Seven Steps Toward Transparency and Replicability in Psychological Science

D. Stephen Lindsay

University of Victoria

Accepted 1 April 2020 (I'm pretty sure it was a real acceptance) for inclusion in a forthcoming special issue of *Canadian Psychology*.

Author Note

D. Stephen Lindsay https://orcid.org/0000-0002-6439-987X

I have no conflicts of interest to disclose.

Correspondence concerning this manuscript should be addressed to D. Stephen Lindsay, Department of Psychology, University of Victoria, Victoria, B.C., Canada V8W 2Y2. Email: slindsay@uvic.ca

I received extraordinarily detailed, well-informed, thought-provoking input on an earlier draft of this work from Kaitlyn M. Fallow, David Mellor, John K. Sakaluk, and Simine Vazire. I am very grateful. Any weaknesses or errors that remain in the paper are mine.

Abstract

Psychological scientists strive to advance understanding of how and why we animals do and think and feel as we do. This is difficult, in part because flukes of chance and measurement error obscure researchers' perceptions. Many psychologists use inferential statistical tests to peer through the murk of chance and discern relationships between variables. Those tests are powerful tools, but they must be wielded with skill. Moreover, research reports must convey to readers a detailed and accurate understanding of how the data were obtained and analyzed. Research psychologists often fall short in those regards. This paper attempts to motivate and explain ways to enhance the transparency and replicability of psychological science. Specifically, I speak to how publication bias and p hacking contribute to effect-size exaggeration in the published literature, and how effect-size exaggeration contributes, in turn, to replication failures. Then I present seven steps toward addressing these problems: Telling the truth; upgrading statistical knowledge; standardizing aspects of research practices; documenting lab procedures in a lab manual; making materials, data, and analysis scripts transparent; addressing constraints on generality; and collaborating.

179 words

Public Significance Statement

Research psychologists often use statistical analyses to interpret and communicate their findings to other researchers. Unfortunately, imperfect understanding of those statistical tools, in combination with pressure to publish or perish, incentivise research practices that tend to yield exaggerated estimates of the strength of evidence. This article argues for seven steps that researchers can take to correct those problems. The underlying theme of the recommendations is that scientists must take pains to provide clear and detailed explanations of how they came to have their data and how they analyzed them.

Seven Steps Toward Transparency and Replicability in Psychological Science

Scholarly works and popular media have in recent years questioned the replicability of findings in psychology. For example, the Open Science Collaboration (2015) reported attempts to replicate 100 published effects, only a minority of which succeeded. Similarly, Camerer et al. (2018) reported attempts to replicate 21 experiments originally published in *Nature* and *Science*, with less than two thirds succeeding. Replication attempts can fail for any of several reasons (e.g., maybe the replication did not properly recreate the essential conditions under which the original effect was obtained, or maybe the failure was just a Type II error), but most attention has focused on the role of faulty research practices in the original research, leading to calls for methodological reform to correct those practices and make psychological science more cumulative and useful (Bishop, 2019; Chambers, 2017; Crüwell et al., 2019; Frith, 2019; Munafo, 2017; Nelson, Simmons, & Simonsohn, 2018; Spellman, 2015).

My aim here is to communicate with psychologists who may have heard about a "replication crisis" but who are not deeply versed on the topic and may be wary of calls for methodological reform. I am not a statistician. I consider myself a follower/promoter of methodological reform, not a leader. My comments draw on more than 30 years of experience as a researcher and two terms as a journal editor, plus intensive exposure in recent years to articles, talks, and workshops on statistical and methodological issues. I explain how and why well-meaning researchers (including my past self) sometimes use methods that exaggerate the size of effects (or the strength of correlations) and rarely adequately highlight limitations on the generalizability of their findings. Then I discuss the relationship between effect-size exaggeration and replication failures. Finally, I describe seven steps that researchers can take to

enhance the transparency and replicability of their work. These ideas are not new, but I hope to communicate them in an effective way.

Effect Size Exaggeration

Some areas of psychology often publish exaggerated estimates of the size of effects. In some such cases, the true effect may be non-trivial but smaller than the published literature indicates. In other cases, the true effect size may be essentially zero (e.g., Bem's 2011 reports of statistically significant ESP; see Francis, 2012). In the following I discuss causes of effect-size exaggeration, then explain how effect-size exaggeration contributes to poor replicability.

