answered the protocol deviation question as a scale, but the decision of whether to consider the study a valid replication or not was not tied to this score.

We can make both of these points (meaning of the scale and connection to lab's decision to consider the replication valid) clearer in the Methods section.

> *Page 6, paragraph 4: How were long-text answers (e.g., justifications) reviewed? Was this done manually by one or more members of the coordinating team, or using any text interpretation tool? What steps were taken to ensure the interpretation of these answers was as objective as possible?*

For the initial analysis of justifications, one reviewer read all answers and flagged those that seemed to concern reproducibility of the methods (e.g. "we replicated the protocol exactly as planned") rather than results reproducibility (e.g. "effects went in the opposite direction"). We then revised these answers among the whole coordinating team to decide whether we should contact the lab asking them to revise them. We can add this information to the Methods section.

For classifications of the justification into categories (i.e. Table S7), justifications were classified by two independent reviewers based on categories created after an initial inspection of the data, and discrepancies were resolved by consensus. We can add this information to the table legend.

> *Page 8, paragraph 1: "If issues were found, the lab and coordinating team reviewed them via email until the sources of errors were identified and corrected (see https://osf.io/58vsx for details)." - Could you please provide information about how often these disagreements arose and briefly explain their causes? I am struggling to understand why these discrepancies occurred and how frequently. Without more detail, the error rate presented in the next paragraph is a little concerning.*

After we extracted data from the lab spreadsheets and summarized the results by code, labs received the results by e-mail and were asked to fill in a form on whether the results were in agreement with what they had found (see details at *https://osf.io/nfr6y*). Discrepancies in results at least 1 experiment were noted by 36% of the 53 (out of 56) labs that responded. Many of these stemmed from the coordinating team misunderstanding issues such as group identity or experimental unit identification in the spreadsheet. Others had to do with different ways to perform calculations (e.g. relative gene expression or % time spent in open arms). In some cases, simple errors in data transcription or typos caused the discrepancy.

We were also surprised (and concerned) by the number of experiments in which we later found data errors that were not detected by this process (e.g. 18% of total). Our best understanding of this is that not every lab checked the results with the necessary care, as some errors were quite obvious, as in experiments in which sample size was different, or in which group labels were reversed. Ultimately, agreeing with a form that says "did you find any discrepancies?" may have been performed as a box-ticking exercise with little attention, and was probably not the ideal way to check data which led us to start reviewing results in live meetings afterwards. This is discussed in more detail in our challenges article (Amaral et al., 2026)

> *Page 8, paragraph 4: Please provide the version of any package or software used throughout, and make sure to cite R appropriately (R Core Team XXX).*

R 4.5.1 was used for the analysis. We can add this information (which was present in the data repository in the R session info.txt file) and provide the R reference in the manuscript as well.

> *In addition, did the authors calculate the log ratio of means (ROM/lnRR) using escalc()? If so, please report this.*
>
> *If not, I would recommend doing so, as escalc() implements recommended small-sample adjustments that produce slightly different values compared to a simple manual*

> *calculation of log(mean1/mean2).*

Yes, we did use the escalc() function for this calculation (for both the replications and the original effect sizes). We can mention this in the manuscript.

> *Page 10, paragraph 1: "Coefficients of variation from the original study were compared to the mean coefficient of variation of its replications using Wilcoxon's signed rank test" - I wonder how these CVs were calculated - whether simply as SD/mean or using escalc() from the R package metafor, which includes a correction for small-sample size. This may affect the fairness of the comparison, particularly since CVs from original studies are expected to be slightly overestimated given their smaller sample sizes relative to the replications.*

We calculated the coefficients of variation as the pooled SD divided by the mean of both group means. The reviewer is correct about the possibility of small-sample effects in this case (which we were not aware of). We will thus look into the possibility of implementing this via the escalc () function in the analysis of the revised manuscript.

