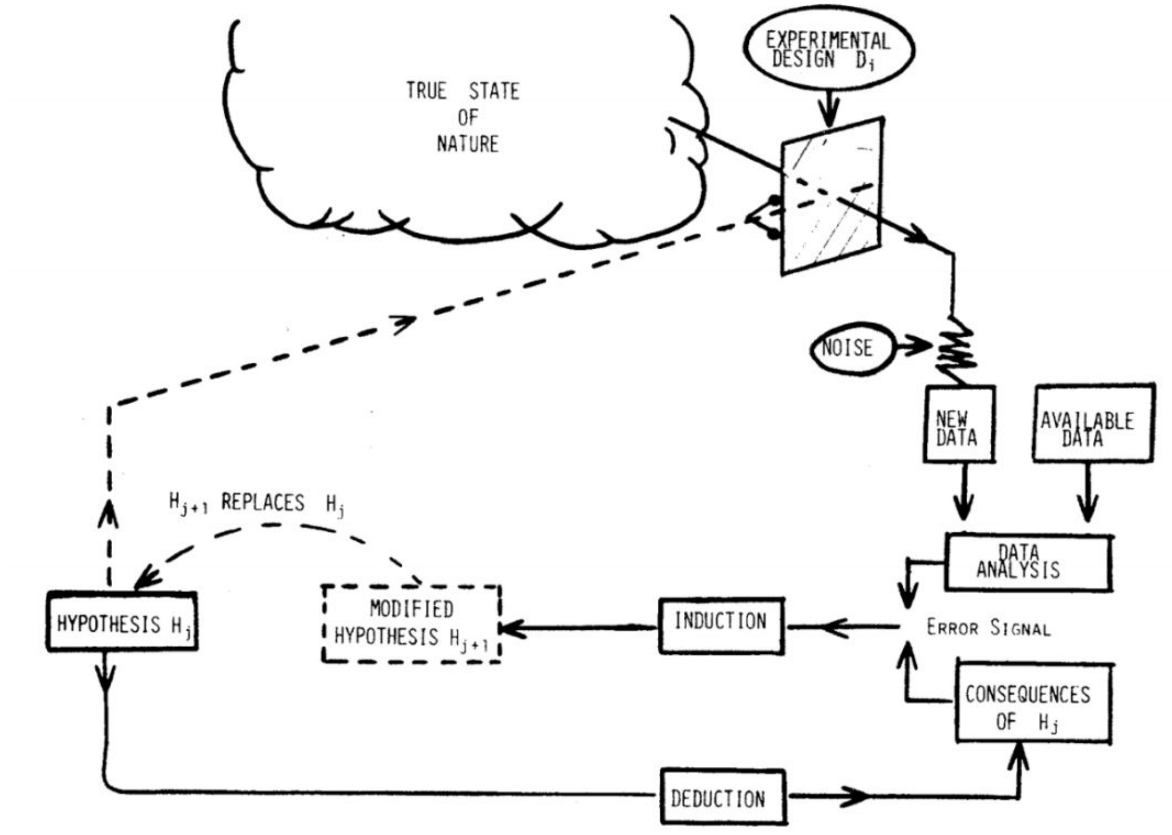
Impact and Reproducibility

A guide to reproducibility for funders of scientific research



George E. P. Box, 1976

Rationally, LLC

Reproducibility Within Reach

www.rationally.io

EXECUTIVE SUMMARY

In 2005, Stanford Professor John Ioannidis published his seminal paper "Why Most Published Research Findings are False," a compelling analysis of how current scientific practice results in findings that are more likely to be false than true. This replication crisis, as it is called, comes at a significant cost - contributing to the estimated $200 billion of wasted R&D dollars every year in biomedical research alone (Pain 2014).

While awareness of this replication crisis is on the rise (Goodman, Fanelli and Ioannidis 2016), the key contributing factors of poor study design and flawed data analysis practices remain ubiquitous (Baker 2016). Some of the best minds in clinical research have designed standards that, when followed, increase the replicability and transparency of research. While these standards have seen success, their use is far from universal and adherence is far from complete (Turner et al. 2012).

Funders should be aware of and concerned about reproducibility. If the research they fund cannot be replicated and thus is likely false, their capital is largely wasted.[1] Meanwhile, approximately 80% of researchers surveyed by *Nature* thought that funders, not just researchers, should do more to help improve reproducibility (Baker 2016). This can be taken as an invitation to collaboratively fix scientific research and increase grantmaking impact.

As holders of the purse, funders are in a unique position to drive the adoption of better practices. In this report, we discuss the causes of the replication crisis, its potential solutions and where grantmakers can make a difference.

[1] Specifically we speak to *preventable* irreplicability, as "some amount of irreproducibility is inevitable: profound insights can start as fragile signals, and sources of variability are infinite" ("Reality check," 2018).

IMPACT AND RESEARCH REPRODUCIBILITY

In 2016, public and private institutions funded $72 billion in academic scientific research (NSB 2015). As grantmaking organizations strive for greater efficiency and accountability in the use of these funds, investments are increasingly evaluated in terms of impact and results. Unfortunately this is easier said than done: 80% of surveyed funders consider measuring impact their greatest challenge (GEO 2009). Scientific research impact is particularly difficult to measure, because research often lays a foundation for future impact rather than delivering a better world itself. As a result, such funders base impact evaluation on proxy indicators such as the importance and dissemination of novel discoveries.

Unfortunately, genuine and useful discovery cannot be guaranteed, particularly not within any single project. Pressure to make novel or conclusive discoveries implicitly incentivizes investigators to find an effect, even if spurious (Edwards and Roy, 2017). Rigorously studied yet uninteresting findings may be disappointing, but at least they avoid a wild goose chase that sends good research dollars after bad. More importantly, research quality is more manageable than discovery: it is feasible to improve research design practices; nature itself, in contrast, must be obeyed before it is commanded.