Publication bias. One major contributor to effect-size exaggeration is publication bias, which selectively favours the publication of studies that obtain statistically significant effects. Many journal editors and reviewers make statistical significance a near-criterion for publication. Responding to that incentive, many researchers conduct multiple studies of a hypothesized effect, each with modest sample sizes, attempting to discover conditions under which the hypothesized effect is strong. They then submit for publication the subset of studies that yielded large effects.

A problem with this approach is that random sampling and measurement error can cause large swings in effect-size estimates. Figure 1 is a screenshot of Cumming's (2011) ESCI program.[1] Understanding the figure takes effort but it is worthwhile (and research indicates that published research psychologists often have poor intuitions regarding the issues the figure aims to illuminate; Bakker, Hartgerink, Wicherts, & van der Maas, 2016).

[1] This figure was made using the Excel 2003 version of ESCI updated in 2012, which can be downloaded from https://thenewstatistics.com/itns/esci/esci-for-utns/, Chapters 5-6, dance p. Thanks to my dear friend Maryanne Garry for introducing me to Geoff Cumming. Changed my career.

The figure shows 25 simulated experiments, each comparing randomly sampled groups of *N* = 20 from control and experimental populations that differ by Cohen's d = 0.50 (half a standard deviation, often described as a medium-sized effect). The population distributions at the top of the figure represent the truth that research seeks to reveal. The 25 solid green circles represent the estimated raw effect sizes (and the 95% confidence interval around those estimates) in 25 experiments. The column on the left shows the *p* value obtained in each simulated experiment.

Most of the 25 stimulated experiments obtained results in the correct direction, but due to low statistical power the difference was often not statistically significant. Because the true average effect is *d* = 0.50, a between-subjects experiment with *N* = 20 in each of two groups has about 33% power. Thus, about two thirds of experiments yield $p > .05$. In this set of simulations, *p* values ranged from $p < .001$ to $p = .98$. This demonstrates the noisiness of *p* values and dramatizes the facts that (a) large *p* values do not compel the null hypothesis and (b) a small *p* value does not necessarily presage replication. Please take some time to study this figure.

In these simulations, every experiment that yielded $p < .05$ obtained an exaggerated estimate of effect size. This is because when the true effect is *d* = 0.50 and there are only 20 subjects per between-subjects condition, it is not possible to get $p < .05$ unless the study happens to yield an exaggerated estimate of the size of the effect.

Conducting multiple studies with smallish samples (especially in designs that include a between-subject factor) and then selecting for publication a subset of studies that yielded statistically significant results contributes to effect-size exaggeration. I did this for years with the best of intentions.