We also acknowledge that this could be a source of bias in the comparisons between original and replication CVs (albeit likely a minor one). That said, we note that sample sizes are not always larger in the replication for some experiments with large original effects, power calculations sometimes yielded lower sample sizes in the individual replication, albeit infrequently. On average, though, replication sample sizes were indeed larger.

> *I also have concerns about using the mean CV of all replications and comparing it to a single CV value, as this ignores the uncertainty around that mean.*

This is indeed the case; that said, the CV of the original effect also has random error relative to the true population CV and in that case, there is no way to estimate the uncertainty, as we have a single measure of that parameter. So there is probably no way around ignoring uncertainty in this case.

We also note that we are looking for evidence of systematic CV inflation across all experiments (rather than for a statistically robust comparison between the CVs of any individual replication). For the sake of measuring this systematic inflation, the use of multiple experiments does allow us to estimate variability at the experiment level which should incorporate the lower-level variability between individual replications if this is not included in the model. Thus, we do not feel that our procedure introduced a systematic bias in the analysis at the experiment-level (although one could argue that it may lead to less precision).

> *An additional check could involve calculating the log coefficient of variation ratio (lnCVR; Nakagawa et al. 2015, Methods in Ecology and Evolution; implemented in escalc()) between the original CV and each replication CV, and running a random-effects (or multilevel) meta-analysis that accounts for shared-control non-independence. I believe this would provide a more robust approach, as it does not ignore the uncertainty around the mean CV of the replications - uncertainty that, if neglected, is expected to increase the likelihood of false positive findings. This concern would also apply to the subsequent analysis on absolute means.*

We thank the reviewer for this suggestion, which indeed seems like an option in this case. We will look into this possibility, although we cannot guarantee at the moment that we will implement it, as we were not previously familiar with the method and will have to study it in more detail.

> *Page 10, paragraph 2: The change in geographical distribution shown in Figure S3 appears rather striking, with western states disappearing step by step. Should the reader be concerned about the eventual geographical representability of the sample?*

Yes, but there are likely different reasons for that. Labs leaving after being included may have been due to those in less privileged regions of Brazil (e.g. the northern and western regions of Brazil, generally speaking) having more difficulty in persisting in the project. That said, most of the "disappearance" happens between registration and inclusion which usually has to do with the labs not working with the methods that were ultimately included in the project. We also note that most of the states that lose representation were those that had a single lab to begin with, which may make the visual pattern more striking than the actual trend (as states in the South/Southeast also lose labs, but don't disappear from the map).

We note again that we never planned to achieve geographical representativeness when recruiting the labs on the contrary, we were aiming to maximize the number of available labs to run the project. That said, we do agree that for the sake of examining whether the population of labs is similar to the one that generated the original experiments (a claim that we do make in the discussion), this representativeness is important to assess. Once more, to allow the reader to evaluate this, we plan to add an additional map to Figure S3 to describe the Brazilian states where the original experiments came from (based on corresponding author affiliations) in which a similar bias towards the South and Southeast Region can be observed.

> *Page 15, Figure 3A: I wonder whether adding 95% CIs calculated from the sampling variance of each ratio would improve interpretation and help readers appreciate the real differences between the dots (i.e., means) - along the lines of a forest plot.*

We agree that this would be useful information, and can experiment with the possibility, but our feeling is that the figure will likely become too noisy in cases where the 95% CIs overlap (which are quite frequent). If this is indeed the case, an option to allow the reader to examine this would be better to add an explicit link to the forest plots for each individual experiment (*https://osf.io/sx9gv*) in the figure legend.

> *Page 17, section "Predictors of replication success": It is unclear to me how the decision was made about which results from Figure 4 to present in the text. Intuitively, given that correlations were calculated for both t values and lnRR (and other metrics), I would have expected that whenever a result is highlighted in the text, the authors also report how it changes depending on the metric used - for example, the interesting result regarding the 5-year number of publications, whose correlation is notably lower when using lnRR (−0.31 vs. −0.18). Presenting this nuance in the text would reduce the risk of inadvertently giving the impression of cherry-picking.*

We selected the highest correlation values for each continuous outcome (t score and lnRR) and presented these separately in the text. This is a systematic way to perform the selection, but is obviously subject to the "winner's curse" effect. We agree that adding both metrics for each predictor would be a fair way to keep this in perspective for the reader, but we would have to think about how to do this without sounding too confusing (as results for the two main outcomes are quite different).

We do note, however, that the outcomes are indeed different and are expected to vary independently in some cases. For the correlation with replication probability predictions, for example, the effects in opposite directions would likely be expected, as larger original effect sizes will likely lead to larger probabilities to be assigned, but also to a higher possibility of effect size decrease. This low correlation between outcomes is probably something that should be pointed out and discussed in the revised manuscript.

> *Page 23, paragraph 1: (this comment should have come during the first % reported, but only in the discussion I realized how important this would be for comparing estimates) I wonder whether the authors should calculate 95% confidence intervals for all their percentages (and those of Errington et al.) using the Wilson method via the function*

> *binom.confint() in R, which handles extreme proportions (0% or 100%) more gracefully. This would ensure that uncertainty around these percentages is not neglected and would aid interpretation when comparisons are made.*

We had given this some thought when writing the manuscript – but ultimately opted not to include confidence intervals for our replication percentages and to use the replication rates as descriptive measures only (as done in other replication studies such as (Errington et al., 2021).

Even though we aimed for our sample of original experiments to be as systematic as possible, it is ultimately constrained by many factors (the choice of methods, the particular expertise of the labs, etc.) thus, adding confidence intervals represents the uncertainty around the replication rate of a very specific population of experiments, which is not directly comparable to those included in other replication efforts in any case.

We will reconsider whether we should include confidence intervals for replication rates: although doing this for every replication rate in Table 1 and Table 2 may end up being too much information, it could probably be done at least for the replication rates of the main analysis in the text. We note that calculating confidence intervals for percentages is straightforward, requiring only the numbers that are in the table thus, any reader that wants to estimate uncertainty for those rates should be able to do it easily.

We will also point out the uncertainty around the percentages mentioned in the discussion when comparing our replication rates with those of other studies, which we agree is an important issue to touch on.

> *In addition, in the next sentence, the authors are comparing correlation coefficients, at least verbally, these could in principle be transformed into Pearson's r and assigned 95% confidence intervals following meta-analytic workflows, which would better allow us to assess whether these correlations are meaningfully larger or smaller, and help avoid potentially misleading arguments.*

Both correlations in that case are non-parametric (e.g. Spearman's ρ), so they cannot be directly transformed into Pearson's r without making assumptions about the distribution (which we would probably avoid doing given the very marked outlier in our own). We can calculate a non-parametric confidence interval for our own correlation coefficient by resampling, but we will have to investigate whether this can be done using the available data from (Errington et al., 2021) (which is probably the case if effect sizes for all experiments have been shared).

> *Page 24, paragraph 2: The following result is really interesting and I would love for the authors to expand on it a little. There must be other meta-research studies that, despite not studying replicability directly, have explored a similar predictor: "Other features of the original article were generally uncorrelated with replication outcome, although large rates of publications by the last author were associated with lower replicability, suggesting that incentivizing publication volume may be counterproductive for the reliability of results."*

It is indeed interesting, and seems to confirm an intuition that has long been present in the reproducibility field, but actually has little evidence to support it: if anything, there is evidence in the opposite direction in psychology (Youyou et al., 2023), although they looked at cumulative publication number, while we used number of publications in a fixed interval.

We can expand a bit further on that finding: that said, we do note that the correlation is relatively weak and has a p value of 0.04. Thus, given the multiplicity of predictors would not be that unlikely to occur by chance, even though it seems intuitive. Thus, even though the

relationship seems intuitive, we think it should be considered tentative at best and would refrain from discussing it in too much detail.

> *Page 25, paragraph 1: I believe the authors could explore if there is evidence for "incorrect labeling of error bars (Cumming et al., 2007; Vaux, 2004)" by plotting log(SD) vs log(mean) across all original studies, and exploring if large outliers (i.e., points largely deviating from the positive regression) exist. That should provide some insights into whether some values reported as SD in the original studies were indeed SE, which I am assuming is what the authors of the study are referring to when they say "incorrect labelling of error bars" here.*

Yes, that is what we mean by “incorrect labeling of error bars” (as can be grasped from the cited references).

We can perform this regression, which seems relatively straightforward to do. That said, we note that another likely cause for outliers at least for cell line studies would be the use of different (and eventually inadequate) experimental units (e.g. having error bars that represent technical replicates of the same measurement rather than truly independent experiments). We suspect that this may have an even greater effect in terms of causing error bars not to express the same thing and the regression will not help in differentiating the two causes.

We should also note that different types of experiments may be expected to have very different SDs, so the regression is likely to have a lot of error associated with it. In particular, it’s probably worth doing separate regressions for each method, to account for the likely difference in CVs between animal and cell line experiments, for example. This could also help tease apart the two causes above, as the experimental unit problem mentioned above will likely only be observed for cell experiments.

> *Code: I could not engage with the data and code, but I would like to highlight that the organisation and clarity of the GitHub repository is of high quality.*

Thanks!

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *The authors conducted a large-scale replication effort of lab-based biomedical experiments with an emphasis on the country of origin and who conducted the replication experiments. The authors aimed to understand this context in both the outcomes produced, but also in the approach. Finally, the authors aimed to conduct multi-lab replications to provide richer data from the replications. Overall, the authors find replication rates that are like other large-scale replication efforts in the biomedical space. The authors provide rich detail into the three experimental techniques that were the focus of this effort, potential moderators of replication success, and challenges in conducting replications and coordinating a large-scale crowd-sourced effort.*
>
> *Strengths:*
>
> *The paper is outstanding in being transparent and calibrated in how the results are presented. While the authors were challenged by mundane aspects (e.g., difficulty with logistics), unexpected aspects (e.g., COVID pandemic), and very insightful aspects unique to conducting replications (e.g., experimental issues). The authors also provide variation in how they present the results, including confirmatory, multiverse, and exploratory analysis. A unique strength for this study is the rich in-depth insights about the process*

> *and interpretation of conducting replications, including predicting replication success in the lab-based biomedical space.*

We thank the reviewer for the compliments. Again, a more extensive list of insights can be found in our challenges article (Amaral et al., 2026), which we will cite in the revised version.

> *Weaknesses:*
>
> *The study has weaknesses that the authors acknowledge in their discussion, such as lower number of replications than originally planned that limited the intended effort to compare multiple experiments with multiple attempts against a single original experiment. Another weakness is the limited discussion connecting these findings to the Brazilian research ecosystem.*

We acknowledge the missing replications as a weakness, and we hope we have made that point clear in the discussion.

Concerning the Brazilian research ecosystem, we could try to explore this in more detail in the introduction. In particular, we believe that a better understanding of the Brazilian academic system, including its regional disparities and the general composition of its workforce (which is largely composed of undergraduate and graduate students), can be useful in interpreting some of the findings.

We can try to provide a bit more context at the end of the introduction (perhaps between the last 2 paragraphs, which would also address a point made by Reviewer #1), and also in different points of the discussion including those comparing replication rates with other studies or discussing infrastructural difficulties, some of which may be specific to the Brazilian context (such as difficulties in acquiring specific reagents or licenses). Still, we reiterate that, due to the lack of studies with comparable samples in other regions, we cannot tease apart the factors that are specific to Brazil from those affecting lab biology as a whole from the data alone.