> “Research findings from underpowered, early-phase clinical trials would be **true about one in four times**, or even less frequently if bias is present.”
>
> -Ioannidis, 2005

What is research reproducibility?

Before discussing solutions, we will delve into the meaning of research reproducibility and how it relates to the broader notion of reliability. In a 2016 paper, Goodman et al. distinguish different types of reproducibility:

1. **Methods** - sufficient description of the methodology such that the experiment could be, hypothetically, precisely repeated.
2. **Results** - conducting an independent study (with #1) and obtaining substantially equivalent results.
3. **Inferential** - Reaching qualitatively similar conclusions as the initial researchers after conducting an independent study or re-evaluating the original research.
4. **Robustness and Generalizability** - Respectively, achieving similar results across variation in procedure, and persistence of effect outside of the experimental context.

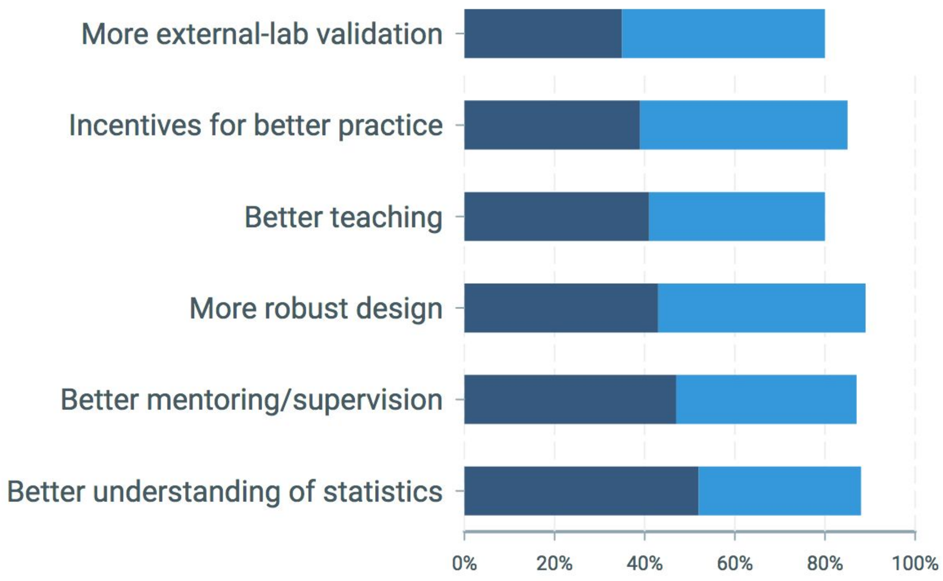
Reproducibility is not an aim in itself so much as it is a proxy for truth: "If a finding can be reliably repeated, it is likely to be true, and if it cannot be, its truth is in question" (Goodman, Fanelli and Ioannidis 2016). Methods and results reproducibility, while essential, should not be taken as the end of the line. All four elements are essential for the "operationalization of truth" - reducing into practice what research has yielded; for example, eventually translating preclinical oncology findings into life-saving cancer treatments. An experiment might meet criteria 1 and 2, yet due to confounders or systematic bias, fail to meet criterion 3. Or, findings that meet criteria 1 through 3 might fail to generalize outside of the experimental context and into real-world applications (e.g., the clinical context) (Treweek and Zwarenstein M, 2009).

What are the obstacles to reproducibility?

In 2014, the journal *Nature* polled 1,500 researchers about the causes and potential solutions to the replication crisis. Commonly cited causes included issues of incentives ("pressure to publish") and various issues around methodology and reporting ("selective reporting", "low statistical power or poor analysis", "poor experimental design" and lack of internal replication). Nearly 90% identified "More robust experimental design", "better statistics", and "better mentorship" as factors that would boost reproducibility (Baker, 2016). Outright fraud is comparatively uncommon (Fanelli 2009).

What could boost reproducibility?

Nature asked 1,500 scientists for their assessment. Below is a breakdown of the top factors researchers said are **very likely** or likely to improve this situation.



Baker, 2016

For those outside academia, all this may come as a surprise. How is it that PhD researchers, having received such extensive education, struggle with a core aspect of their function? More fundamental than the perverse incentives and methodological flaws listed above is that the attainment of knowledge about this world is vastly more difficult than our intuition would have us believe (Tversky and Kahneman 1982; Kahneman and Tversky 1982). In other words, “I saw it with my own eyes” has far less epistemic import than we tend to think. This applies to all humans, including researchers (Elliott and Resnik, 2015); but it is the inferences of researchers, unlike those of the general population, that are (sometimes) subject to flaw-revealing scrutiny.

Methodological flaws can be thought of as a failure to account for statistical noise: both the noise that naturally arises given the experimental circumstance, and that which researchers implicitly make more likely (Gelman and Loken, 2013). *Multiplicity* is a such a design flaw, wherein many hypotheses are explicitly or implicity tested in the same experiment. Even if these hypotheses are established before collection of data, multiplicity causes a significant risk of false positives: any empirical research runs some risk of false positive (i.e., an ostensible “effect” that is nothing more than statistical coincidence), and multiple comparisons amplifies this risk. Excessive subgroup analysis implicitly yet combinatorially increases the number of hypotheses being tested. For example, investigators reported (as a pedagogical demonstration) that certain zodiac signs “did not experience the same reduction in vascular mortality attributable to aspirin” as compared to other patients (p = .003) (Sun et al. 2014). Most reports of subgroup effects are not so readily identified as spurious - if the mechanism sounded plausible, this underlying statistical error is more difficult for the lay (and professional) reader to identify.