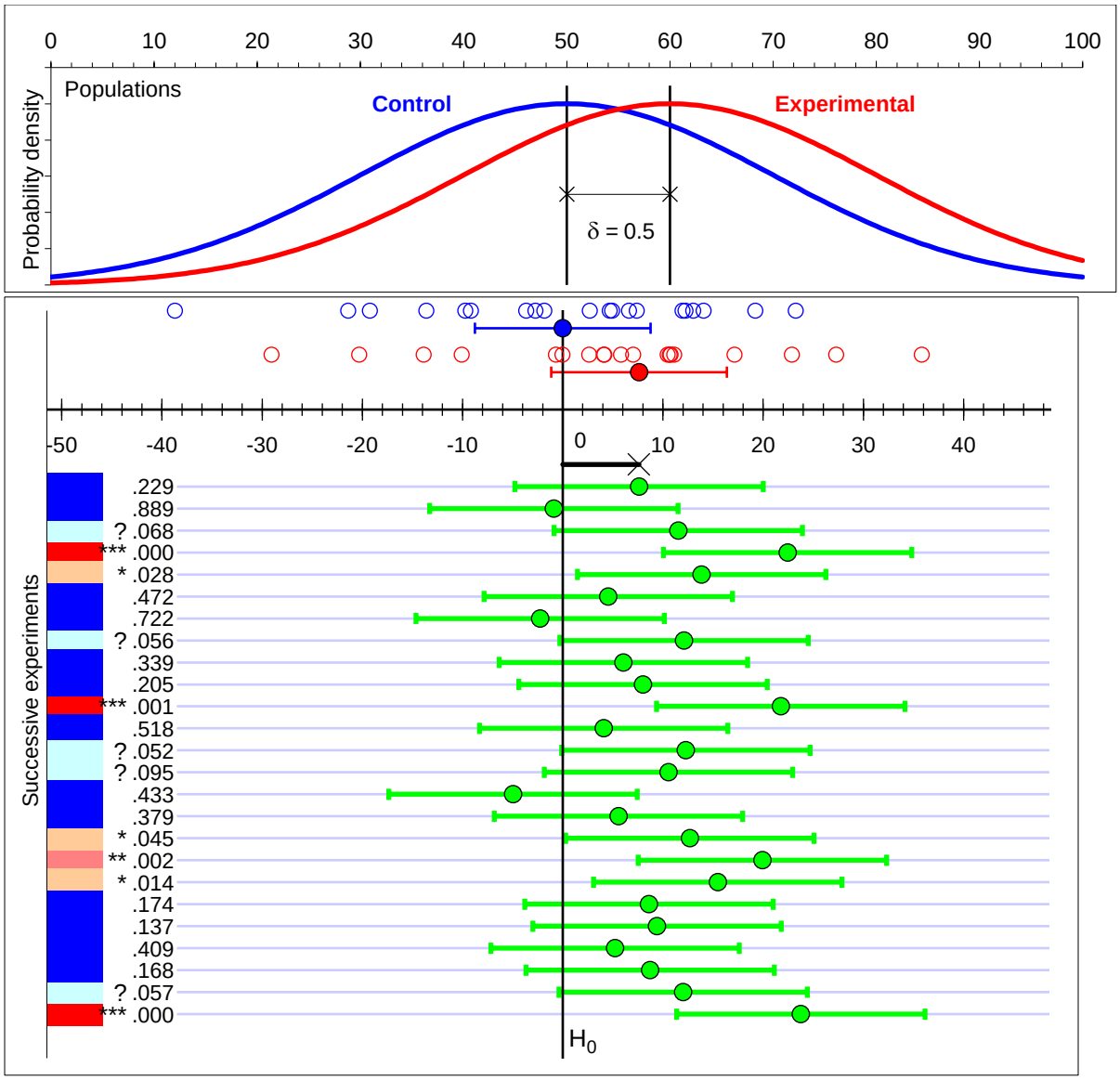

Figure 1. Screenshot of Cumming's ESCI. The blue and red distributions at the top depict population values for control and experimental conditions that differ by Cohen's *d* = 0.5 (or 10 raw points); this represents the reality research aims to reveal. The 20 open blue and 20 open red circles were randomly sampled from the control and experimental conditions, respectively. Each solid green circle represents the result of a simulated experiment: A raw effect size (experimental mean minus control mean) from random draws from the two populations, with the 95% confidence interval around that effect-size estimate. The first (topmost) solid green circle represents the difference between the open blue and red circles. In that random draw, the difference between conditions was not statistically significant (*p* = .229, as shown in the left margin), a Type II error. Of the 25 simulated experiments shown here, 3 came out in the wrong direction due to random sampling error; 7 experiments detected the effect and every one of those overestimated the size of the effect.

p hacking. In addition to publication bias, a variety of common research practices can further inflate effect-size estimates. These involve deciding which hypotheses to test and how to test them based in part on the results in hand. A researcher might, for example, decide to collect more data; to exclude subjects, conditions, measures, or observations; to transform a measure; or to add a covariate to an analysis, with the aim of finding a pattern that is statistically significant. Such practices have been variously referred to as p hacking (Nelson, Simmons, & Simonsohn, 2018), hypothesizing after the results are known (HARKing; Kerr, 1998), questionable research practices (John et al., 2012), and researcher degrees of freedom (Wicherts, Veldkamp, Augusteijn, Bakker, van Aert, & van Assen, 2016). John et al. reported survey evidence that such practices are widespread. The problem is that when decisions about analyses are guided by the data in hand *p* values cannot be interpreted in a straightforward way and Type I error rates can be vastly higher than alpha.

Effect size exaggeration and replication failures. Suppose a journal article reported an experiment in which two groups of 20 people differed a dependent variable by Cohen's *d* = 1.1, as in the fourth simulated experiment from the top in Figure 1. Intrigued, you set out to conduct follow-up research. You might decide to test the same number of subjects as in the original study. Testing 40 subjects would give you 91% power to detect an effect of Cohen's d = 1.1 by a two-tailed test at alpha = .05. But if the average effect size under the conditions tested is only Cohen's d = 0.50 (as was in fact the case in the simulation underlying Figure 1), then power in your follow up study will not be 91% but only about 33%. Your follow-up would be expected to fail two thirds of the time even if it was a *perfect* direct replication. Because

publication bias and p hacking often lead to inflated estimates of effect size, follow-up studies that use sample sizes comparable to the original work often fail to attain statistical signifance.

One might hope that meta-analysis (i.e., efforts to estimate effect size by combining results across many experiments) could correct for effect-size exaggeration. The problem is that studies that (for whatever reason) obtained large, statistically significant effects are much more likely to be discoverable by meta-analysts than are other studies of the same effect that yielded null results. Experts in meta-analysis have proposed methods to correct for effect-size exaggeration arising from publication bias and p hacking, but simulations indicate that such corrections are often grossly inadequate, so meta-analyses must be interpreted with caution (e.g., Carter, Schönbrodt, Gervais, & Hilgard, 2019; Corker, in press).

How can psychologists who use null hypothesis significance testing (NHST) decide how many subjects to test? There is no certain way to answer that question. From a statistical point of view more is better, but practical and ethical constraints limit sample size. One rule of thumb is to test a sample large enough to have a high probability of detecting an effect half as large as that reported in the to-be-replicated study (Schäfer & Schwarz, 2019). Setting out to replicate a finding of Cohen's d = 1.10, for example, one could calculate the *N* required to have 80% power to detect an effect of d = 0.55 with a one-tailed test (which, according to G*Power 3, is 84 subjects; Faul, Erdfelder, Lang, & Buchner, 2007). Taking a different approach, Simonsohn (2015) recommended testing 2.5 times more subjects than were in the to-be-replicated study (i.e., in this case, 80). For yet another approach that takes into account the uncertainty of the estimate of the size of the effect, see Anderson, Kelley, and Maxwell (2017).

Seven Steps Toward Increasing Transparency and Replicability

Effect-size exaggeration in the published literature can be reduced by changing norms in two broad ways. One is to improve statistical sophistication among psychologists who test hypotheses about populations based on samples. Another is to shift our culture to reward quality and accuracy, rather than quantity and flashiness.

Even together, these two changes—better understanding of statistical tools and emphasizing quality over quantity--cannot deliver on the full promise of psychological science. For that, we also need better theory and better measures (Devezer, Nardin, Baumgaertner, & Buzbas, 2019; Flake & Fried, 2019; Pexman & Jamieson, in press; Szollosi et al., 2019; Yarkoni, 2020). But although the changes advocated here are not sufficient to make psychology a robustly useful science, they would at least reduce the frequency with which we confuse matters by publishing inflated claims of effects that in reality are tiny or nonexistent. Here I propose seven steps that I believe would be particularly helpful.

1. Tell the truth. Take pains to disclose rather than to cover up. Transparency is the core of methodological reform. Vazire (2017) drew an analogy between reading a scientific report and shopping for a used car. Slick sales jobs and jerry-built repairs can generate quick profits, but they tend to lower the overall value of the market because they undermine buyers' confidence (see also Vazire, 2019). It is appropriate to be an advocate of your research, but not a huckster. So, don't imply that you had an a priori hypothesis if really the idea was inspired by the data. Don't p hack. Report your findings in ways that disclose them in detail, warts and all. Use richly detailed graphs, such as frequency histograms, scatterplots, or box plots with jittered data points (which can be easily made with the free JASP program, https://jasp-stats.org/).

Report measures of effect size (and/or relationship strength) with 95% confidence intervals (CIs) around them. If editors or reviewers push you to tell a story that makes your evidence appear stronger than it is, push back. If you are a reviewer or editor, reward frankness and don't encourage p hacking. For further tips on writing and reviewing transparently, see Davis et al. (2018) and Mellor, Vazire, and Lindsay (2018).

2. Assess your understanding of inferential statistical tools. There is evidence that many psychologists who use NHST do not have a solid understanding of its core concepts, such as what *p* values mean or how to determine sample size (Belia, Fidler, Williams, & Cumming, 2005; Wicherts et al., 2006). Cassidy et al. (2019) reported that 89% of introductory psychology texts that attempted a definition of *p* got it wrong. I misunderstood key aspects of NHST for many years. I'm still no stats maven, but I have learned a lot recently and I believe it has made me a better scientist.

You don't have to become a mathematician. A few core insights will take you a long way. As a first step, I recommend *The Dance of the p Values*, a video by Geoff Cumming (there are several versions on YouTube, such as www.youtube.com/watch?v=5OL1RqHrZQ8&t=12s). Then, Simmons, Nelson, and Simonsohn's (2011) article "False Positive Psychology" and a more recent *Annual Reviews in Psychology* chapter by Nelson et al. (2018). As another useful resource, Makin and Orban de Xivry (2019) recently presented a concise review of 10 common statistical mistakes in journal submissions.

3. Consider standardizing aspects of your approach to conducting hypothesis testing research. Strive to reduce the extent to which your decisions about what hypotheses to test and how to test them are biased by the data you happened to obtain. That's cheating! It is

great to be alive to serendipitous patterns in your data, but it is not helpful to mistake them for predicted patterns.

One useful practice is to create a detailed research plan stating *a priori* hypotheses and specifying sample size, data exclusion rules, analyses, transformations, covariates, etc. That plan can be "registered" on a website such as the Open Science Framework or AsPredicted.org, thereby creating an immutable, date-stamped record of the plan (called a preregistration). Researchers can, if they wish, give editors and reviewers access to their preregistered research plan. See Lindsay, Simons, and Lilienfeld (2016) for a brief introduction to preregistration. For a step-be-step guide to creating a preregistered research plan on the Open Science Framework, see https://help.osf.io/hc/en-us/articles/360019738834-Create-a-Preregistration.

Transparency is a key aim of creating and registering a research plan. And transparency is a central goal of the methodological reform movement (which is partly what it is often called "open science"). The key idea is that readers of a scientific article should be able to gain a detailed and accurate understanding of how the researcher(s) obtained and analyzed the data. Such knowledge is essential for assessing the meaning of the reported findings.

A preregistration does not need to specify all details of a research plan. Indeed, at an early stage of conducting empirical work in a new area, a preregistration might be quite vague, leaving many decisions to be made on the fly. Even a vague plan can help the researcher think through a project in advance and protect them from later mistakenly believing they had *a priori* hypotheses that really only developed in view of the data.

Researchers are free to deviate from a preregistered plan for a research project—but they can only do so knowingly (and, if they share the plan with others, openly). Researchers are

always free to conduct exploratory analyses – they just cannot so easily present them as if they had been planned.

Writing a good preregistration is not easy. The task is especially arduous in the very domains in which preregistrations are most useful, namely those in which there are many decisions the researcher must make in the absence of strong consensus and tightly constraining theory. It is often difficult to anticipate all the judgments that will need to be made, and often difficult to determine the best decisions *a priori* (which is part of the value of the practice). Decisions you cannot make in advance can at least be honestly identified as such.

It is not always easy for readers to assess the completeness of a preregistration and the extent to which the research reported was consistent with the preregistered plan. Efforts are under way to develop tools that make it easier to develop and evaluate preregistered research plans (e.g., Aczel et al., 2019, who introduced a web app called the Transparency Checklist).

Preregistration is sometimes confused with the Registered Report (RR) approach to publishing. The two are related but different. In the RR model, a detailed research proposal is submitted to a journal before data are collected. The editor sends this Stage 1 proposal for peer review. The Stage 1 proposal is judged on perceptions of (a) the importance of the question to be addressed and (b) the rigour and appropriateness of the methods. Often the review process leads to revisions to the proposal. If the editor eventually judges the proposal worthy, then it receives in-principle Stage 1 acceptance. If the author completes and writes up the work as planned then it will receive Stage 2 acceptance and be published regardless of whether or not the primary hypotheses are supported (in some cases acceptance is contingent on other criteria, such as avoiding floor and ceiling effects, passing manipulation checks, etc.).

In my opinion, RRs are better suited for some kinds of projects than for others. The RR model seems particularly appropriate for large, labour-intensive hypothesis tests that require lots of resources to complete and for which results will very likely be of value whether they support the experimental hypothesis or the null hypothesis (as in a well-motivated, double-blind, randomized clinical trial). For present purposes, my main point is that preregistering a research plan does not in itself entail using the Registered Report approach to publishing. Most preregistered research projects are submitted *after* data collection, not as Registered Reports.

Preregistration has some powerful critics (e.g., Szollosi et al., 2019; cf. Nosek et al., 2019, and Wagenaar, 2019). Preregistering is more useful for hypothesis testing research than for computational modeling or hypothesis generation. Preregistration is particularly important in domains in which theories are underspecified, studies afford many researcher degrees of freedom (e.g., because a large number of variables are measured), and direct replications are rare. Preregistering a research plan does not guarantee that the research is worthwhile; excellent research can be done without preregistration, and preregistered studies can be daft. Nonetheless I believe that preregistration is a helpful practice, especially for hypothesis testing in domains that allow many researcher degrees of freedom.

4. Consider developing a lab manual. Transparency and replicability may be supported by setting standard procedures for routine tasks in your lab (e.g., planning sample size; backing up data; naming conventions for projects, files, and variables; rules for excluding subjects or observations; data-cleaning procedures; determining authorship order; resolving conflicts with other lab members, etc.). For examples, see osf.io/3jcrd/; psyarxiv.com/gxcy5/;

ccmorey.github.io/labHandbook/; https://handbook-public.themusiclab.org/. New members of your lab can be given the portions of the lab manual that pertain to their roles.

5. Make your materials, data, and analysis scripts transparent. To the extent that ethical and practical constraints allow, make it easy for readers to access your materials, de-identified data, and analysis scripts directly, rather than having to go through you for this information. This facilitates direct replications, attempts to reproduce your analyses, and explorations of the robustness of your claims across alternative analyses. As per the FAIR principles, make your materials, data, and analysis Findable, Accessible, Interoperable, Reusable.

Researchers have a primary responsibility to protect the wellbeing of their participants, and it is not always ethical to post data on the open web. But it is often possible to find ethical ways of sharing data with other scientists. In some cases, for example, it might be appropriate to post data on a Protected Access site that is moderated by a third party (https://osf.io/tvyxz/wiki/8.%20Approved%20Protected%20Access%20Repositories/). For further information about this complex topic, please see Michelle Meyer's (2018) very helpful article.

Sometimes the findings reported in an article are based on a small subset of data from a larger study. The desired practice is to give other scientists easy access to the data upon which the analyses reported in an article were based. That does not require sharing other raw data that were collected as part of the same project (although it might be appropriate to disclose the existence of the other data).

The recommendation here is to make directly accessible to other researchers the processed data on which statistical scripts were run, so that researchers can examine them, reproduce the analyses, and explore alternative analyses. Often, such data are not the rawest form of the data. A researcher might, for example, have video recorded participants performing a task and then scored aspects of their behaviour and analyzed those scores. In that situation, making the scores directly

accessible to other researchers would meet criteria for open data (although, of course, when ethical and feasible it would be great if other researchers could also directly access the videos so as to be able to reproduce the original scoring or explore alternative scoring schemes).

In my experience, many researchers are reluctant to give readers direct access to materials, analysis scripts, and data. They prefer to provide such information on request. But authors often fail to comply with such requests. Wicherts (2006) emailed requests for data to authors of articles in APA journals; only 27% eventually complied (see also Hardwicke & Ioannidis, 2018). Kidwell et al. (2016) attempted to obtain and reproduce the analyses of data from a (smallish) set of recently published articles that claimed that data were available on request. As shown in Figure 2, except for articles published in *Psychological Science* that earned a data badge (the blue line), it was rare for authors to provide complete and usable data. And these authors were still alive.

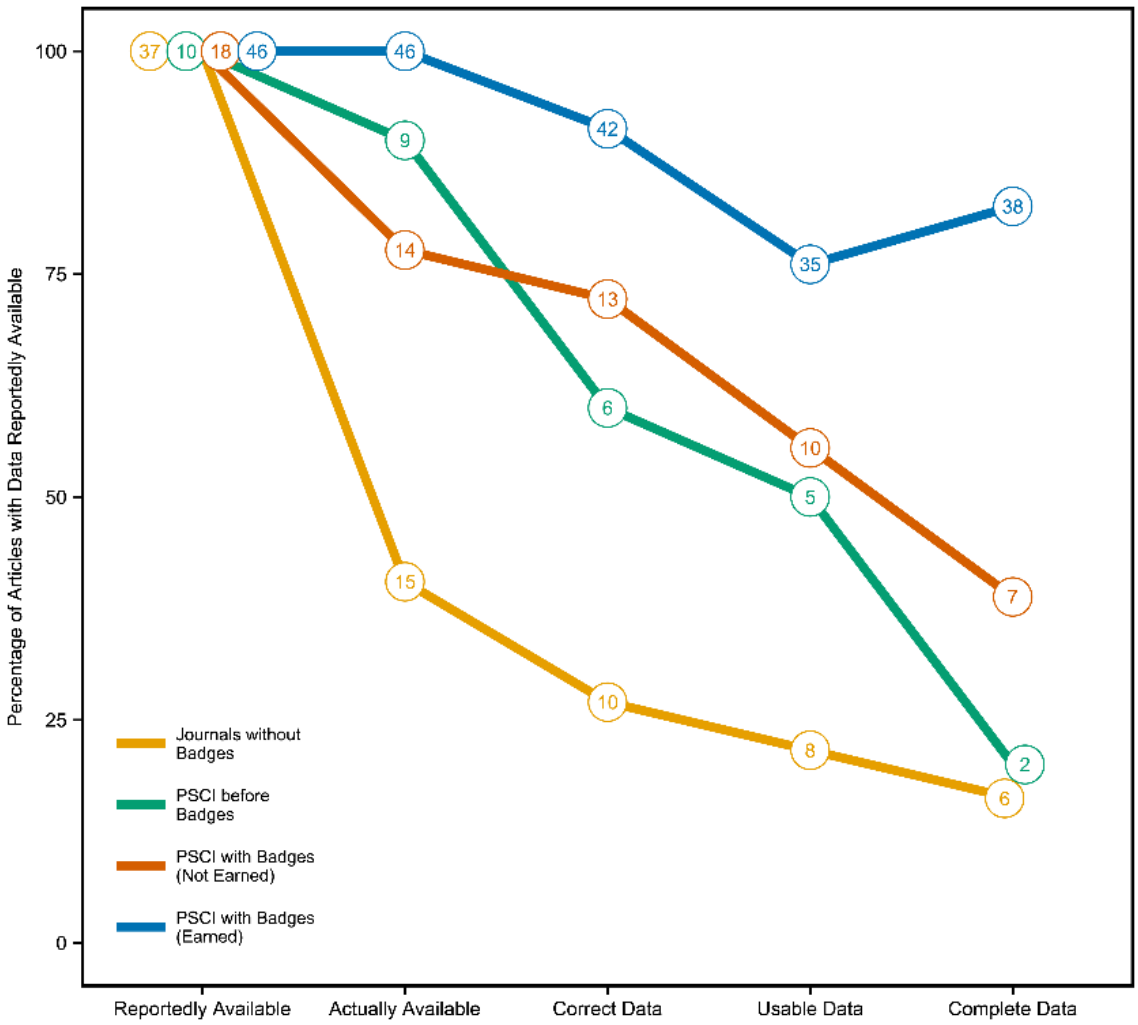

Figure 2. From Kidwell et al. (2016), who examined articles published in psychology journals between 2012 and 2015, with a focus on *Psychological Science* before and after that journal adopted data badges midway through 2015 to encourage authors to share data (see Eich, 2014). For comparison, articles published during the same period in four other prestigious psychology journals were also examined. The graph depicts measures for articles that explicitly claimed that data were available; the numerals in circles indicate the number of articles (e.g., of 37 articles published in the comparison journals that explicitly claimed data were available, only 6 provided Kidwell et al. with complete data).

It may be that one reason researchers are reluctant to invest time and effort in preparing their data, analysis scripts, and materials for sharing is because they think the labour is likely to be in vain. I suspect that most of us rarely receive requests for such information, so we may assume that few scientists would avail themselves of our materials and data if we took

the trouble to make them directly accessible. But I believe psychologists would more often scrutinize one another's data, analysis scripts, and materials if it was easy to access them. As one piece of anecdotal evidence supporting that belief, I point to a half dozen Corrigenda and a couple of Retractions of articles in *Psychological Science* that came about during my editorship because authors had posted their data and/or scripts (enabling others to discover errors therein). It is not fun to have errors in your work brought to light, but it is better than having them stay in the literature. The value of science is overturned if researchers value appearance over truth. Scientists who set the record straight when they learn of errors are to be lauded.

I hope that in future it will be normative to reward researchers for providing data and materials that are used by other researchers. Perhaps, for example, CVs might include a section listing articles by other researchers who used one's data or materials.

6. Address constraints on the generality of your findings. Currently dominant incentives encourage researchers to imply boundless generality because a finding seems more consequential if there are no limits on the conditions under which it occurs. But when a published finding fails to replicate in a follow-up study, the authors of the original work often attribute the failure to differences between the original and replication procedures, even though the original work did not hint that such differences would modulate the effect. Here again, I think that we need to somehow shift the culture away from overclaiming. If you are aware of (or have suspicions about) constraints on the generality of your findings, make them clear (see Simons, Shoda, & Lindsay, 2016). If you are a reviewer or editor, encourage authors to be appropriately circumspect in the ways they generalize their findings.

7. Consider collaborative approaches to conducting research. This recommendation is particularly relevant if the nature of your research makes it difficult to conduct high-powered tests of hypotheses. An example is research comparing neurotypical children and children diagnosed with Fetal Alcohol Syndrome (FAS). It is difficult to test young children on many trials (reducing measurement reliability) and it is often not feasible for a lone researcher to collect large samples of children diagnosed with FAS. As shown above, unless true effects are very large, small samples often yield Type II errors and when they don't they often exaggerate effect size. Multiple labs working together can mitigate these problems, as shown by the trail-blazing ManyBabies project led by Michael Frank (ManyBabies Consortium, in press). The Psychological Science Accelerator is a platform for distributing data collection across an international consortium of labs. StudySwap is another web-based platform for exchanging data-collection with other researchers. Such collaborative projects foster transparency, increase statistical power, and can help assess generality/robustness of findings across labs.

Conclusion

I have argued that psychological scientists should strive to tell the truth; upgrade their statistical knowledge; standardize aspects of their research practices; document lab procedures in a lab manual; make materials, data, and analysis scripts transparent; address constraints on generality; and collaborate with other scientists. Other proponents of methodological reform might emphasize different steps (indeed, while working on this paper I read an excellent 2019 article in which Crüwell et al. proposed an overlapping but somewhat different "seven steps"). And in truth these are not "steps," but rather disciplines, practices, even aspirations. Probably you already do some or all of these to some extent. I'm still working on them myself (e.g., as of

today “my” lab manual is a work in progress drafted by doctoral student Kaitlyn M. Fallow). My claim is that most of us would do better science if we worked on these practices.

There is no one-size-fits-all recipe for doing good science. The steps I have emphasized pertain particularly to hypothesis testing. Not all science involves hypothesis testing; exploratory research, hypothesis generation, theory development, modeling, descriptive research, etc. are crucial to science.

Even within the realm of hypothesis testing, science (like the rest of life) is full of trade-offs. Some important hypotheses cannot feasibly be tested with large samples, tightly controlled procedures, highly reliable measures, etc. For example, field work often demands compromises that reduce reliability and internal validity, but that doesn’t mean psychologist shouldn’t do field work. It just means they must be appropriately modest in their conclusions.

The aim of the methodological reform movement is not to restrict psychological research to procedures that meet some fixed criterion of replicability. Replicability is not in itself the goal of science. Rather, the central aim of methodological reform is to make research reports more transparent, so that readers can gain an accurate understanding of how the data were obtained and analyzed and can therefore better gauge how much confidence to place in the findings. A second aim is to discourage practices that contribute to effect-size exaggeration and false discoveries of non-existent phenomena. As per Vazire’s analogy, the call is not for car dealerships to sell nothing but new Ferraris, but rather for dealers to be forthcoming about the weaknesses of what they have on the lot. The grand aim of science is to develop better, more accurate, and more useful understandings of reality. Methodological reform cannot in and of itself deliver on that goal, but it can help.

References

Aczel, B., Szaszi, B., Sarafoglou, A. et al. (2020). A consensus-based transparency checklist. Nat Hum Behav 4, 4–6. https://doi.org/10.1038/s41562-019-0772-6

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*, 1547–1562.

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425.

Bishop, D. V. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology, 73*, 1–19. https://doi.org/10.1177/1747021819886519

Camerer, C.F., Dreber, A., Holzmeister, F. et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat