Preregistration: the good, the bad, and the confusing

Alain de Cheveigné

v230214

Preregistration: the good, the bad, and the confusing

Alain de Cheveigné, CNRS, ENS/PSL, UCL, alain.de.cheveigne@ens.fr

Abstract

Preregistration is a tool to enhance the reliability of science. I argue that: (a) preregistration addresses an important need, (b) it offers considerable benefits, (c) those benefits partially cover the need, (d) they are accompanied by costs and side effects. The decision to make preregistration a *normative* requirement should be carefully assessed for its potential side-effects, and alternative models and norms should be considered. I discuss factors that affect the reliability of science and how preregistration can influence them, and I make a few suggestions to enhance its efficacy while limiting its risks.

key words: preregistration, replicability, transparency, statistics, p-values

word count: 9700

Introduction

A while back I overheard a discussion. A student and a colleague had found an issue with the design of their experiment that could be addressed by adding a control, however that was not an option, they said, because *the study had been preregistered*. This, I thought, cannot be right. Conversations at the coffee machine or around lunch, once about science, cool apps or gossip, were increasingly devoted to preregistration, its gifts, and its burdens. Among the gifts is a heightened awareness of issues of statistical reliability, hypothesis tests and how they go wrong, human foibles and flaws, and so on. Among burdens is the time and effort discussing the tradeoffs of various schemes, and whether the whole thing is a good idea or not. This article counts among the 'burdens' for you, the reader, and I, the writer, as our time could be spent otherwise, but hopefully it comes with some gifts too. It is an attempt to steer a course through the choppy seas around preregistration.

The ‘replicability crisis’ has attracted lots of attention, but the real concern is *credibility*, for which replicability (or lack thereof) is an indicator. It is a concern for the *individual* who answered the call of science over a different career, for a *field* trying to defend its status (and funding) relative to other disciplines, and for *science* trying to defend its status (and funding) within society. It is a concern for *society* because science plays an essential role in enabling progress and addressing crises. Loss of credit attributed to science diminishes society as a whole.

Science is designed to maximize the reliability of knowledge. For that it draws on logical, mathematical, empirical and statistical tools, established practices of communication, peer review, and an ethos of subjecting past results to future scrutiny (Popper 1935). These are widely assumed to 'weed out' unreliable results, in a form of self-correcting mechanism that maintains the validity of science as a whole. However, certain of these tools can be weakened by flawed behavior, and the efficacy of self-correction has been questioned (Ioannidis 2012, Pashler and Harris 2012, Vazire 2021, Stewart and Plokin 2021).

An important tool is *replication* in which a study is repeated in similar conditions, or with a similar goal. Failure to replicate casts doubt on the generality of the original results, whereas success might bolster them (Nosek and Errington 2020). This, one hopes, ensures that flawed results are gradually detected, flagged, and eliminated from subsequent consideration. However, there are weak or negative incentives to publish the negative results of a failed replication, and instead strong incentives to output work that is 'original' (hence not a replication), at a sustained rate and with little subsequent scrutiny, and a premium for more surprising (hence more likely wrong) results (Nosek et al 2012). Non-replicable results were found to be *more* likely to be cited, even after a replication failure had been published (Serra-Garcia and Gneezy 2021), and their influence may linger after they are retracted (Berenbaum 2021). According to a simple model, reluctance to publish negative results is guaranteed to canonicize false facts (Nissen et al 2016).

Attempts to evaluate empirically the level of replicability within a field have produced sobering results (Open Science Collaboration 2015, Minocher et al 2021, Chang and Li 2022), and in some domains there are reports of fraud on an industrial scale (Holly and Van Noorden 2021). Even in the absence of fraud or conscious manipulation, 'researcher degrees

of freedom' can greatly increase the probability that an inexistent effect tests significant (Simmons et al 2011, Gelman and Loken 2013).

This motivates the drive (itself form of a 'self-correction') to promote new forms of scientific practice to improve reliability (Nelson et al 2018). These practices follow two approaches: one operates *upstream* before publication, the other relies on *downstream* control via critical appraisal (Gelman 2017, Nosek and Errington 2020, Vazire 2021). The former includes preregistration and registered reports, the latter includes full disclosure of data and analysis scripts, and better documentation so that details of the analysis can be scrutinized and reproduced. Both require changes in publication culture, which may entail various forms of 'nudging', including guidelines, standards, badges, and so-on, to overcome inertia and contrary incentives. Each of these changes (and the nudging itself) has costs as well as benefits, and that is what this paper is about.

My colleague and his student were perhaps wrong to hesitate – we know that "*preregistration is not a prison*" (Dehaven 2017). Perhaps it was OK to choose a pristine design and slightly flawed preregistration, over a pristine preregistration but slightly flawed design? Whatever the answer, they were checked in their momentum, unsure what to do. If this had been an opportunity to think deeply, and glimpse some truth about science, epistemology, or statistics, all would be well. Instead, they were probably just wondering how to get the paper past the reviewers. Is it *really* safe to void the preregistration, or might the reviewers be fastidious about it? This is what I call the "*cost of confusion*" (see below).

Preregistration is motivated by an important need

The situation is dire. Published results cannot be trusted because, when challenged, they fail to replicate or reproduce (Artner 2021, Chang and Li 2021), because data or experimental details needed for scrutiny are lacking (Wicherts et al 2016, Miyakawa 2020), because the power, significance threshold and sample size imply that significant results are too common to all be true (Cohen 1962, Ioannidis 2005, Bertamini and Munafò 2012, Button et al 2013), or because of known effects of publication bias or incentives (Gelman and Loken 2013, Higginson and Munafò 2016, Smaldino and McElreath 2016, Edwards and Roy 2017), or various shades of fraud (John et al 2011, Holly and Van Noorden 2021). This motivates the diagnostic of *crisis* (Gelman and Vazire 2021), although its reality or severity has also been questioned (Gilbert et al 2016, Shiffrin et al 2018, Redish 2018).

A simple sketch of the scientific process is that theory generates predictions that empirical tests might find wrong (falsify) – or not. A theory that makes incorrect predictions, or is unable to make predictions, should be discarded (Popper 1935), or downgraded relative to competing theories (Lakatos, Stanford Encyclopedia of Philosophy 2021). Critical to this process is the empirical test that seals the fate of the theory. It must be trustworthy, and this trust must be reliably communicated to, and evaluated by, the community. Box (1976) described this process as a loop that iterates between theory-driven experimentation and evidence-driven theory update, progressively refining the theory (Fig. 1).

