Reporting Field Trials
Protocols, Registration, and Reporting Guidelines

Larry V. Hedges

Northwestern University

Prepared for the 2025 IES/NCER Summer Research Training Institute at Northwestern University

Challenges to Empirical Science

The introduction of randomized trials vastly improved medical science

Randomized trials became the “gold standard” of evidence

But just doing randomized trials alone does not guarantee sound and cumulative science

Medical science has slowly been confronting its own crisis in empirical work

The smart money says education and the social sciences will confront the same challenges

But we can benefit from what medicine has learned

Ioannidis in *PLOS Medicine*, 2005, 2(8):e124

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2

Ioannidis in *JAMA*, 2005, 294(2), 218-228

JAMA The Journal of the American Medical Association

Home Current Issue All Issues Online First Specialties & Topics CME Multim

July 13, 2005, Vol 294, No. 2 >

< Previous Article Next Article >

Original Contribution | July 13, 2005

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research FREE

John P. A. Ioannidis, MD

[+] Author Affiliations

JAMA. 2005;294(2):218-228. doi:10.1001/jama.294.2.218.

Text Size: A A A

The Crisis in Medical Research

Anomalies have occurred that bring the scientific *process* into question

Contradicted and initially stronger effects in highly cited clinical research (Ioannidis, 2005)

> *Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged. Five of 6 highly-cited nonrandomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials (p = .008).*

Replication Problems in medicine are not limited to published clinical trials!

Nearly 2/3 of 67 in-house Projects Could Not Replicate Published Data

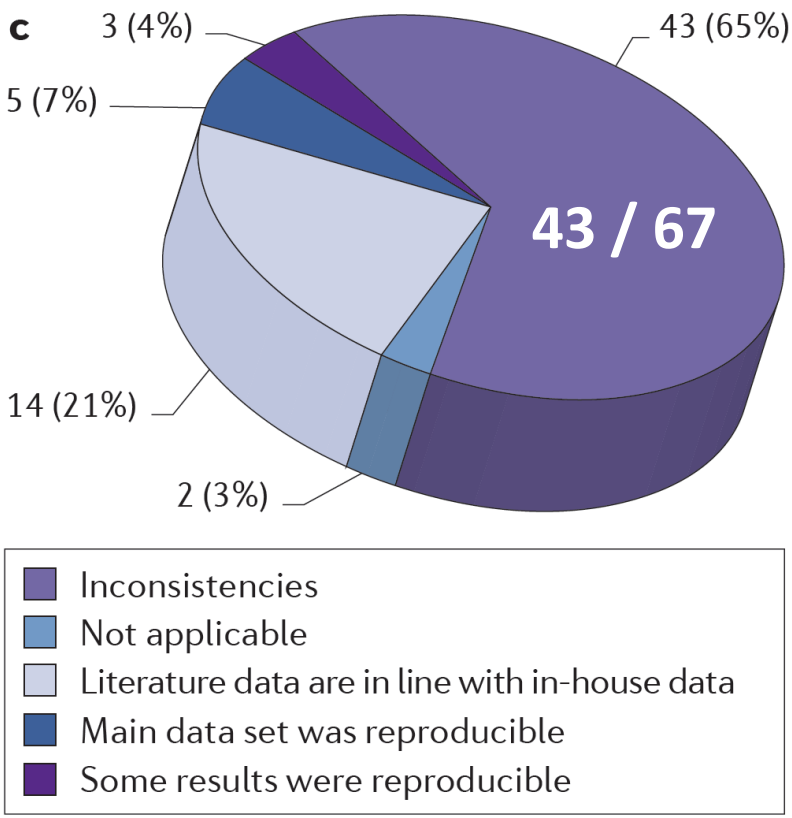
(Silberberg, *Nature Drug Discovery*, 2011, 10, 712-713)

Believe it or not: how much can we rely on published data on potential drug targets?

Prinz, Schlange and Asadullah

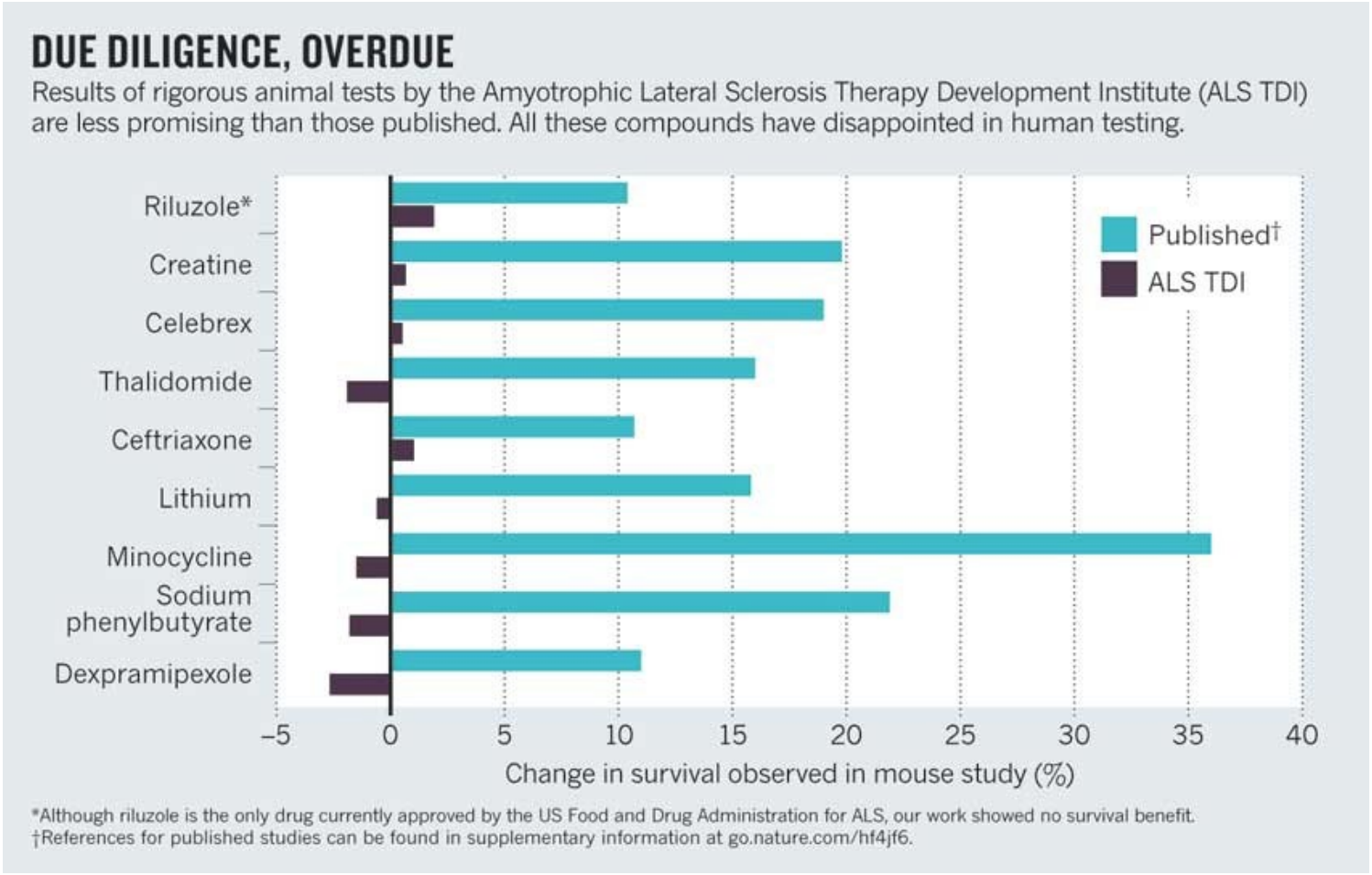
Bayer HealthCare

Nature Reviews Drug Discovery
2011; 10:712-713



How Frequent are Failures to Replicate in Preclinical Studies?

(Perrin, *Nature*, 2014, 507, 423-425)



Medical Science Subject to Public Scrutiny

The Economist, October 19, 2013

The Economist | World politics | Business & finance | Economics | Science & technology | Cu

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition

Like 13k Tweet 1,017

A SIMPLE idea underpins science: "trust, but verify". Results should always be subject to challenge from experiment. That simple but powerful idea has generated a vast body of knowledge. Since its birth in the 17th century, modern science has changed the world beyond recognition, and overwhelmingly for the better.

But success can breed complacency. Modern scientists are doing too much trusting and not enough verifying—to the detriment of the whole of science, and of humanity.

The Economist | World politics | Business & finance | Economics | Science & technology | Culture