The prevalence of underpowered studies is another source of unreproducible research. For instance, Ioannidis analyzed 159 economics studies and found that their median statistical power was a mere 18% (Ioannidis, Stanley and Doucouliagos, 2017). Put another way, the median study had an 18% chance of finding an effect if present and a corresponding 82% chance of a *false negative*. The net effect defies intuition: selective reporting combined with underpowered research causes published research to contain a higher percentage of *false positives*, and is a key indictment in “Why Most Published Research Findings are False” (Ioannidis 2005).

Richard Feynman famously said, "The first principle is that you must not fool yourself and you are the easiest person to fool." The means by which scientists can fool themselves - as in, seeing what they want to see and finding what they want to find - is a list far lengthier than discussed here (Lindquist, 2018). Good methodological design helps researchers avoid fooling themselves, but it can be claimed that graduate education insufficiently equips PhD candidates with these tools (Bosch and Casadevall 2017).

SOLUTIONS

With an emphasis on the role of funders, we will now discuss opportunities for improvement in research reproducibility.

Encourage reproducible research via standards

In the early 1990s, a group of clinical researchers, methodologists, epidemiologists and journal editors convened with the goal of producing a quality rubric for randomized control trial (RCT) design. However, they agreed that this would be of little use, because the elements of good RCT design were not reported consistently in published research. Over the course of the next few years, this group produced a checklist for reporting called the CONSORT (Consolidated Standards of Reporting Trials) statement. Hailed as one of the "top health research milestones of the 20th century," this standard has achieved widespread endorsement and adoption. Due to CONSORT's success, more standards have been developed and are now available for nearly all research fields and designs.

While standards are an existing and compelling solution to increase research reproducibility, adoption is far from universal and adherence far from perfect (Guowei, 2018). Funders can influence culture by encouraging or even mandating the use of standards, but many researchers lack the expertise to do so. Solutions such as Rationally, a "TurboTax of research design," helps researchers adhere to these guidelines in a manner readily verifiable by grant-making institutions.

Beyond reporting

Opportunities to improve methodology exist beyond standards adoption. For example, CONSORT is self-described as *minimum* standards and only prescribes what researchers should *report* - not what they should *do*. The Patient-Centered Outcomes Research Institute has produced a methodology standard that goes beyond reporting requirements (PCORI, 2019). Such standards may be of interest to funders looking to further ensure the importance, reproducibility and generalizability of the research they fund.

Another area of interest is statistical inference: how are data interpreted to support a research claim? Null-hypothesis significance testing (NHST) and the arbitrary $p < .05$ threshold for so-called 'statistical significance' are increasingly under fire and for good reason (McShane et al. 2018; Amrhein 2019). Bayesian methods are often suggested as an alternative, but the question of how, precisely, to implement such methods is another subject of controversy (Efron 2013). Funders may wish to avoid stepping into this controversy and instead focus their efforts by encouraging the adoption of widely-accepted standards. CONSORT, PCORI and other standards offer some guidance around statistical methods, and more recently, guidelines for the statistical analysis plans (SAPs) for clinical trials have been published (Gamble et al. 2017). Future revisions may speak to adaptive designs and Bayesian analysis.

Promote open science practices

Open Access is gaining steam, with Plan S and the UC systems' Elsevier discontinuation making waves (Rabesandratana 2019; Fox 2019). These momentous changes may be controversial, but even those who oppose Plan S acknowledge benefits of open access (e.g., Plan S Open Letter, 2018). However, "Open Science" does not stop at open access - it refers broadly to an increase of transparency on many fronts, including the following (Nosek et al 2015):

1. **Data transparency** - Full research data should be made available to facilitate reproductions and inclusion in subsequent analysis. Additionally, data should be

preserved in a standard/open form and made machine-readable via metadata and structure.

2. **Code transparency** - All code involved in data collection and analysis should be available, ideally in a readily accessible, easily executable form (such as with containerization or computational notebooks like Jupyter).
3. **Preregistration** - Essential components of the study design should be disclosed prior to data collection, such as hypotheses and outcome measures, to prevent HARKing (Hypothesizing After the Results are Known - a common practice that substantially increases false positive rates).

Funders can require that the data and code from research they fund be made available, and that all confirmatory research be pre-registered (via clinicaltrials.gov, OSF, AsPredicted, etc) before data collection.

Take a hard look at feasibility and efficiency

Feasibility plays an indirect but important role in research reliability: if scope exceeds funds, corners will necessarily be cut. For this reason and others, investigators need and want funders as their allies in delivering reproducible research (Baker 2016). What follows are specific areas in which funders can influence feasibility as an essential element of research reproducibility.

Invest in signal over noise: ensure adequate sample size

Clinical trial costs per patient increased by approximately 4.5x between 1989 and 2011 (Berndt and Cockburn 2014), putting a substantial price tag on well-powered research. These high per-sample costs are not limited to clinical research: proteomics (Pascovici et al. 2016), neuroscience (Poldrack 2019), research involving animal models (Doke and Dhawale 2015) and other fields experience high per-subject costs. Creating a dangerous tension between costs and reproducibility, studies with too few participants have a high likelihood of yielding false positives and negatives.

> “It is difficult to get a man to understand something, when his salary depends on his not understanding it.”
>
> -Upton Sinclair

Revealingly, the rate of reporting prospective power analysis (from which sample size is determined) has been as low as 3% even in high impact factor journals (Tressoli and Giofré 2015). As illustrated by the Upton Sinclair quote, we wonder if researchers may be implicitly discouraged from taking this crucial step because it might show that proper study power isn’t within budget.

Funders have a vested interest in funding research that reduces risk of false discovery. At minimum, they can ask investigators to justify their sample size and corresponding assumptions. Ideally, funders would also work with the scientific community to understand the costs of confirmatory, reproducible research and seek to minimize the rate at which the projects they fund yield a false “eureka!” Given the low rate of external replication, it is unlikely that a false discovery will be revealed as such (Ioannidis 2012). More likely, we suspect, is that future projects building upon that latent false discovery will fail to produce the desired results. Tools such as Rationally can help funders manage their portfolio with regard to reproducibility and the corresponding impact on ROI and risk.

Increase efficiency with study design win/win

While reproducibility can come at the cost of feasibility, certain experimental designs can increase research validity while reducing costs. As one example of such a win/win, within-subject designs can deliver higher study power for a given sample size (Wellek and Blettner 2012) and produce less-biased results (Normand 2016; Gelman 2016). However, these designs have limited applicability and add complexity to statistical analysis. Still, if the requirements for applicability are met, then statistical analysis should present no barrier; it seems a safe assumption that the cost of procuring the necessary statistical expertise is lower than the cost of increasing the sample size. If funders lack the in-house expertise to assess the efficiency of the proposed experimental design, tools such as Rationally can bring both the investigators' and funders' attention to opportunities for optimization.

Manage uncertainty with adaptive design

We suggest that the current model of research funding makes a risky implicit assumption: that researchers know enough about the data they will find that they can estimate how many dollars are required to obtain reproducible results. For example, sample size calculations depend upon the predicted effect size, which is notoriously difficult to estimate even with prior research and pilots (Gelman 2017, 2018). Formal adaptive designs are growing in popularity; for example, such methods are recommended in the FDA's draft guidance for drug and biologic developers for their efficiency and flexibility in the face of unknowns (FDA 2018). Informal adaptive designs, such as 2-Stage exploratory/confirmatory, can be overlaid on the funder's formative evaluation process. While adaptive designs deliver a host of benefits, they present a reproducibility risk if done improperly: all adaptations must be planned prior to data collection and designed according to best practices. Otherwise they may create even more degrees of freedom with which false positives can be generated.

2-Stage Designs

Research can be exploratory (hypothesis generating), confirmatory (hypothesis confirming) or a combination thereof (i.e., a primary confirmatory but measuring secondary outcomes and subgroups for which findings should be reported as exploratory). However, this distinction is often made improperly (Berendt 2016). Furthermore, Gelman (2018) argues that researchers demonstrate overconfidence about their findings arising from pilot studies despite embracing the label of preliminary research: such studies are so noisy, he argues, that they should not be taken as an estimation of treatment effect and only as a demonstration of feasibility.

What can funders do to avoid implicitly encouraging overconfidence (e.g., exploratory research presented as confirmatory, or overstated pilot findings)? If investigators and funders wish for a research outcome beyond mere feasibility assessment, but don't yet have the requisite data to justify or budget for a rigorously designed confirmatory study, 2-Stage designs (as described in Nosek, Spies and Motyl 2012; Gelman and Loken 2013; see Appendix) may be a good solution. Such designs have an initial, flexible exploratory phase in which findings can inform subsequent hypotheses and iterations. An interesting finding from Stage 1 (feasibility) informs the Stage 2 (confirmatory) hypothesis and study design, which are ideally preregistered and tested on freshly collected data. Grantmakers can fit into this model by funding an initial, flexible exploratory Stage 1 and offering up follow-up funding for the confirmatory Stage 2.

Formal Adaptive Designs

Adaptive study designs can optimize and occasionally even save confirmatory research. Such designs involve predetermined decision criteria to be made at certain milestones to reassess sample size, adjust study arm allocation, and determine whether the study is even worth continuing. If funders work with investigators on these types of designs, grant funds can be saved by cutting losses early or augmenting funding to avoid what would otherwise be unpowered and thus unreliable research. As mentioned above, adaptive methods require diligent upfront design with careful controls to avoid biased results. Tools such as Rationally help with adaptive study design and the timing and content of the funder's formative evaluation.

CASE STUDY

To demonstrate the role of reproducibility in funder ROI, we performed a simulation study based on the work of Ioannidis (2005). Our model takes in cost-per-participant to calculate the *sample size* possible at various grant amounts and at different levels of *bias* (an inverse measure of research reliability - more bias means lower reproducibility), based on other parameters such as estimated effect size and the prior probability of the hypothesis.

The average value of a grant is calculated by considering four scenarios:

1. A real effect exists and study finds it (**1A**)
2. A real effect exists but the study doesn't find it (**1B**)
3. There is no effect, yet one is spuriously found (**2A**)
4. There is no effect and none is found (**2B**)

The values of each scenario are calculated as follows:

	Positive Finding (A)	**Negative Finding (B)**
A Real Effect (1)	True positive rate **x** Genuine discovery value	False negative rate **x** False negative opportunity cost
No Effect (2)	False findings rate **x** Goose chase & real-world costs	True negative rate **x** True negative value

Yielding a total study value calculation:

Total study value

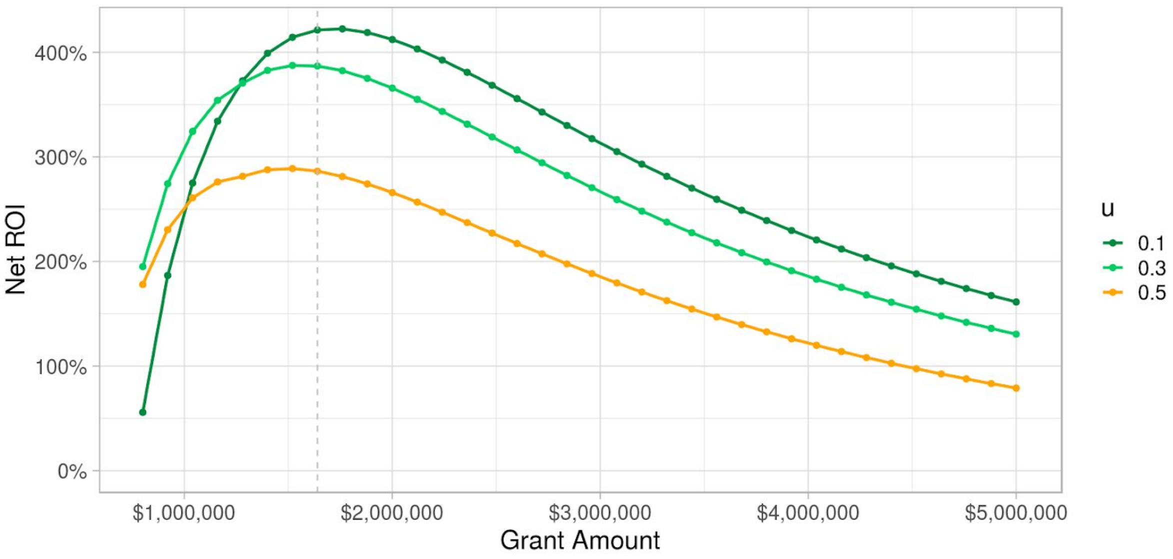
= Value of positive findings + Value of negative findings

= 1A + 2B - 2A - 2B

Before going into scenarios, we'd like to issue this disclaimer: while this model demonstrates the significant impact of research reproducibility on grant ROI, this model is limited by both its simplicity, complexity and sensitivity to hard-to-estimate composite parameters. Its numbers should not be taken as conclusive. Further details can be found on our ShinyApp model.

Scenario 1: Randomized Control Trial

Using prior probability estimates from Ioannidis 2005, we evaluated a randomized control trial with predicted medium effect size and a prior odds that would typically exist for a phase I/II trial (20%). 80% of the grant amount goes to fund a sample with a cost per patient of $10k against. Assuming that a genuine discovery is worth $80m (a parameter representing the future benefit of this discovery minus subsequent R&D costs, averaged across a portfolio) and that “false discoveries making it to the real world” cost $30m (a number that represents diminished reputation, lawsuits, angry patients, etc arising from a false discovery making it through all stages of research and into production), below is the ROI curve at different funding levels.



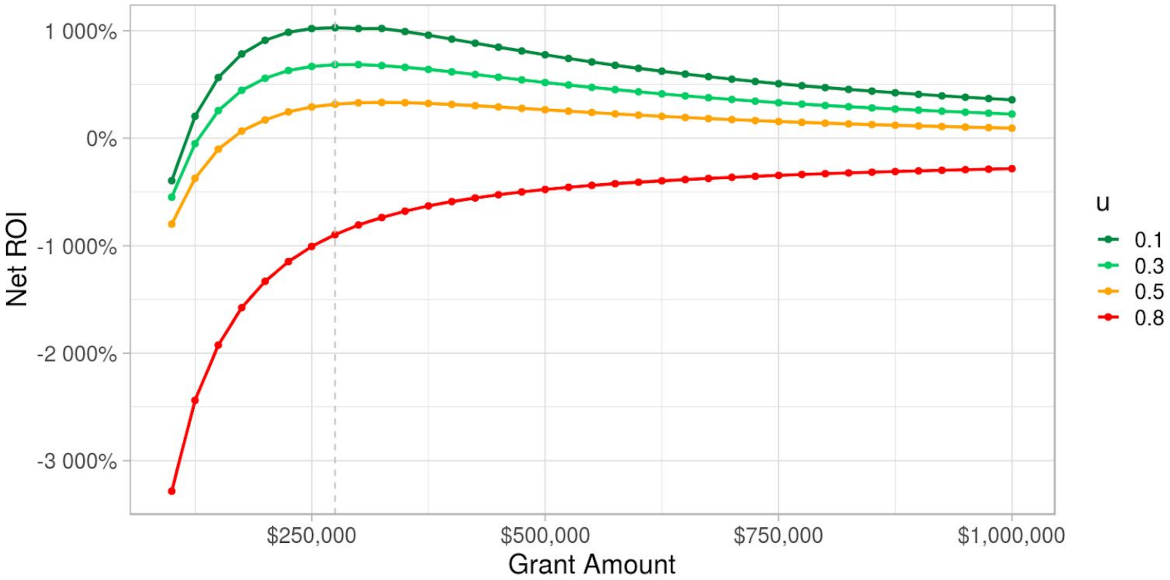
Bias, represented by *u*, is as defined as: “the proportion of probed analyses that would not have been ‘research findings,’ but nevertheless end up presented and reported as such, because of bias” (Ioannidis 2005). *u* = 0.1 is the bias one could expect of a well-designed, well-executed RCT, while on the other end of the spectrum *u* = 0.8 (not shown) is the bias one could expect of a poorly-designed, poorly-executed RCT or of a high multiplicity exploratory study.

Notice that for all bias curves, the far left side of the chart shows lower ROI, reflecting the larger number of false positives and negatives in smaller sample sizes. Beyond a certain sample size, however, ROI declines as the additional study power achieved no longer justifies the cost of more study participants. The gray dashed line indicates the first grant amount funding a study of 80% power or greater.

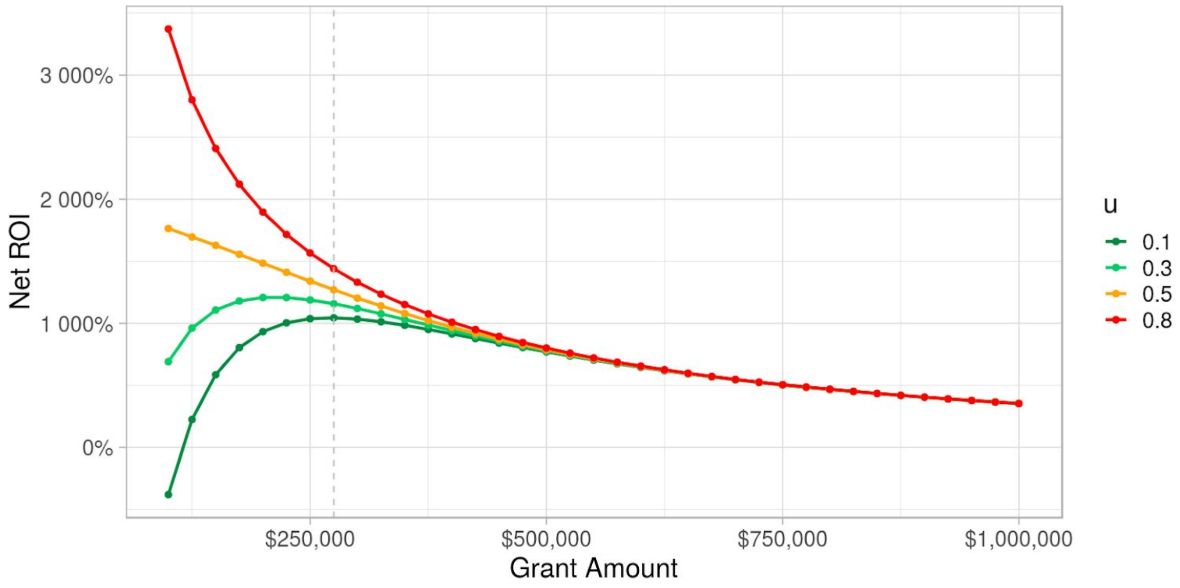
Also noteworthy is the role of bias in ROI, which impacts overall ROI but also the shape of the curve. For *u* = 0.3 and *u* = 0.5, peak ROI is achieved before a study reaches 80% power, because the "signal" arising from higher power does less to offset the "noise" caused by bias. In this scenario, good methods win: *u* = 0.1 achieves the highest ROI.

Scenario 2: Bias pays off?

Let's take another scenario, this time with lower prior odds (10% instead of 20%). The following includes a *u* = 0.8 (very high bias, like a poor quality RCT or high multiplicity exploratory research).



In a world with no consequences, however, bias wins. Below is this same scenario, except where false discoveries are "free" and true negatives aren't valued.



Thus if the costs of false discovery are truly low, or more likely, those costs are borne by "someone else" (a Prisoner's Dilemma scenario), then one could reach the conclusion that poor methods and underpowered studies pay off (Smaldino and McElreath 2016). Thus, it is important to consider the costs of false discoveries, internal and external, in grantmaking decisions.

Additionally, the value of true negatives should be considered. Ioannidis (2005) says:

> *A negative finding can then refute not only a specific proposed claim, but a whole field or considerable portion thereof. Selecting the performance of large-scale studies based on narrow-minded criteria, such as the marketing promotion of a specific drug, is largely wasted research.*

If employed as a means by which to focus on more profitable future investments, negative findings are potentially a gold mine.

To conclude this case study, reproducibility plays a direct and significant role in grant ROI and therefore impact. Costs arising from the goose chases following false discoveries are formidable, as is the opportunity cost of generating false negatives. Luckily, funders can improve this situation by encouraging grant applicants to design low-bias studies (i.e. according to best practices), to work with researchers on achieving proper study power, and to broaden their view of ROI to consider the costs of false discoveries and gains of true negatives.

CONCLUSION

The critical takeaways of this white paper are:

1. **If research is not reproducible, it is not impactful.** While impact cannot always be measured, much less guaranteed, reproducibility is both a critical target and a manageable one.
2. **Most research isn't reproducible**. While this differs from field to field, theoretical and empirical meta-research has shown that a shocking amount of research cannot be reproduced and thus is likely false.

In a 2014 poll, approximately 80% of researchers believed that funders should do more to improve research reproducibility (Baker 2016), suggesting an openness to collaboration. Grantmakers are in a unique position to drive adoption of the best practices in research design and reporting to yield more reliable research, and in doing so, ensure that they are contributing more signal than noise to the research ecosystem.

Summary of recommendations

1. Encourage reproducible research via standards
2. Promote open science practices
3. Take a hard look at feasibility and efficiency
    a. Invest in signal over noise: ensure adequate sample size
    b. Increase efficiency with study design win/win
4. Manage uncertainty with adaptive designs
5. Recognize research reproducibility as a key performance indicator in impact evaluation

ABOUT RATIONALLY

Rationally puts replicability within reach. As a framework for experimental design, we make reproducible research design easier, more fundable and more visible. Our flexible SaaS framework guides grant applicants through the research design portion of proposal submission, ensuring compliance with best practices and a funder's specific requirements. Once research proposals are funded, Rationally helps researchers follow reliable research practices on an ongoing basis and report findings in a transparent, standards-compliant fashion. Finally, we help grant-making institutions add reproducibility to their impact evaluations for individual projects and across their portfolio, to ensure that their funds are creating signal rather than noise.

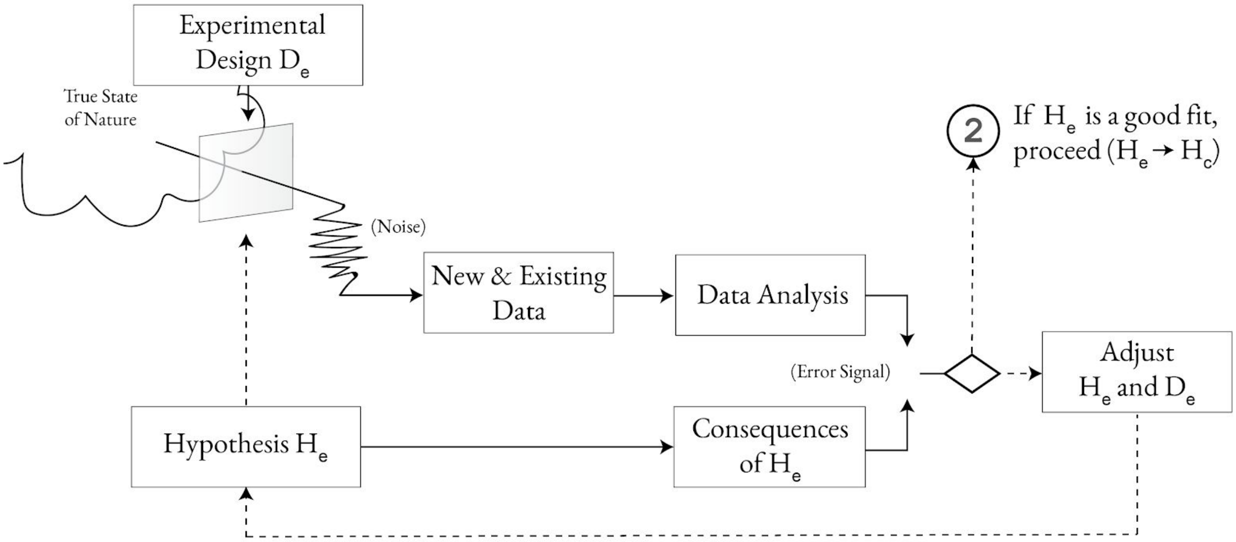
Learn more at www.rationally.io or by contacting kristin@rationally.io. Learn more about the replication crisis via this infographic.

APPENDIX

2-Stage Design

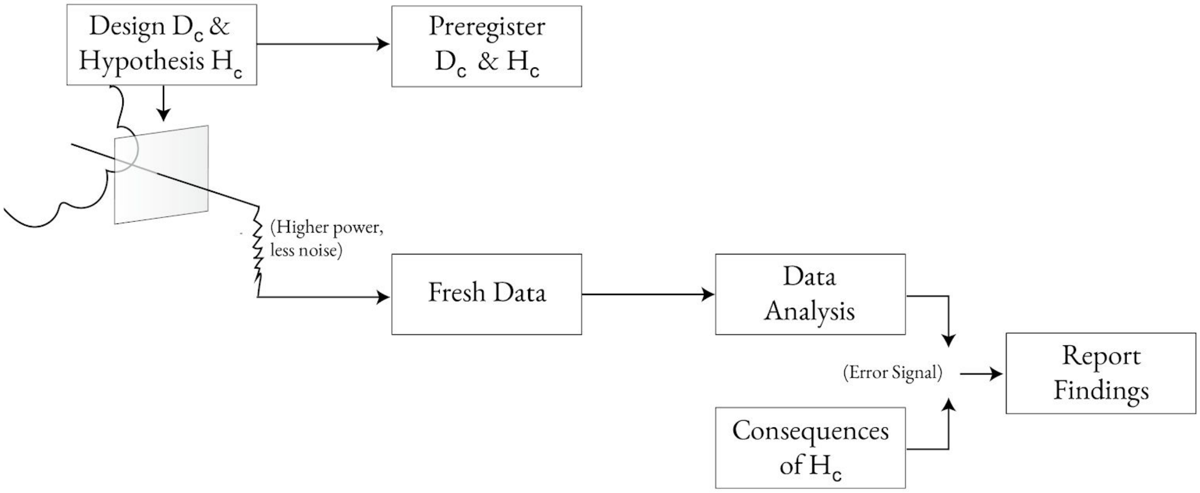
Stage 1: Exploratory

(Based on Box 1976)



Stage 2: Confirmatory

(Inspired by Nosek, Spies and Motyl 2012; Gelman and Loken 2013; Box 1976)


Design D_c & Hypothesis H_c
Preregister D_c & H_c
(Higher power, less noise)
Fresh Data
Data Analysis
(Error Signal)
Report Findings
Consequences of H_c

REFERENCES

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in Science. *Circulation Research*, *116*(1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Berendt, L., Callréus, T., Petersen, L. G., Bach, K. F., Poulsen, H. E., & Dalhoff, K. (2016). From protocol to published report: a study of consistency in the reporting of academic drug trials. *Trials*, *17*(1), 100. https://doi.org/10.1186/s13063-016-1189-4

Berndt, E., & Cockburn, I. (2014). Price indexes for clinical trial research: a feasibility study. *Monthly Labor Review*. https://doi.org/10.21916/mlr.2014.22

Boja, E. S., Kinsinger, C. R., Rodriguez, H., Srinivas, P., & Participants, on behalf of O. I. W. (2014). Integration of omics sciences to advance biology and medicine. *Clinical Proteomics*, *11*(1), 45. https://doi.org/10.1186/1559-0275-11-45

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

CONSORT. (n.d.). CONSORT History. Retrieved January 4, 2019, from http://www.consort-statement.org/about-consort/history

Edwards, M. A., & Roy, S. (2017). Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environmental Engineering Science*, *34*(1), 51–61. https://doi.org/10.1089/ees.2016.0223

Efron, B. (2013). Bayes' theorem in the 21st century. *Science*, *340*(6137).

Elliott, K. C., & Resnik, D. B. (2015). Scientific Reproducibility, Human Error, and Public Policy. *BioScience*, *65*(1), 5–6. https://doi.org/10.1093/biosci/biu197

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLoS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

FDA. (2018). Adaptive Designs for Clinical Trials of Drugs and Biologics - Guidance for Industry. In *Fda*. https://doi.org/10.1037/0894-4105.19.2.223

Fox, A. (2019). University of California boycotts publishing giant Elsevier over journal costs and open access. *Science*. https://doi.org/10.1126/science.aax1895

Gamble, C., Krishan, A., Stocken, D., Lewis, S., Juszczak, E., Doré, C., … Loder, E. (2017). Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *JAMA*, *318*(23), 2337. https://doi.org/10.1001/jama.2017.18556

Gelman, A. (2018). The purpose of a pilot study is to demonstrate the feasibility of an experiment, not to estimate the treatment effect. *Statistical Modeling, Causal Inference, and Social Science*. Retrieved from https://statmodeling.stat.columbia.edu/2018/03/20/purpose-pilot-study-demonstrate-feasibility-experiment-not-estimate-treatment-effect/

Gelman, A. (2017). Statistical Modeling, Causal Inference, and Social Science. *Statistical*

Modeling, Causal Inference, and Social Science. Retrieved from https://statmodeling.stat.columbia.edu/2017/12/04/80-power-lie/

Gelman, A. (2018, June 18). Power analysis and NIH-style statistical practice: What's the implicit model? *Statistical Modeling, Causal Inference, and Social Science*. Retrieved from https://statmodeling.stat.columbia.edu/2018/06/18/power-analysis-nih-style-statistical-practice-whats-implicit-model/

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time**. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Grantmakers for Effective Organizations. (2009). Evaluation in Philanthropy: Perspectives From the Field - IssueLab. Retrieved March 28, 2019, from Grantmakers for Effective Organizations website: https://www.issuelab.org/resource/geo-2009-evaluation-in-philanthropy-perspectives-from-the-field.html

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science*, *5*(1), 171511. https://doi.org/10.1098/rsos.171511

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, *127*(605), F236–F265. https://doi.org/10.1111/ecoj.12461

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*(2), 123–141. https://doi.org/10.1016/0010-0277(82)90022-1

Kimmelman, J., Mogil, J. S., & Dirnagl, U. (2014). Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biology*, *12*(5), e1001863. https://doi.org/10.1371/journal.pbio.1001863

Li, G., Bhatt, M., Wang, M., Mbuagbaw, L., Samaan, Z., & Thabane, L. (2018). Enhancing primary reports of randomized controlled trials: Three most common challenges and suggested solutions. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2595–2599. https://doi.org/10.1073/pnas.1708286114

Lindquist, K. (2018). The Replication Crisis. Retrieved January 4, 2019, from Rationally website: https://app.rationally.io/replication

Luce, B. R., Kramer, J. M., Goodman, S. N., Connor, J. T., Tunis, S., Whicher, D., & Schwartz, J. S. (2009). Rethinking Randomized Clinical Trials for Comparative Effectiveness Research: The Need for Transformational Change. *Annals of Internal Medicine*, *151*(3), 206. https://doi.org/10.7326/0003-4819-151-3-200908040-00126

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, *73*(sup1), 235–245.

https://doi.org/10.1080/00031305.2018.1527253

National Science Board. (2018). Report - S&amp;E Indicators 2018 | NSF - National Science Foundation. Retrieved April 1, 2019, from Science & Engineering Indicators website: https://www.nsf.gov/statistics/2018/nsb20181/report/sections/academic-research-and-development/expenditures-and-funding-for-academic-r-d#sources-of-support-for-academic-r-d

Normand, M. P. (2016). Less Is More: Psychologists Can Learn More by Studying Fewer People. *Frontiers in Psychology*, *7*, 934. https://doi.org/10.3389/fpsyg.2016.00934

Nosek, B. A., Alter, G., Banks, G., Borsboom, D., Bowman, S. et. al. (2015). Promoting an open research culture. *Science*, *348*(6242). https://doi.org/10.1126/science.aab2374

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Pain, E. (2014). Taking the "Waste" Out of Biomedical Research. *Science*. https://doi.org/10.1126/science.caredit.a1400030

Pascovici, D., Handler, D. C. L., Wu, J. X., & Haynes, P. A. (2016). Multiple testing corrections in quantitative proteomics: A useful but blunt tool. *PROTEOMICS*, *16*(18), 2448–2453. https://doi.org/10.1002/pmic.201600044

PCORI. (2019). *PCORI METHODOLOGY STANDARDS*. Retrieved from https://www.pcori.org/sites/default/files/PCORI-Methodology-Standards.pdf

Poldrack, R. A. (2019). The Costs of Reproducibility. *Neuron*, *101*(1), 11–14. https://doi.org/10.1016/J.NEURON.2018.11.030

Rabesandratana, T. (2019). Will the world embrace Plan S, the radical proposal to mandate open access to science papers? *Science*. https://doi.org/10.1126/science.aaw5306

Reality check on reproducibility. (2018). *Nature,* 533(7604), 437–437. https://doi.org/doi:10.1038/533437a

Research Community. (2018). Plan S Open Letter. *Zendo*. https://doi.org/10.5281/ZENODO.1477914

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Sun, X., Ioannidis, J. P. A., Agoritsas, T., Alba, A. C., & Guyatt, G. (2014). How to Use a Subgroup Analysis. *JAMA*, *311*(4), 405. https://doi.org/10.1001/jama.2013.285063

Tressoldi, P. E., & Giofré, D. (2015). The pervasive avoidance of prospective statistical power: major consequences and practical solutions. *Frontiers in Psychology*, *6*, 726. https://doi.org/10.3389/fpsyg.2015.00726

Treweek, S., & Zwarenstein, M. (2009). Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials*, *10*(1), 37. https://doi.org/10.1186/1745-6215-10-37

Turner, L., Shamseer, L., Altman, D. G., Schulz, K. F., & Moher, D. (2012). Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Systematic Reviews*, *1*, 60. https://doi.org/10.1186/2046-4053-1-60

Turner, L., Shamseer, L., Altman, D. G., Weeks, L., Peters, J., Kober, T., ... Moher, D. (2012). Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database of Systematic Reviews*, *11*, MR000030. https://doi.org/10.1002/14651858.MR000030.pub2

Tversky, A., & Kahneman, D. (1982). Causal schemas in judgments under uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty* (pp. 117–128). https://doi.org/10.1017/CBO9780511809477.009

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Wellek, S., & Blettner, M. (2012). On the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International*, *109*(15), 276–281. https://doi.org/10.3238/arztebl.2012.0276