References:

Amaral OB, Neves K, Wasilewska-Sampaio AP, Carneiro CF. 2019. The Brazilian Reproducibility Initiative. eLife 8:e41602. DOI: *https://doi.org/10.7554/eLife.41602*

Amaral OB, Valério B, Carneiro CFD, Mota GPS, Neves K, Abreu M, Tan PB. 2026. Challenges for building up confirmatory science in lab biology: lessons learned from the Brazilian Reproducibility Initiative. MetaArXiv, DOI: *https://doi.org/10.31222/osf.io/8y3tg_v1*

Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA. 2021. Investigating the replicability of preclinical cancer biology. eLife 10:e71601. DOI: *https://doi.org/10.7554/eLife.71601*

Fanelli D. 2010. Do pressures to publish increase scientists' bias? An empirical support from US states data. PLoS One 5:e10271. DOI: *https://doi.org/10.1371/journal.pone.0010271*

Fanelli D, Schleicher M, Fang FC, Casadevall A, Bik EM. 2022. Do individual and institutional predictors of misconduct vary by country? Results of a matched-control analysis of problematic image duplications. PLoS One 17:e0255334. DOI: *https://doi.org/10.1371/journal.pone.0255334*

Ioannidis jpa. 2005. why Most Published Research Findings Are False. PLoS Medicine 2. DOI: *https://doi.org/10.1371/journal.pmed.0020124*

Serghiou S, Contopoulos-Ioannidis DG, Boyack KW, Riedel N, Wallach JD, Ioannidis JPA. 2021. Assessment of transparency indicators across the biomedical literature: How open is open?

PLOS Biology 19:e3001107. DOI: *https://doi.org/10.1371/journal.pbio.3001107*

Smaldino PE, McElreath R. 2016. The natural selection of bad science. R Soc Open Sci 3:160384. DOI: *https://doi.org/10.1098/rsos.160384*, PMID: 27703703

Tyner AH, Abatayo AL, Daley M, Field S, Fox N, Haber NA, Hahn KM, Struhl MK, Mawhinney B, Miske O, Silverstein P, Soderberg CK, Stankov T, Abbasi A, Aberson CL, Aczel B, Adamkovič M, Albayrak N, Allen PJ, Andreychik M, Awtrey E, Axxe E, Azevedo F, Bader MD, Bago B, Bailey J, Bakker M, Banik G, Banks GC, Baskin E, Batruch A, Beatteay A, Behr SM, Berente N, Berry Z, Białkowski J, Bodroža B, Boeschoten L, Bognar M, Bokhove C, Bonfiglio D, Bouwman R, Brady TF, Braithwaite SR, Briceño Jiménez G, Brick C, Bricka T, Briker R, Brown AN, Brown GDA, van Aert RCM, Caldwell K, Capitan S, Capitán T, Chandler J, Charles T, Chartier CR, Chawdhary R, Cheng KJ, Chopik WJ, Clark B, Colvin VE, Comer CC, Costantini G, Coupé T, Cummins J, Czernatowicz-Kukuczka A, de Leeuw J, Dobolyi D, Druckman JN, Duan J, Dujmović M, Dunleavy DJ, Durkee PK, Emery C, Esterling KM, Evans TR, Fedor A, Fernández-Castilla B, Fiala N, Field JG, Fong N, Fonseca MA, Freeman ALJ, Freese J, Geiger SJ, Geng J, Getz LM, Geven LM, Gleibs IH, Gonzales DP, Gooty J, Gourdon-Kanhukamwe A, Greculescu C, Griffin SM, Grigoryan L, Grunow M, Gunby N, Hall B, Hanel PHP, Hannon EE, Harper S, Held MJ, Hickman L, Higgins NC, Hippel S, Hoeppner S, Hong S, Hostler TJ, Inzlicht M, Izydorczak K, Jaeger B, Jankowsky K, Jarke-Neuert J, Jensen M, Jokić B, Jolles D, Jolly P, Jones AM, Juanchich M, Kačmár P, Kapoor H, Keljanovic A, Koirala S, Kołczyńska M, Kouroupaki D, Kühnen U, Landgrave M, Larson MJ, Laulié L, Lawrence ACE, Le Forestier JM, Leahy KE, Lee S, Leslie J, Lewis SC, Limnios C, Lin H, Liu A-C, Lloyd JW, Ludvig EA, Lynott D, MacDonald J, Mallik P, Mallinson DJ, Marinazzo D, Martarelli CS, Matacotta J, McBride A, McHugh C, McMillan G, Méndez E, Metzger M, Michaelides MP, Michalak J, Micheli L, Miller JK, Milyavskaya M, Molden DC, Monjaras AG, Moreau D, Morrow A, Moya C, Mudrik L, Mulder LB, Munt KA, Nandi A, Nason K, Nast C, Nave G, Nax HH, Neubauer F, Nguyen PLL, Nichols AL, Nilsonne G, O'Boyle E, Oettinghaus J, Oh J, Oshana A, Ostermann T, Ostrowski RP, Oyebanjo A, Panczak R, Patrianakos J, Pavez I, Pavlov YG, Persson S, Perugini M, Peters K, Pieters C, Ponizovskiy V, Porter ND, Prenoveau JM, Purić D, Purol MF, Puthillam A, Quinn KA, Ramljak M, Reed WR, Ritchie M, Ritzau M, Roche SP, Rodela R, Röer JP, Ropovik I, Rothschild J, Saal J, Safadi H, Samaha J, Sanchez M, Sankaran S, Santos D, Sargent AC, Sauter M, Schmidt K, Schnabel L, Schroeder AN, Schuetz SW, Schuetze BA, Schulte-Mecklenbeck M, Schütz A, Sevigny EL, Shackleton E, Shafranek RM, Shaki S, Shakya S, Sirota M, Sisco MR, Sitnikov MM, Slevc LR, Smalarz L, Smith CT, Snyder JS, Sommet N, Sonmez F, Spellman BA, Stanulewicz-Buckley N, Stock G, Street CNH, Strømland E, Sundelin T, Syed M, Szabelska A, Szaszi B, Szumowska E, Tagat A, Täuber S, Tay L, Thapa S, Thatcher J, Tsaklakidou D, Tummers L, Turkovich E, Tutor MV, Urbanska K, van 't Veer AE, van Assen M, van de Ven N, van den Goorbergh R, Vargo EJ, Vaughn LA, Vazire S, Vermeulen JM, Vo DTH, Volkman V, Wagenmakers E-J, Wagner D, Walasek L, Walter F, Warmelink L, Wei L, Weißflog MI, Weller N, Wichman AL, Wilbiks J, Williams JR, Wolfe K, Wort F, Wright R, Wulff JN, Xue X, Yan VX, Yang Y, Yoon S, Žeželj I, Zhang Y, Ziano I, Zogmaister C, Zupan Z, Zwaan RA, Nosek BA, Errington TM. 2026. Investigating the replicability of the social and behavioural sciences. Nature 652:143–150. DOI: *https://doi.org/10.1038/s41586-025-10078-y*

Westlake H, David F, Tian Y, Krakovic K, Dolgikh A, Juravlev L, Bournonville TE de, Carboni A, Melcarne C, Shan T, Wang Y, Mu Y, Kotwal A, Pirko N, Boquete JP, Schüpfer F, Rommelaere S, Poidevin M, Liu Z, Kondo S, Ratnaparkhi GS, Chakrabarti S, Liu G, Masson F, Xiaoxue L, Hanson MA, Jiang H, Cara FD, Kurant E, Lemaitre B. 2026. Reproducibility of scientific claims in Drosophila immunity: A retrospective analysis of 400 publications. eLife 15. DOI: *https://doi.org/10.7554/eLife.108404.1*

Youyou W, Yang Y, Uzzi B. 2023. A discipline-wide investigation of the replicability of Psychology papers over the past two decades. Proceedings of the National Academy of Sciences 120:e2208863120. DOI: *https://doi.org/10.1073/pnas.2208863120*

https://doi.org/10.7554/eLife.111001.1.sa0