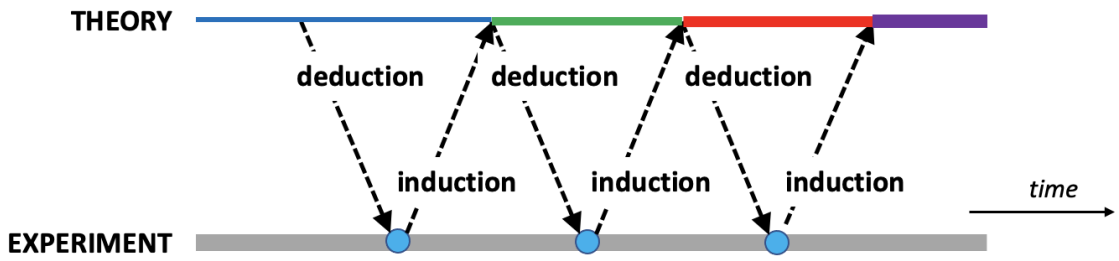

Figure 1. Box's iterative loop between theory and experiment, unrolled over time. After each new experiment (blue dot) the theory is updated and the new theory's predictions are worked out so they can be empirically challenged once more (next blue dot), gradually refining our theoretical understanding (adapted from Box 1976).

This neat sketch is upset by the inevitable noisiness of observations. An experiment might require averaging over a sample of observations to reduce the noise, and the outcome might still be noisy. A prediction might wrongly fail to materialize (false negative), or else be wrongly corroborated (false positive) (Amrheim 2019). Worse, an interesting pattern might suddenly appear, purely by chance, prodding us to make unwarranted changes to the theory. Occasionally one does see a rabbit in the clouds, or a teddy bear. *Statistical hypothesis testing* is designed to protect us from this risk, by sorting – however imperfectly – the chance rabbits from the real.

The ubiquitous *null-hypothesis significance test* (NHST) estimates the probability that an observation occurred by chance in the absence of any effect. This probability (the 'p-value') is compared to a conventional threshold (e.g. 0.05, or 0.005), and the observation is taken seriously only if the p-value is smaller. In terms of Fig. 1, only such a 'statistically significant' observation should trigger an upward arrow, thus providing a form of stability to Box's iterative loop. The process is far from perfect, and has been criticized roundly on many accounts. However, we have come to rely on it as part of the machinery that we use to decide

how much credit to attribute to empirical observations (ours or those of others): the ‘p-value’ is how we communicate – however imperfectly – this trust to the readers of our work. It is this process that seems to be in crisis: the process by which we evaluate and communicate trust cannot itself be trusted.

There are several ways a significance test can go wrong. The best known is ‘multiple testing’, akin to rolling the dice several times and taking the best outcome, but there are others less obvious. For example, adding a few more observations after the negative outcome of a significance test on an initial set of observations can vastly inflate the risk of a false positive (Fig. 2). This is unnerving, because it seems a very natural thing to do. Isn’t it reasonable to make a few more observations rather than discard what we've observed so far and lose the investment made so far? More observations imply more information, no? Indeed, but we lose the ability to test whether that information is statistically reliable... There are many such mistakes to be made, some deliberate (‘p-hacking’), others innocent (‘the garden of forking paths’, Gelman and Loken 2013, Nuzzo 2015).

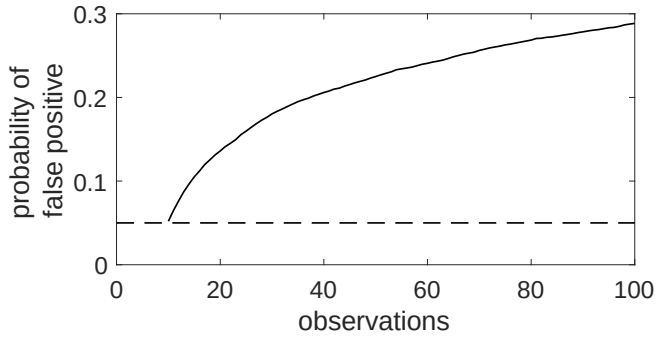

Figure 2. Probability of a false positive for a significance test with threshold p=0.05 if additional observations are made after an initial 10 observations. After each new observation the test is repeated until it yields p<0.05, at which point the process is stopped. The false positive rate (full line) soars well above the nominal threshold (dashed

line). A similar effect plagues 'early stopping': if the initial plan is to gather 100 observations, but we allow ourselves to stop as soon as the test is significant after at least 10 observations, the false positive rate approaches p=0.3 (rightmost point on the full line). This simulation is based on a Gaussian process with zero mean, 10000 repeats. Lakens (2021) gives pointers to valid approaches to early stopping.

Uncontrolled, these *researcher degrees of freedom* undermine the trust we should otherwise lend to published results based on the statistical tests that they report. If these degrees of freedom were brought under control, a major factor of unreliability (and thus non-replicability) would be removed. This is a prime motivation for preregistration, as part of an ensemble of measures to enhance reliability of results in the literature (Wichert et al 2016, Nosek et al 2018).

Preregistration is beneficial

The benefits of preregistration have been laid out in many publications (Wicherts et al 2016, Davis et al 2018, Nelson et al 2018, Nosek et al 2012, 2018, 2019).

The primary benefit is to reduce the researcher degrees of freedom that invalidate hypothesis testing. If all steps of data collection and analysis are specified in advance, the investigator cannot adjust them. This preserves the validity of statistical tests applied to the data, removing a major factor of unreliability of published results. Replication failure becomes less likely because the rate of 'false positives' among original findings is reduced, although a replication might still fail due to the stochastic nature of observations, at a rate possibly inflated if multiple studies and teams probe the same hypothesis.

A secondary benefit is that preregistration offers the opportunity to lay out the *theoretical* assumptions to be tested, and ensure that they are not subject to an opportunistic adjustment to empirical outcomes (e.g. HARKing, hypothesizing after results are known, Kerr 1998). Taking an example from cognitive neuroscience, the hypothesis of predictive coding is consistent with both an *increase* in brain activity (reflecting the process of prediction) or a *decrease* of brain activity (reflecting a smaller prediction error). Either outcome is consistent with the theory, so the experiment does not actually test it.

A third benefit is more diffuse. Preregistration slows the pace, encouraging the investigator to think more carefully about hypotheses, assumptions, and methodological details. To formulate a plan, each step must be considered, which may be beneficial whether or not the plan is actually followed during execution. To paraphrase General Dwight D. Eisenhower, "*plans are worthless, but planning is of essence*".

The benefits partially cover the need

Preregistration addresses one of several things that can go wrong with a study. In addition to invalid statistical tests, inadvertent or intentional errors in bookkeeping, analysis, or reporting can occur whether or not the study is preregistered. A sample may not be representative of the heterogenous population (Bryan et al 2021). Data can be faked (Felgenhauer 2021). The preregistered plan may be incomplete, or obscure, or it may not be followed exactly (Bakker et al 2020, Ofosu et al 2021, Claesen et al 2021). Deviations may or may not be reported by the authors, or spotted by the reviewers or editor. Analogous to HARKing, PARKing (preregistration after results are known) may invalidate the exercise (Yamada 2018). If the pursuit of reliability is likened to a game of 'whack-a-mole', preregistration whacks one mole.

A motivation for preregistration is to counter the "file drawer bias" that results from selective publication of successful but not failed experiments. Unfortunately, it only partly fills this role, as the fate of a preregistered but unpublished experiment is undetermined. Did the experiment fail, or did it succeed but was not published, or was it not undertaken? Failure is not explicitly documented.

Preregistration is mainly applicable to *confirmatory*, and less useful for *exploratory* research (although the pertinence of this distinction has been questioned, Rubin 2020, Szollosi and Donkin 2021). Arguably, exploratory results too need to be reliable, given that a false lead might trigger a wild goose chase at the expense of more fruitful investigations. Preregistration, like hypothesis testing, takes the focus away from the theory and evidence themselves and diverts it towards *indicators* of their validity and robustness, rules to guarantee the validity of those indicators, and norms to enforce those rules.

The benefits come with costs

Preregistration entails costs for the investigator, the reviewer/editor, and the reader. These may be associated with the process (e.g. time required to preregister or follow up), or with the 'nudging' deployed to get researchers or publishers to change their practices (norms, guidelines, transparency ratings, badges, etc.). Benefits and costs should be weighed together to guide our choice of which preregistration model to choose, if any.

Preregistration takes time and attention. An initial cost is to grasp the concept, select a scheme, figure out the requirements (witness the numerous workshops and tutorials devoted to the question), and fill in the required fields. Some schemes, such as the Registered Report (Chambers 2013, Grand et al. 2018) entail a review process during which the project is

paused and its future uncertain. The investigator might curtail his/her objectives to make the plan easier to write or more likely to be accepted, resulting in a possibly less fruitful outcome.

There is a cost also for the reader. Conventionally, the bolus of information transfer is the scientific paper, prepared with care by the investigator, reviewed critically, refined iteratively, and published for the record and the benefit of readers. The unity of place (the article) and time (the submission/review/publication cycle) is somewhat fragmented by preregistration. The reader might need to access both preregistration and report, especially if the author decides (reasonably) to omit from the paper methodological details provided in the preregistration.

There is a cost for the reviewer and editor if tasked with checking that preregistration and report match (Davis et al 2018). It is unclear what to do if there is a mismatch. If the preregistration is obscure (c. f. Bakker et al 2020), should it be rewritten to serve the interests of the reader, at the cost of voiding the immutability of preregistration? As a reviewer, I don't want to release a poorly written paper, nor get into an argument with an author who feels entitled to publication. Review/editorial manpower is a scarce and precious resource.

An element of each of these costs is uncertainty. This is a paradoxical effect of flexibility of the norms, as the investigator must not only choose which standards to adopt, but also guess which standards he/she will be held to by the reviewers of a paper or grant proposal.

Alternatives

Preregistration intervenes upstream, in the early phases of a study. Like Ulysses tying himself to a mast to resist the call of the Sirens, the investigator preregisters to avoid bamboozling his/herself, or us.

A downstream alternative is *replication*, the process by which experiments or observations are repeated by the same investigator or others (National Academy of Sciences 2019). Chance factors that tipped the balance in the original study are unlikely to benefit the replication. Moreover, the prospect of replication failure, if sufficiently prominent, may suffice to 'keep the investigator honest'. Also downstream is *scrutiny*, which involves detective work by other researchers to reveal flaws within a study. Making data and analysis scripts publicly available to aid scrutiny are an important part of 'Transparency and Openness Promotion' (Nosek 2012, Davis et al 2018, Aczel et al 2020, Nelson et al 2021).

Thus, responsibilities shouldered by preregistration can in part be offloaded in part to other measures, with a different tradeoff between costs and benefits. Whether doing so is desirable is a complex question that depends on how costs and benefits are weighted (Rubin 2020).

In summary, the pros and cons of preregistration have been pointed out by many authors. The rest of this paper attempts to dig a bit deeper.

What is the goal?

A symptom, not a malady

Replicability Crisis is a bit of a misnomer. The real concern is credibility, or more bluntly, the innuendo that "*most published research findings are false*" (Ioannides 2005). A focus on

replicability per se diverts attention towards different issues. Isn't replication failure inevitable given the vagaries of statistically variable effects (Ulrich and Miller 2020, Miller and Ulrich 2022)? Should we favor exact (or direct) replication that challenges data robustness, or conceptual replication that challenges theoretical implications (Nosek and Errington 2020)? How is this related to Popperian falsification (Fidler et al 2018)? These questions are interesting, but peripheral to the core issue of credibility. Replication failure is a canary in the coal mine: the main concern is credibility, and thus "*replicability is not an aim that needs to be independently pursued"* (Szollosi and Donkin 2021). Replicability is however a prime metric by which we measure credibility, and 'non-replicable' is often used as polite way of saying 'flawed'.

Credibility of theory or data?

When writing a paper, I cite two categories of studies: those that propose *ideas* (possibly supported by data) useful to inspire and elaborate a new idea, and those that report *data* (possibly reported together with a theory) useful to argue for - or against - the plausibility of the new idea. My main concern is the reliability of those data: whether or not the theory that accompanies them is trustworthy is less important. There is a need for theory-independent measures of credibility. In this context, the p-value is attractive as a theory-free measure of data quality, function only of the values observed in an experiment.

A test of the report 'manipulation ***A*** results in effect ***a***' helps me decide whether this report is reliable. In Bayesian terms, the *evidence* is more important than the *posterior probability* of whatever hypothesis was reported together with that evidence. This is in contrast with the view that the hypothesis testing should focus on theoretical hypotheses (Meehl 1990, Ashton 2013, Proulx et al 2021, Stewart and Plotkin 2021, Szollosi et al 2021). It is not to suggest

lack of interest in theoretical positions, but rather that it is easier to aggregate evidence across studies, if reliable. That said, the story (theory) offered together with the evidence is useful to advertise its potential relevance (Stewart and Plotkin 2021), and it might serve also as a hint to its plausibility. For example, if a paper claims support for extrasensory perception, which I find implausible, I might be suspicious of the evidence it reports. This is slippery terrain, as it may prevent us from taking seriously new and unexpected phenomena (Jaynes 1979, chapter 5).

I suspect that the literature contains more evidence than theory. Countless empirical studies have tested the theory according to which the brain employs predictive coding (Friston 2018), usually each with a handful of alternative theories to allow model selection. Theory is the tip of the iceberg, the rest consists of data, and at issue is the credibility of those data.

Science as Search

Box (1976) compared science to a loop that iterates between theory and evidence gathering, progressively refining the theory (Fig. 1). This can be likened to a search within the high-dimensional space of plausible theories and relevant evidence. From Computer Science we know that two factors are critical for efficient search: the speed with which each item can be inspected before discarding it, and the ability to *prune* the search space, or at least prioritize different parts of that space. Search is slow if each inspection takes too long, and infeasible in a large space unless guided by pruning or prioritizing. The outcome of a hypothesis test (p-value) can be understood as an aid for search: non-significant effects are ignored, i.e. pruned. The aim is not so much to assess ‘truth’ or ‘confidence’ as to avoid time wasted on unpromising parts of the search space.

The binary nature of a hypothesis test has been lamented (Greenland et al 2016, Held and Ott, 2018, Amrhein et al 2019, Brent et al 2020) because it offers just one bit of information. However, the same is true for a decision, forced on us by the need to act, but also by space, time, or cognitive limits. Our mind cannot entertain unlimited hypotheses so we need to *discard* some of them. We won't get around to reading every paper, or all of each paper. The hour gets late, attention wanes, and we give up half way through the stack of PDFs, effectively pruning the space of knowledge on which we base our thinking. A librarian clears shelves of books less likely to be requested, to make space, and sensory input is thinned by attentional processes before it reaches the brain.

The loss of information is offset by a benefit similar to regularization or dimensionality reduction in machine learning, akin to Occam's razor. Maintaining a stack of observations so that we can at some later point combine them, each with its own weight, is akin to maintaining a complex model, prone to overfitting. Discarding produces a simpler model. The call for 'complete reporting' of all experiments, successful or not (Amrhein et al 2019, van Assen et al 2014) ignores this need. From the point of view of a reader, that sounds like a terrible idea! We need to prioritize and prune, if not, information overload will do the pruning for us.

P-values or other inferential statistics simplify our papers and reduce time wasted on writing or reading about dubious results, and preregistration may help by making the pruning criteria crisper and more to the point. On the other hand, instead of reading or writing dubious papers, we might spend time reading or writing preregistration plans, and reading or writing or reviewing sections reporting how they (or we) did (or did not) adhere to them.

Certainty vs discovery

Before basing a costly experiment, a theory, or a career on a phenomenon, we would like to be *certain* that it is real. Initiatives to enhance that certainty (such as preregistration) seem of obvious benefit. On the other hand, Popper (1935) denies that that goal is achievable: every result is provisional, liable to be overturned or superseded. From that perspective, a guarantee of certainty is less useful than a dynamic process of discovery that can – possibly – show the result to be *untrue*. Factors that favor the former may hinder the latter. Lakatos (1976) made a similar remark in the context of mathematics: formal methods that attempt to guarantee truth may be sterile in terms of discovery and refinement of mathematical concepts. This is a fundamental tension: truth *must* be pursued with rigor, yet rules to *guarantee* truth may thwart discovery.

Alternatives

Replication

An alternative to upstream registration is downstream replication, but several obstacles hamper replication in this role. One is a reputation of dullness, that makes results hard to publish (Koole and Lakens 2012). Another is lack of clarity on whether we should insist on a 'direct' (or 'exact') replication, or also value 'conceptual' replications. A third is the vexing question of replicability of phenomena that are stochastic in nature, yielding false positives and false negatives at random. A fourth is the perception of replication as a hostile act, that may cause investigators to pussy-foot around a dubious study for fear of alienating its authors (Ioannides 2012).

A possible solution is offered by Nosek and Errington (2020) who define replication as "*a study for which any outcome would be considered diagnostic evidence about a claim from*

prior research". This casts replication as a refinement of the original study in the spirit of Box's loop (Fig. 1): success extends the domain of validity of the claim, failure more tightly marks its limits. New information is learned, making the replication *interesting* rather than dull, also sidestepping the debate as to whether 'direct' or 'conceptual' replication should be preferred. Exact replication, which tests the generality of the observation that *manipulation **A** results in effect **B*** made in the original study, cannot be attained: "*just as it is impossible to bathe in the same river twice, it is impossible to run the same study twice*" (Nelson et al 2018). Fortunately, thanks to Nosek and Errington's definition, *inexact* replication usefully tests for generality with respect to any mismatch between studies. Their definition also accommodates *conceptual* replication, by which a theoretical hypothesis is challenged by a deliberately different experiment, similar to the concept of 'triangulation' (Lawlor et al 2017). Direct and conceptual replication thus appear to belong to the same continuum, the one testing an empirical claim, the other a theoretical claim.

The definition goes some way to solve the conundrum of probabilistic success or failure, in the sense that results of original and replication can be combined statistically using bayesian techniques (Ly et al 2019, Held 2020, see also Simonsohn 2015, Stefan et al 2022). Pitching a failed replication as a *constraint on generality* of the original study softens the stigma but does not eliminate it completely: failure of one's results to replicate in other researcher's hands remains mildly embarrassing.

Also valuable are reports of *lack of effect* (Oldehinkel 2018). Replication of a high-profile study might be made a requirement to obtain a PhD (Frank and Saxe 2012). Conversely, onerous constraints on replication, such as that they must be of high power or submitted only as Registered Reports (Davis et al 2018, Wagenmakers and Forstmann 2014, Mayo-Wilson et

al 2021, TOP guidelines https://www.cos.io/initiatives/top-guidelines) are possibly less helpful. Replication is the go-to tool to enhance reliability: in doubt, "*science's most effective solution is to replicate, again*" (Nosek and Errington 2020).

Skepticism and severity

Discussions of the replication crisis take for granted that an investigator's goal is to *prove* hypotheses. For Popper or Lakatos, the goal should be to *disprove* them, the corroboration of a hypothesis being in proportion to the severity of tests deployed to disprove it (Popper 1935). Statistical tests contribute to this severity, questionable research practices finesse those tests, and preregistration helps avoid the finessing, thus restoring the severity (Lakens 2019).

Severity, here, is a property of the methodology. Interestingly, Szollosi and Donkin (2021) argue that the target of severity should be theory. "*A good theory designates both what we should have observed in the past and what we should observe in the future—there is no difference*" (hence no need to preregister). A good theory is "*hard to vary*", i.e. not easily adaptable to explain everything, and "*inflexible in a way that maximally allows for criticism*", which amounts to theoretical severity, rather than methodological.

I strive to *disprove* phenomena before I report them because I'd rather find a fault myself than let others find it. Preregistration helps by more effectively flagging statistically weak patterns. However, there are other ways of being wrong, such as artifacts, confounds, or bias, that may be recognized late after data are recorded. Analyses designed to reveal or control for them are ad-hoc, and the investigator may be unsure of their status and reluctant to carry them out if the study is preregistered. As an example, a classic study (Yuval-Greenberg and Deouell 2011) found that certain electroencephalograph (EEG) signals in a high frequency

band (gamma), hitherto attributed to cognition-related cortical activity, could be explained as an artifact of ocular microsaccades. That result challenged the interpretation of countless studies, and the research programs based on them, saving countless researcher-years of barking up the wrong tree. The frame of mind that led to it, skepticism, differs from that of studies bent on 'proving' effects.

In the extreme, hypothesis tests could have a paradoxical effect of shielding claims from skepticism: "*I did the stats, the effect was significant, why should you (or I) doubt me?*". This puts the onus on the doubter to explain how a significant p-value might have emerged. Few of us are skilled, tenacious and courageous enough to do so, particularly if details are lacking. Transparency makes checking easier, and preregistration closes loopholes. However, it adds a new task for the doubter (check that the report matches the preregistration), and a new shield for the transparency-badged doubtee: "*I preregistered, the effect was significant, why should you (or I) doubt me?*". That shield is even thicker for a registered report (*"The reviewers approved the report,..."*). The answer to these questions is, of course, that other things might be wrong that the reports don't control for.

We are at the cutting edge of science, every new phenomenon that we investigate may, or may not, be real. With high probability, the rabbit in the clouds is not really a rabbit. The prior distribution of each new and surprising effect is mostly concentrated at zero, and skepticism should be foremost in our mind.

Ethos

Preregistration assumes that the researcher is prone, in the absence of guardrails, to misleading his/herself and others. To some extent this is justified by examples of, to put it

mildly, 'cheerful blindness', positively selected because it confers a competitive edge. On the other hand, nobody is more qualified than the researcher to ferret out the many ways a study can go wrong, and we cannot but trust them to do so. Signaling that trust is not expected may be counterproductive: "*when you rely on incentives, you undermine virtues*" (Schwartz 2009). Questionable research practices are driven by strong incentives, profit-driven business models of scientific publishers, and cut-throat professional competition (Nosek et al 2012, Edwards and Roy 2017, Munafò 2017, Vazire et al 2021). We lack leverage to address those factors, but it is worth keeping in mind that they are at the root of the problem. It is unfair to lay *all* the blame on researchers' shortcomings.

It may seem naïve to count on the researcher's ethos, in the face of the many incentives to publish and promote regardless of truth (Poldrack 2019). On the other hand, why else work in science?

Opportunities for improvement

Overhead for the investigator

The overhead depends on the scheme adopted, with at one extreme a simple scheme such such as AsPredicted (https://aspredicted.org), and at the other Registered Reports. It is larger if the ambition is to exhaustively specify *all* details in advance, or if proof of compliance is required at other phases, such as grant approval. This is a matter of taste, but I feel we are better served by light-weight schemes than more comprehensive schemes, that promise greater rigor but require more time to prepare and report, and carry the risk of turning the process into a 'bureaucratic exercise' (Munafò et al 2017, Chambers and Tzavella 2021, Errington et al 2021). There is a point of diminishing returns: lighter is better.

Overhead for the reader

Space is required in the paper to mention the plan and report deviations (or lack thereof). A report of every ill-fated (but preregistered) experiment or hypothesis (van Assen et al 2014, Amrhein et al 2019) is not of prime interest to most readers. On the other hand, such details should be documented and discoverable. An interesting idea is that of a 'post-preregistration' report (Banerjee et al 2020) that tersely describes the outcome of each planned experiment, or informs us that it was abandoned. This would address a shortcoming of the current model, mentioned earlier, that the fate of a preregistered but unpublished experiment is unspecified. The follow-up report, archived together with the plan, would close, as it were, the parenthesis opened by the preregistration plan. A worthy 'nudging' target might be to convince researchers to add a follow-up to each preregistered plan, even just to say that the plan didn't materialize. This would avoid bloating papers with preregistration-related details: the cognitive bandwidth of the reader is a resource worth preserving.

Epistemological bias

In my own work, I would have found it hard to provide a "*specific, precise, and exhaustive plan*" of each study in advance (Wicherts et al 2016). In some instances, I was learning tools of the trade at the same time as the topic, in others forging new tools. I have also observed teams, often large, that work productively on well-defined topics using standard tools. Stringent preregistration standards fit the latter style but perhaps less the former. In an attempt to cure one bias (statistical), we risk introducing another bias (epistemological). One could argue, perhaps facetiously, that a field for which every experiment can be planned in full detail is not just mature, but overripe (Wilson and Wixted 2018, National Academy of

Science 2019, see Lakatos 1976 and Redish et al 2018 for analogous arguments in the context of mathematical proofs).

Preregistration offers a neat algorithm to label results as *exploratory* or *confirmatory*: if a preregistered experiment and analysis yield a statistically significant result, that result is confirmatory, otherwise it is exploratory (Nosek 2018). However, for Popper (1935) 'confirmatory' is an overstatement: "*a positive decision can only temporarily support the theory, for subsequent negative decisions may always overthrow it*". A statistically significant result from a preregistered study can still be wrong. What today we judge to be confirmatory might tomorrow appear to have been exploratory, suggesting a continuum rather than a dichotomy (see Szollosi and Donkin 2021, and Rubin 2020 for additional arguments). If we follow Popper, the bulk of science consists of exploratory results.

Perverse incentives

Rules designed to promote one behavior may trigger others, unintended. Required to write (and stick to) a detailed plan, an investigator may tweak the study to make that task easier, possibly at the expense of a more ambitious and interesting study. Having submitted the plan, he/she may be reluctant to make reasonable adjustments to the design, even when scientifically advisable, to avoid having to downgrade the results from confirmatory to exploratory. An unintended effect of the algorithm mentioned earlier is to induce this hierarchy, since 'exploratory' is what happens to a study when we are forced to void its preregistration. If 'confirmatory' became a label of excellence, journals might come to adopt the policy of publishing only confirmatory results, or funding agencies insist that they be delivered. A norm induces a landscape of incentives which people then learn to navigate.

The Cost of Confusion

Should I go for a Registered Report or just preregistration? Should I favor AsPredicted, or OSF and if so, which template? Should I use the 36-item checklist of Aczel et al (2020), or just the 12-item list? How seriously should I take this exercise? If the preregistered plan turns out to be unworkable, should I stop the experiment and re-preregister? I saw some small fraction of the data, is that really bad? I made a minor adjustment, must I label the results as exploratory and remove the p-values? A reviewer objected to my preregistered analysis, their objection is valid, what should I do? Should I repeat methods in the paper methods that were described in the preregistration? I noticed some typos and the preregistration is really unclear, should I revise it to make it more useful for the reader? I'm a reviewer, the preregistered plan is poorly written or self-contradictory, should the authors be encouraged to fix it? Or should I recommend rejection, despite the results being otherwise sound and the conclusions interesting? The study involves a Registered Report, which is supposed to guarantee publication, but the writing is abysmal, may I recommend rejection? The Registered Report was flawed, despite successful review, what should I do? Deciding these points may require lengthy head-scratching. A wrong decision might lead to good work being rejected, or poor work accepted.

Flexibility for the researcher might seem ideal, but there are other actors in the loop, such as supervisors, editors, and reviewers of grants and papers, and they too have flexibility. That is probably what was worrying my colleague and his student, prompting them to 'play safe' to avoid the risk of running afoul of someone else's interpretation of the rules.

In summary

Preregistration rides on a strong cultural current and is supported by an active community. It is here to stay. At the same time, its norms, practices and tools are in flux and not universally agreed across disciplines, and there may be a window of opportunity to tune them to further enhance its benefits, keeping in mind that costs and benefits are in a tradeoff, that norms interact with each other, and that an adjustment (however justified) might increase the confusion. The following suggestions are based on my personal reading of the situation.

Keep it light

Preserve cognitive bandwidth

Cognitive resources are scarce. The call to ‘report everything’ should be resisted. The ‘post-preregistration follow-up report’ (Banerjee et al 2020) might be a way to ensure transparency without bloating papers.

Leverage skepticism and trust

The investigator is ideally placed to spot problems and act upon them. It must be clear that this is the right thing to do. This boils down to an ethos interiorized by the investigator and expected by the field.

Go easy on enforcement

Norms demand compliance, but side-effects are in proportion to enforcement. The tradeoff between the benefits of greater compliance and the associated side effects may be adjusted based on the availability of alternatives, such as ethos, replication, and scrutiny.

Consider alternative tools

Replication and scrutiny serve directly to test reliability, and indirectly by promoting skepticism, strengthening norms, and signaling that rigor is expected. Replication can be encouraged by recognizing its scientific value, lobbying to lower barriers to publication, and leveraging Nosek and Errington's (2020) definition that pitches it as a positive, knowledge-seeking endeavor. Scrutiny is encouraged by strengthening requirements to share well-documented data and code.

Go easy on nudging

Behavior modification techniques (UK Behavioral Insights Team 2014, Nosek 2019, Rubin 2020, Mayo-Wilson 2021) help overcome inertia and perverse incentives. However, they also weaken our ability to check that new norms are indeed beneficial and resist them if not. Nudging overrides the 'wisdom of crowds'. 'Transparency factors' and 'badges' may lead to gaming strategies to maximize reward along one particular dimension at the expense of others within the complex, high-dimensional value landscape of science. By conferring prestige to their awardees, they may shield them from criticism and induce overconfidence.

Conclusion

Preregistration addresses an important need, and constitutes a powerful tool as part of a 'transparency' bouquet that also includes tools to promote data sharing and documentation. Its benefits should be balanced with the burden and risks associated with making preregistration mainstream, in particular possibly pernicious effects of 'nudging' designed to hasten the uptake. The purpose of this paper was to sift through these effects, weighing each

as best I could, to hopefully find a good balance between freedom to do good research, and protection from freedom to mislead others and ourselves.

Acknowledgments

Daniel Pressnitzer, Andrew Gelman, Daniël Lakens gave useful comments on an early draft of this paper.

Funding

This work was supported by grants ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL, and ANR-17-EURE-0017.

Ethics

n/a

Competing interests

none

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M.A., Ioannidis, J.P., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S.O., Lindsay, D.S., Morey, C.C., Munafò, M., Newell, B.R., ... Wagenmakers, E.-J. (2020). A consensus- based transparency checklist. *Nature Human Behaviour*, 4 (1), 4-6. doi: 10.1038/s41562-019-0772-6

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician*, 73 (sup1), 262–270. doi: 10.1080/00031305.2018.1543137

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, 26 (5), 527–546. doi: 10.1037/met0000365

Ashton, J. C. (2013). Experimental power comes from powerful theories – the real problem in null hypothesis testing. *Nature Reviews Neuroscience*, 14 (8), 585–585. doi: 10.1038/nrn3475-c2

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., ... Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, 18 (12), e3000937. doi: 10.1371/journal.pbio.3000937

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L. F., Olken, B. A., & Sautmann, A. (2020). IN PRAISE OF MODERATION: SUGGESTIONS FOR THE SCOPE AND USE OF PRE-ANALYSIS PLANS FOR RCTS IN ECONOMICS (Working Paper No. 26993). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w26993

Benjamin, D. J., & Berger, J. O. (2019). Three Recommendations for Improving the Use of p - Values. *The American Statistician*, 73 (sup1), 186–191. doi: 10.1080/00031305.2018.1543135

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K.A. Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10. doi: 10.1038/s41562-017-0189-z

Berenbaum, M. R. (2021, August). On zombies, struldbrugs, and other horrors of the scientific literature. *Proceedings of the National Academy of Sciences*, 118 (32), e2111924118. doi: 10.1073/pnas.2111924118

Bertamini, M., & Munafò, M. R. (2012, January). Bite-Size Science and Its Undesired Side Effects. *Perspectives on Psychological Science*, 7 (1), 67–71. doi: 10.1177/1745691611429353

Box, G. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71, 791–799. Retrieved from https://www.jstor.org/stable/2286841

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò & M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. doi: 10.1038/nrn3475

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49(3), 609–610. doi: 10.1016/j.cortex.2012.12.016

Chambers, C. D., & Tzavella, L. (2021). The past, present and future of Registered Reports. *Nature Human Behaviour* . Retrieved doi: 10.1038/s41562-021-01193-7

Chang, A. C., & Li, P. (2015). Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not". *Finance and Economics Discussion Series*, 2015 (83), 1–26. doi: 10.17016/FEDS.2015.083

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, 8 (10), 211037. doi: 10.1098/rsos.211037

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65 (3), 145–153. doi: 10.1037/h0045186

DataColada (2015), http://datacolada.org/44.

Davis, W. E., Giner-Sorolla, R., Lindsay, D. S., Lougheed, J. P., Makel, M. C., Meier, M. E., Sun J., Vaughn, L.A. & Zelenski, J. M. (2018). Peer-Review Guidelines Promoting Replicability and Transparency in Psychological Science. *Advances in Methods and Practices in Psychological Science*, 1 (4), 556–573. doi: 10.1177/2515245918806489

de Cheveigné, A. (2021). Harmonic Cancellation - A Fundamental of Auditory Scene Analysis. *Trends in Hearing*, 25. doi: 10.1177/23312165211041422

de Cheveigné, A. (2022). Local Subspace Pruning (LSP), *bioRxiv*. https://doi.org/10.1101/2022.02.27.482148

DeHaven, A. (2017) `Preregistration: a Plan not a Prison'. https://www.cos.io/blog/preregistration-plan-not-prison

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. doi: 10.1089/ees.2016.0223

Else, H., & Van Noorden, R. (2021). The battle against paper mills. *Nature*, 591, 516–519.

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in pre-clinical cancer biology. *eLife*, 10, e67995. doi: 10.7554/eLife.67995

Felgenhauer, M. (2021). Experimentation and manipulation with preregistration. *Games and Economic Behavior*, 130, 400–408. doi: 10.1016/j.geb.2021.09.002

Fidler, F., Thorn, F. S., Barnett, A., Kambouris, S., & Kruger, A. (2018). The Epistemic Importance of Establishing the Absence of an Effect. *Advances in Methods and Practices in Psychological Science*, 1, 237– 244. doi: 10.1177/25152459187704

Frank, M. C., & Saxe, R. (2012). Teaching Replication. *Perspectives on Psychological Science*, 7, 5. doi: 10.1177/1745691612460686

Friston, K. (2018). Does predictive coding have a future? *Nature Neuroscience*, 21 (8), 1019–1021. doi: 10.1038/s41593-018-0200-7

Gelman, A. (2017). Ethics and Statistics Honesty and Transparency Are Not Enough. *CHANCE*, 30 (1), 37–39. doi: 10.1080/09332480.2017.1302720

Gelman, A., & Vazire, S. (2021). Why Did It Take So Many Decades for the Behavioral Sciences to Develop a Sense of Crisis Around Methodology and Replication? *Journal of Methods and Measurement in the Social Sciences*, 12(1). doi: 10.2458/jmmss.3062

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, 351 (6277), 1037–1037. doi: 10.1126/science.aad7243

Grand, J. A., Rogelberg, S. G., Banks, G. C., Landis, R. S., & Tonidandel, S. (2018). From Outcome to Process Focus: Fostering a More Robust Psychological Science Through Registered Reports and Results-Blind Reviewing. *Perspectives on Psychological Science*, 13 (4), 448–456. doi: 10.1177/1745691618767883

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337–350. doi: 10.1007/s10654-016-0149-3

Hackley, S. A. (2015). Evidence for a vestigial pinna-orienting system in humans: Pinna orienting. *Psychophysiology*, 52 (10), 1263–1270. doi: 10.1111/psyp.12501

Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2), 431–448. doi: 10.1111/rssa.12493

Held, L., & Ott. M. 'On P-Values and Bayes Factors'. *Annual Review of Statistics and Its Application* 5, no. 1 (7 March 2018): 393–419. https://doi.org/10.1146/annurev-statistics-031017-100307.

Higginson, A. D., & Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLOS Biol*ogy, 14 (11), e2000995. doi: 10.1371/journal.pbio.2000995

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

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

Jaynes, E.T. (1979) Probability, the Logic of Science, edited by Bretthorst, G.L., Cambridge University Press (2002).

John, L., K, Loewenstein, G., & Prelec, D. (2011). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23, 254–232.

Kidwell, M. C., Lazarevi\'c, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T.M., Fiedler, S. & Nosek, B. A. (2016, May). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, 14 (5), e1002456. doi: 10.1371/journal.pbio.1002456

Koole, S. L., & Lakens, D. (2012). Rewarding Replications: A Sure and Simple Way to Improve Psychological Science. *Perspectives on Psychological Sci*ence, 7(6), 608–614. doi: 10.1177/1745691612462586

Lakatos, I (1976). Proofs and refutations, Cambridge University Press, ISBN 978-0-521-29038-8.

Lakens, D. (2019). The Value of Preregistration for Psychological Science: A Conceptual Analysis. *Japanese Psychological Review*, 62. doi: 10.31234/osf.io/jbh4w

Lakens, D. (2021). Invited commentary: Comparing the independent segments procedure with group sequential designs. *Psychological Methods*, 26 (4), 498–500.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Argamon, S.E., Baguley, T., Becker, R.B., Benning, S.D., Bradford, D.E., Buchanan, E.M, Caldwell, A.R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L.J., Collins, G.S., Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, 2(3), 168–171. doi: 10.1038/s41562-018-0311-x

Lawlor, D. A., Tilling, K., & Davey Smith, G. (2017). Triangulation in aetiological epidemiology. International *Journal of Epidemiology*, dyw314. doi: 10.1093/ije/dyw314

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods*, 51 (6), 2498–2508. doi: 10.3758/s13428-018-1092-x

Mayo-Wilson, E., Grant, S., Supplee, L., Kianersi, S., Amin, A., DeHaven, A., & Mellor, D. (2021). Evaluating implementation of the Transparency and Openness Promotion (TOP) guidelines: the TRUST process for rating journal policies, procedures, and practices. *Research Integrity and Peer Review*, 6(1), 9. Retrieved 2022-01-26, doi: 10.1186/s41073-021-00112-8

Meehl, P. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defence and Two Principles that Warrant It. *Psychological Inquiry*, 1, 108–141.

Miller, J., & Ulrich, R. (2022). Optimizing Research Output: How Can Psychological Research Methods Be Improved? *Annual Review of Psychology,* 73 (1), annurev–psych–020821–094927. doi: 10.1146/annurev-psych-020821-094927

Minocher, R., Atmaca, S., Bavero, C., McElreath, R., & Beheim, B. (2021). Estimating the reproducibility of social learning research published between 1955 and 2018. *Royal Society Open Science*, 8 (9), 210450. doi: 10.1098/rsos.210450

Miyakawa, T. (2020). No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain*, 13:24, 6. doi: 10.1186/s13041-020-0552-2

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1 (1), 0021. doi: 10.1038/s41562-016-0021

National Academies of Sciences, Engineering, and Medicine (2019). Reproducibility and Replicability in Science. Washington, DC: *The National Academies Press*, https://doi.org/10.17226/25303.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, 69 (1), 511–534. doi: 10.1146/annurev-psych-122216-011836

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *eLife*, 5, e21451. doi: 10.7554/eLife.21451

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A.E., &Vazire, S. (2019) Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, 23 (10), 815–818. doi: 10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115 (11), 2600–2606. doi: 10.1073/pnas.1708274114

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, 18 (3), e3000691. doi: 10.1371/journal.pbio.3000691

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, 7 (6), 615–631. doi: 10.1177/1745691612459058

Nuzzo, R. (2015). Fooling Ourselves. *Nature*, 526, 182–185.

Ofosu, G. K., & Posner, D. N. (2021). Pre-Analysis Plans: An Early Stocktaking. *Perspectives on Politics*, 1–17. doi: 10.1017/S1537592721000931.

Oldehinkel, A. J. (2018). The importance of taking no for an answer. *Nature Human Behaviour*, 2 (8), 533–534. doi: 10.1038/s41562-018-0393-5

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, 7 (6), 531–536. doi: 10.1177/1745691612463401

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

Proulx, T., & Morey, R. D. (2021). Beyond Statistical Ritual: Theory in Psychological Science. *Perspectives on Psychological Science*, 16 (4), 671–681. doi: 10.1177/17456916211017098

Redish, A. D., Kummerfeld, E., Morris, R. L., & Love, A. C. (2018) Reproducibility failures are essential to scientific inquiry, *Proceedings of the National Academy of Sciences*, 115, 5042–5046, https://doi.org/pnas.18063701.

Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology*, 16 (4), 376–390. doi: 10.20982/tqmp.16.4.p376

Schwartz, B (2009) https://www.wired.com/2009/02/ted-barry-schwa/.

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, 7(21), eabd1705. doi: 10.1126/sciadv.abd1705

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, 115(11), 2632–2639. doi: 10.1073/pnas.1711786114

Simmons, J., Nelson, L., & Simonsohn, U. (2021). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22, 1359–1366.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, 12, 1123–1128.

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, 26 (5), 559–569. doi: 10.1177/0956797614567341

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

Stefan, A. M., Lengersdorff, L. L., & Wagenmakers, E.-J. (2022). A Two-Stage Bayesian Sequential Assessment of Exploratory Hypotheses. *Collabra: Psychology*, *8*(1), 40350. https://doi.org/10.1525/collabra.40350

Stewart, A. J., & Plotkin, J. B. (2021). The natural selection of good science. *Nature Human Behaviour*, 5 (11), 1510–1518. doi: 10.1038/s41562-021-01111-x

Szollosi, A., & Donkin, C. (2021). Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research. *Perspectives on Psychological Science*, 16 (4), 717–724. doi: 10.1177/1745691620966796

UK Behavioural Insights Team. (2014). *EAST: Four simple ways to apply behavioural insights*. https://www.bi.team/publications/east-four-simple-ways-to-apply-behavioural-insights/.

Ulrich, R., & Miller, J. (2020). Questionable research practices may have little effect on replicability. *eLife*, 9, e58237. doi: 10.7554/eLife.58237

van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014). Why Publishing Everything Is More Effective than Selective Publishing of Statistically Significant Results. *PLoS ONE*, 9 (1), e84896. doi:10.1371/journal.pone.0084896

Wagenmakers, E.-J., & Forstmann, B. U. (2014). Rewarding high-power replication research. *Cortex*, 51, 105–106. doi: 10.1016/j.cortex.2013.09.010

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, 7 . doi: 10.3389/fpsyg.2016.01832

Yuval-Greenberg, S., & Deouell, L. Y. (2011). Scalp-Recorded Induced Gamma-Band Responses to Auditory Stimulation and Its Correlations with Saccadic Muscle-Activity. *Brain Topography*, 24 (1), 30–39. doi: 10.1007/s10548-010-0157-7

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41, e120. doi: 10.1017/S0140525X17001972