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

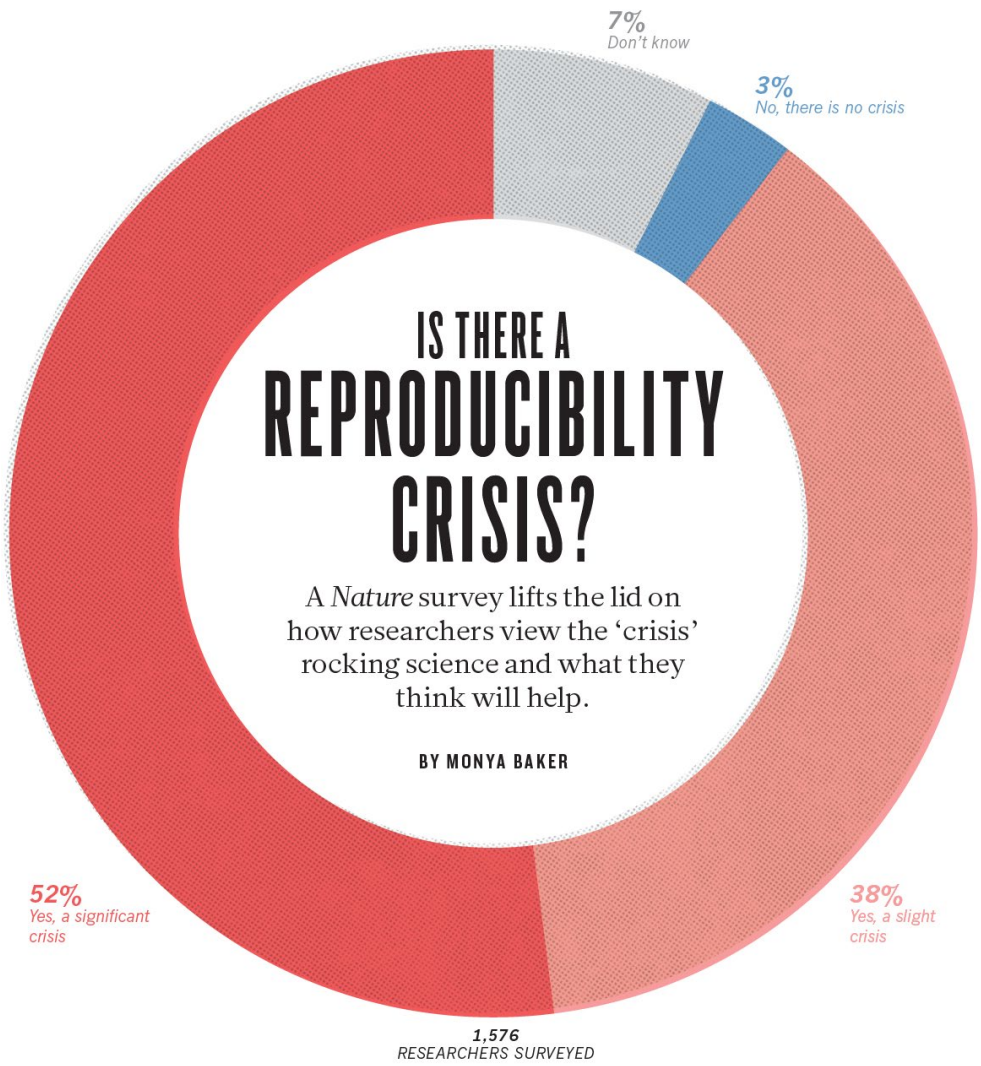
Oct 19th 2013 | From the print edition

Like 11k Tweet 1,227

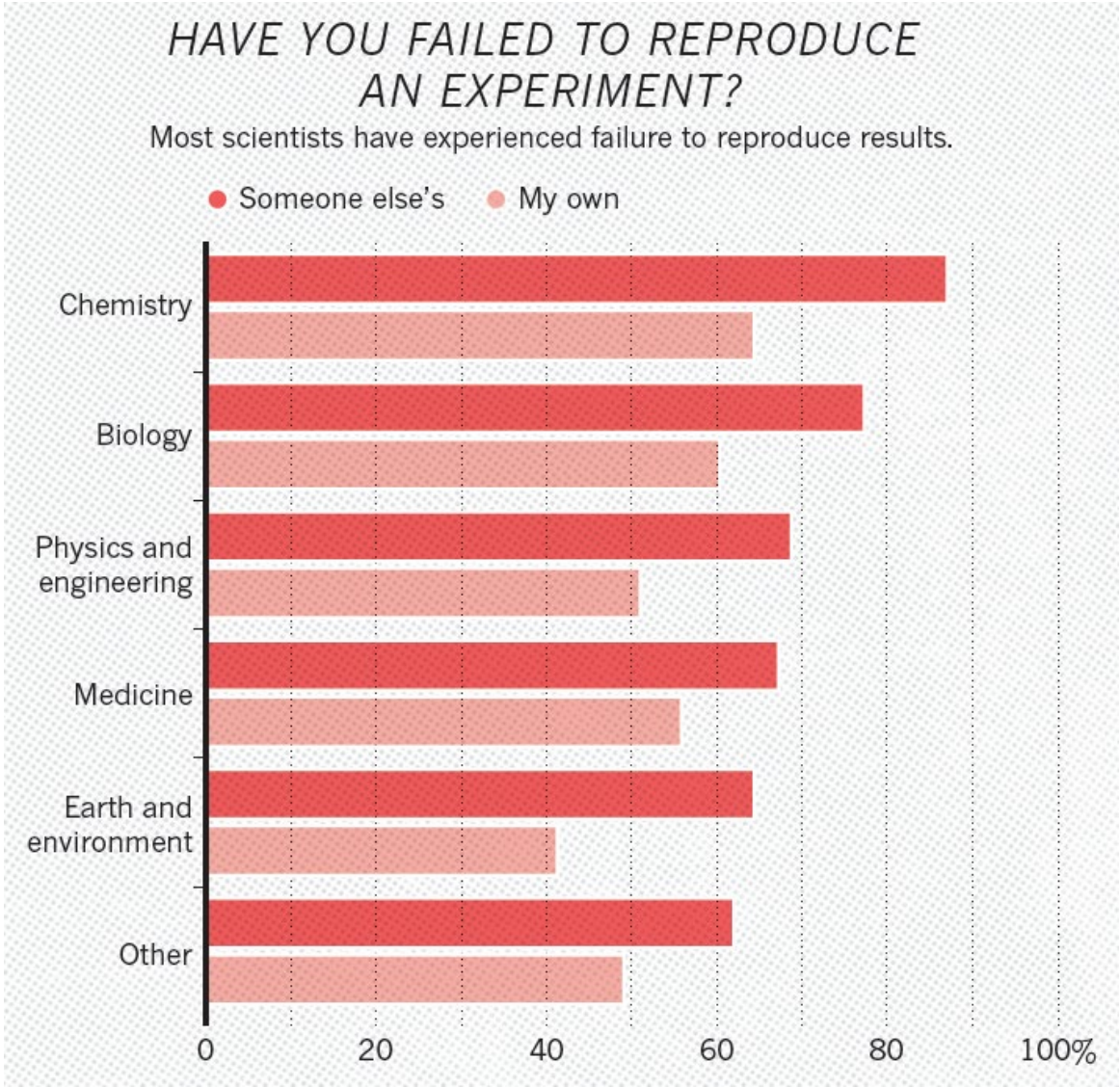
"I SEE a train wreck looming," warned Daniel Kahneman, an eminent psychologist, in an open letter last year. The premonition concerned research on a phenomenon known as "priming". Priming studies suggest that decisions can be influenced by apparently irrelevant actions or events that took place just before the cusp of choice. They have been a boom area in psychology over the past decade, and some of their insights have already made it out of the lab and into the toolkits of policy wonks keen on "nudging" the populace.

Dr Kahneman and a growing number of his colleagues fear that a lot of this priming

Baker (2016) in *Nature*



Baker (2016) in *Nature*



The Crisis in Medical Research

How could Ioannidis get these results?

There are many things that can influence validity of research findings: Selection issues, multiplicity issues, *p*-hacking

We have begun to realize that to think about scientific process in terms of ***ensembles of studies*** (Fisher would not have been surprised, he emphasized robust replication in scientific process)

NIH, NSF, and increasingly, IES, are worried about this

So is the rest of the scientific community

We must learn from medicine (which is ahead of us on this)

NIH *is* Taking the Problem Seriously: (from *Nature*, 2014, 505 1/14/2014)

NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

A growing chorus of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring[1,2]. As leaders of the US National Institutes of

shorter term, however, the checks and balances that once ensured scientific fidelity have been hobbled. This has compromised the ability of today's researchers to reproduce others' findings.

Let's be clear: with rare exceptions, we

Selection Issues

Outcomes of research studies are uncertain (that is why we use statistical inference to understand them)

An observed result may be bigger or smaller than the true effect of a treatment

If a study finds a particularly big effect (notable, likely to be highly cited) it may be partly due to chance: A lucky break—it is bigger than the true effect

A replications is not likely to be as lucky, hence it finds a smaller effect (maybe even not statistically significant)

Multiplicity Issues

Expensive trials measure multiple outcomes (or look at multiple subgroups)

It makes sense to look at all of them

There is a tendency to report what is significant, as if it is the only outcome

If you multiple significance tests at the 5% level, the chance that at least one (of several) comes out significant is greater than 5% —it may be *much* higher than 5%

There are ways to adjust significance levels, but it is tricky

Less obvious is that multiplicity has effects on size of effects we estimate too

p-Hacking

Torture the data until it confesses

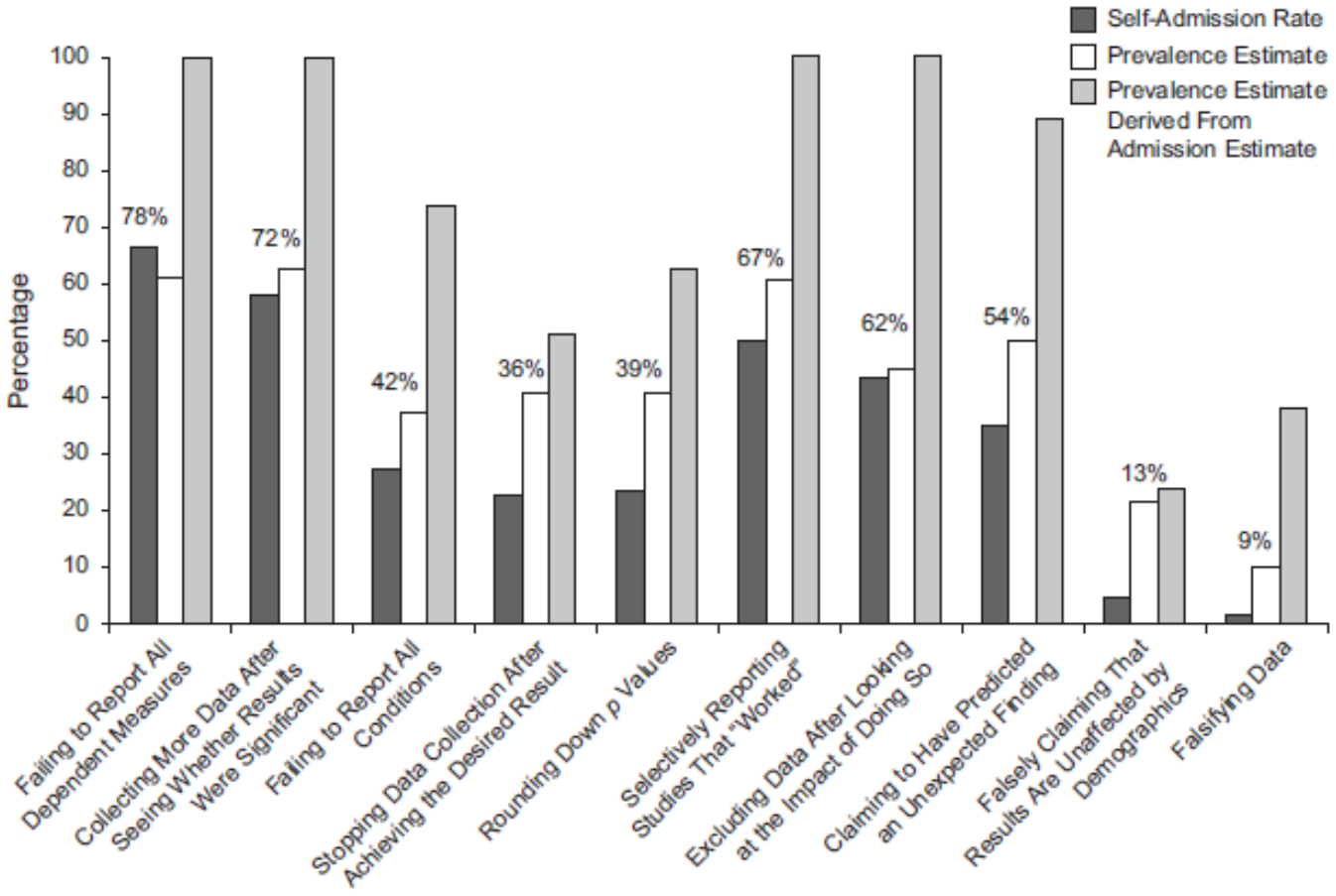
Otherwise known as *p*-hacking

Lots of different analyses are possible and they give slightly (or substantially) different answers

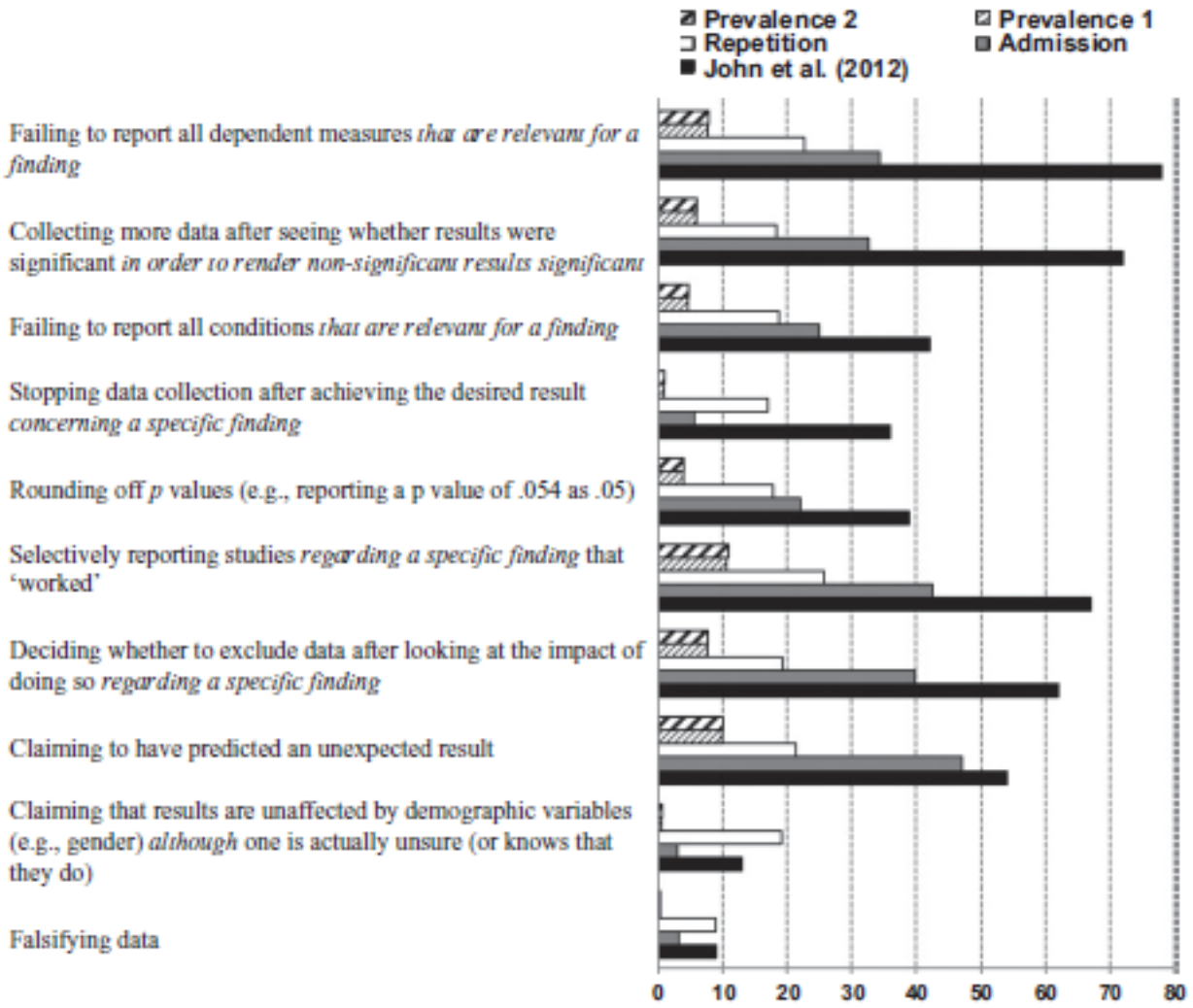
It makes sense to try lots of analyses, but picking the one you like *because* it gives the answer you want is not an appropriate way to select an analysis

Of course it is hard to distinguish (even for yourself) what is a sensible modification of an analysis plan and an opportunistic one

Rates of Questionable Research Practices (from John, Loewenstein, & Prelec, 2012)



Rates of Questionable Research Practices (from Fiedler & Schwarz, 2015)



The Problem

All of these research practices make the scientific literature unreliable

They undermine the credibility of scientific practice and the scientific knowledge base

This is serious!

Approaches to Solutions

Constraints on scientific practice

Protocols for studies that define methods to be used

Public ***Registration*** of protocols in advance

Reporting Standards that ensure transparency of research practice

All of these ideas are new in education, but increasingly important in medicine

We will borrow from the SPIRIT guidelines for protocols and the CONSORT standards for reporting trials

CONSORT and SPIRIT

Both CONSORT and SPIRIT have multiple parts

A checklist or guideline and flow diagram

One or more “explanations and elaborations” documents that illustrate the use of the checklist or guideline

The explanations and elaborations documents are intended to be used along with the reporting guidelines to flesh out and illustrate concepts

Protocols

Protocols

A protocol is a detailed description of the methods that will be used in a study

It is intended for public scrutiny

It is intended to be written before the study is undertaken

It may be modified, but the modifications and the reasons for them should be part of amendments to the protocol that are also available for public scrutiny

The idea is to constrain the methods that are used in eventual research publications from trials and make them transparent

SPIRIT

The SPIRIT movement (like the CONSORT movement) originated in medicine but has reached beyond it

SPIRIT stands for **S**tandard **P**rotocol **I**tems: **R**ecommendations for **I**ntervention **T**rials

They also have an extensive website

http://www.spirit-statement.org/

The items in the SPIRIT guidelines generally make sense for education and provide guidance about what to put in a protocol that is registered

SPIRIT

The SPIRIT statement has several pieces

A checklist giving key features to be reported

A figure to facilitate reporting

An explanation and elaborations document illustrating the use of the SPIRIT checklist

You may think that a 35 item checklist and a figure are silly—but this kind of standardization can dramatically improve the transparency of reporting both protocols and results

Registration

Registration

Registration is the process of “publishing” the protocol so that it open to public scrutiny

Protocols are published in special archives called registries (e.g., clinicaltrials.gov)

WWC currently has a kind of registry, which is being phased out

The Society for Research on Educational Effectiveness (SREE), with the support of IES, has developed a registry of efficacy and effectiveness studies (now housed at the University of Michigan’s ICPSR), see

https://sreereg.icpsr.umich.edu/sreereg/

Some day (hopefully soon), registration will be a *requirement* for publication (as it is in many medical journals that publish clinical trials)

How Did Medicine Get Scientists to Register Trials?

The NEW ENGLAND JOURNAL of MEDICINE

EDITORIALS

Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors

Altruism and trust lie at the heart of research on human subjects. Altruistic individuals volunteer for research because they trust that their participation will contribute to improved health for others and that researchers will minimize risks to participants. In return for the altruism and trust that make clinical research possible, the research enterprise has an obligation to conduct research ethically and to report it honestly. Honest reporting begins with revealing the existence of all clinical studies, even those that reflect unfavorably on a research sponsor's product.

Unfortunately, selective reporting of trials does occur, and it distorts the body of evidence available for clinical decision-making. Researchers (and journal editors) are generally most enthusiastic about the publication of trials that show either a large effect of a new treatment (positive trials) or equivalence of two approaches to treatment (non-inferiority trials). Researchers (and journals) typically are less excited about trials that show that a new treatment is inferior to standard treatment (negative trials) and even less interested in trials that are neither clearly positive nor clearly negative, since inconclusive trials are registered in a public repository at their inception, every trial's existence is part of the public record and the many stakeholders in clinical research can explore the full range of clinical evidence. We are far from this ideal at present, since trial registration is largely voluntary, registry data sets and public access to them varies, and registries contain only a small proportion of trials. In this editorial, published simultaneously in all member journals, the International Committee of Medical Journal Editors (ICMJE) proposes comprehensive trials registration as a solution to the problem of selective awareness and announces that all eleven ICMJE member journals will adopt a trials-registration policy to promote this goal.

The ICMJE member journals will require, as a condition of consideration for publication, registration in a public trials registry. Trials must register at or before the onset of patient enrollment. This policy applies to any clinical trial starting enrollment after July 1, 2005. For trials that began enrollment

Catherine De Angelis, M.D., M.P.H.
Editor-in-Chief, *JAMA*

Jeffrey M. Drazen, M.D.
Editor-in-Chief, *New England Journal of Medicine*

Prof. Frank A. Frizelle, M.B., Ch.B., M.Med.Sc., F.R.A.C.S.
Editor, *The New Zealand Medical Journal*

Charlotte Haug, M.D., Ph.D., M.Sc.
Editor-in-Chief, *Norwegian Medical Journal*

John Hoey, M.D.
Editor, *CMAJ*

Richard Horton, F.R.C.P.
Editor, *The Lancet*

Sheldon Kotzin, M.L.S.
Executive Editor, MEDLINE
National Library of Medicine

Christine Laine, M.D., M.P.H.
Senior Deputy Editor, *Annals of Internal Medicine*

Ana Marusic, M.D., Ph.D.
Editor, *Croatian Medical Journal*

A. John P.M. Overbeke, M.D., Ph.D.
Executive Editor, *Nederlands Tijdschrift voor Geneeskunde (Dutch Journal of Medicine)*

Torben V. Schroeder, M.D., D.M.Sc.
Editor, *Journal of the Danish Medical Association*

Hal C. Sox, M.D.
Editor, *Annals of Internal Medicine*

Martin B. Van Der Weyden, M.D.
Editor, *The Medical Journal of Australia*

Reporting Standards

CONSORT

The CONSORT movement originated in medicine but has reached far beyond it

Consort stands for **Co**nsolidated **S**tandards **o**f **R**eporting **T**rials

They have an extensive website
http://www.consort-statement.org/

CONSORT started with standards for reporting individually randomized trials but have extended their standards to include cluster randomized trials and social and policy interventions (about to be released)

CONSORT Standards

Note that CONSORT has many materials:

A checklist for Abstracts

A general checklist for randomized trials

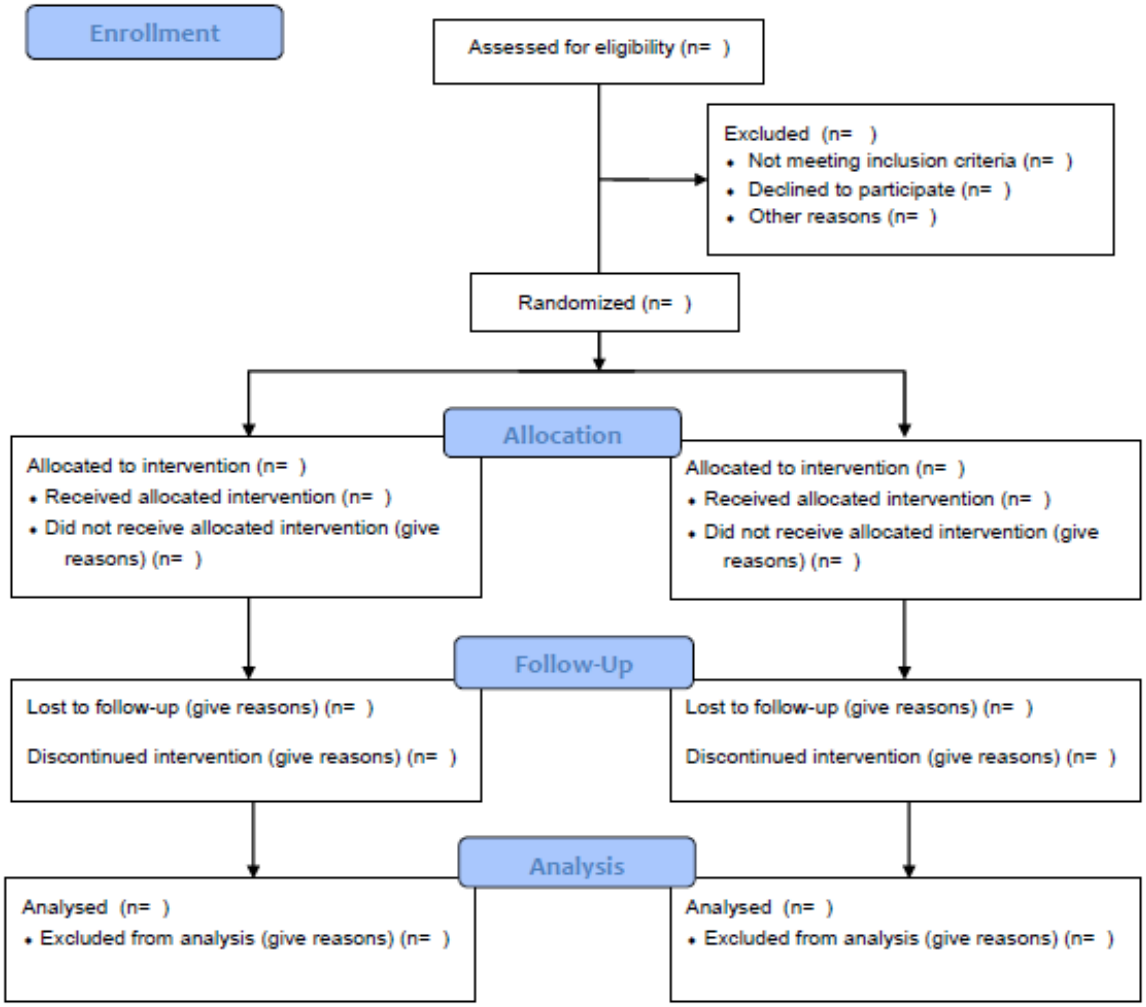
A flow diagram to help with reporting

An extension for cluster randomized trials

An extension for social and policy interventions (CONSORT SPI)

Elaboration and explanations documents explaining how to use the checklists

The Original CONSORT Flow Diagram



The CONSORT SPI Flow Diagram

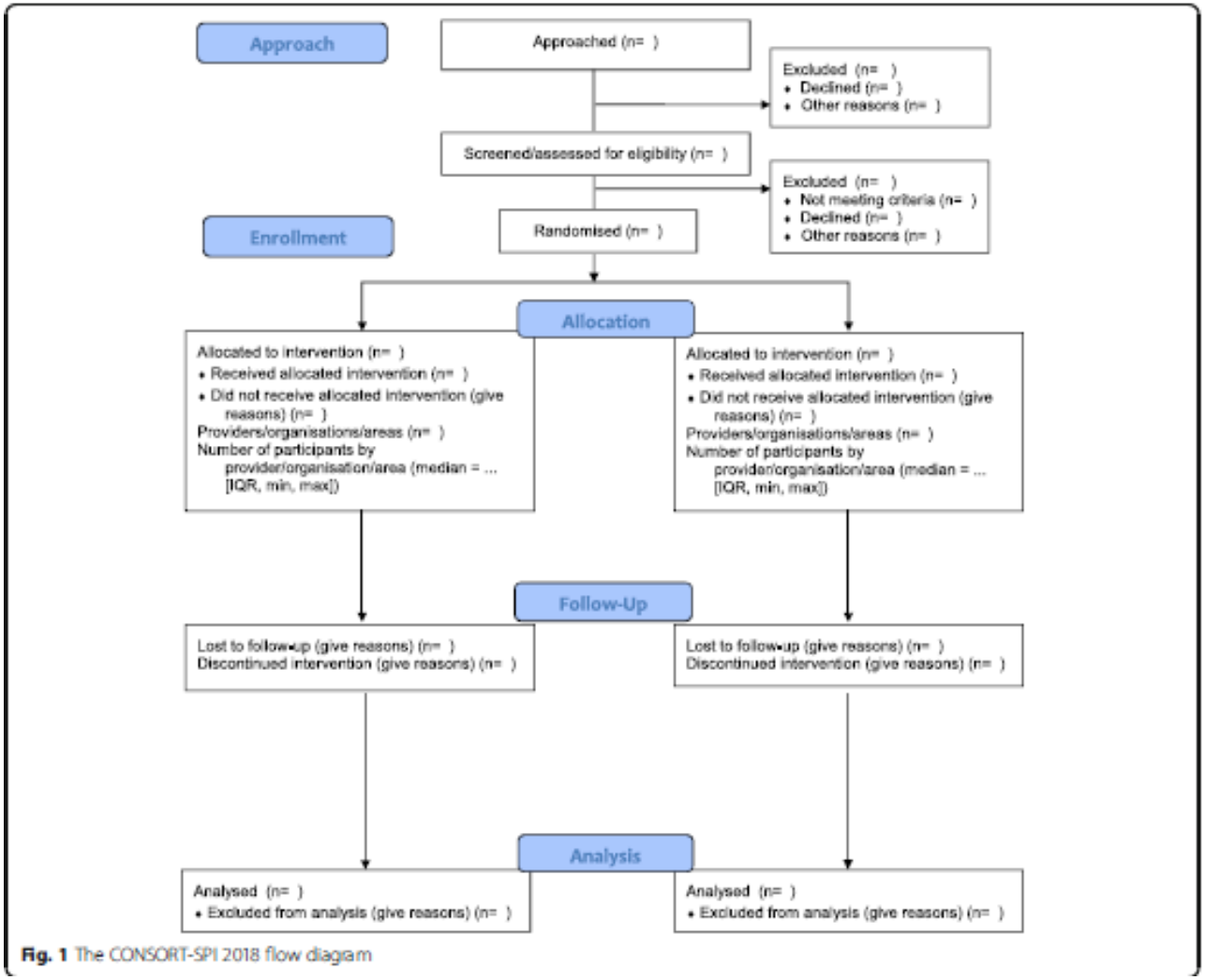


Fig. 1 The CONSORT-SPI 2018 flow diagram

The CONSORT SPI Checklist

Section	Item #	CONSORT 2010	CONSORT-SPI 2018
Title and abstract	1a	Identification as a randomised trial in the title[§]	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance, see CONSORT for Abstracts)[§]	Refer to CONSORT extension for social and psychological intervention trial abstracts
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale[§]	
	2b	Specific objectives or hypotheses[§]	If pre-specified, how the intervention was hypothesised to work
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio[§]	If the unit of random assignment is not the individual, please refer to CONSORT for Cluster Randomised Trials [8]
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	
Participants	4a	Eligibility criteria for participants[§]	When applicable, eligibility criteria for settings and those delivering the interventions
	4b	Settings and locations where the data were collected	
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered[§]	
	5a		Extent to which interventions were actually delivered by providers and taken up by participants as planned
	5b		Where other informational materials about delivering the intervention can be accessed
	5c		When applicable, how intervention providers were assigned to each group
Outcomes	6a	Completely defined pre-specified outcomes, including how and when they were assessed[§]	
	6b	Any changes to trial outcomes after the trial commenced, with reasons	
Sample size	7a	How sample size was determined[§]	
	7b	When applicable, explanation of any interim analyses and stopping guidelines	
Randomisation			
Sequence generation	8a	Method used to generate the random allocation sequence	
	8b	Type of randomisation and details of any restriction (such as blocking and block size)[§]	
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence, describing any steps taken to conceal the sequence until interventions were assigned[§]	
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions[§]	
Awareness of assignment	11a	Who was aware of intervention assignment after allocation (for example, participants, providers, those assessing outcomes), and how any masking was done	
	11b	If relevant, description of the similarity of interventions	
Analytical methods	12a	Statistical methods used to compare group outcomes[§]	How missing data were handled, with details of any imputation method
	12b	Methods for additional analyses, such as subgroup analyses, adjusted analyses, and process evaluations	

The CONSORT SPI Checklist (continued)

Section	Item #	CONSORT 2010	CONSORT-SPI 2018
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers randomly assigned, receiving the intended intervention, and analysed for the outcomes[§]	Where possible, the number approached, screened, and eligible prior to random assignment, with reasons for non-enrolment
	13b	For each group, losses and exclusions after randomisation, together with reasons[§]	
Recruitment	14a	Dates defining the periods of recruitment and follow-up	
	14b	Why the trial ended or was stopped	
Baseline data	15	A table showing baseline characteristics for each group[§]	Include socioeconomic variables where applicable
Numbers analysed	16	For each group, number included in each analysis and whether the analysis was by original assigned groups[§]	
Outcomes and estimation	17a	For each outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)[§]	Indicate availability of trial data
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses, adjusted analyses, and process evaluations, distinguishing pre-specified from exploratory	
Harms	19	All important harms or unintended effects in each group (for specific guidance, see CONSORT for Harms)	
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	
Generalisability	21	Generalisability (external validity, applicability) of the trial findings[§]	
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	
Important information			
Registration	23	Registration number and name of trial registry	
Protocol	24	Where the full trial protocol can be accessed, if available	
Declaration of interests	25	Sources of funding and other support, role of funders	Declaration of any other potential interests
Stakeholder involvement	26a		Any involvement of the intervention developer in the design, conduct, analysis, or reporting of the trial
	26b		Other stakeholder involvement in trial design, conduct, or analyses
	26c		Incentives offered as part of the trial

The Future of Our Science

Registration of protocols for studies intended to provide causal evidence is essential to ensure the validity of our science

(It is even more important for quasi-experiments, but expect resistance there)

Reporting guidelines that we generally adhere to are also crucial

You can be part of assuring the future of education science by being an early adopter of these methodological innovations

Likelihood of Null Results in Medicine has Increased Since 2005

PLOS | ONE

RESEARCH ARTICLE

Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time

Robert M. Kaplan[1]*, Veronica L. Irvin[2]

1 Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services, Rockville, Maryland, United States of America, **2** Oregon State University, Corvallis, Oregon, United States of America

* Robert.Kaplan@ahrq.hhs.gov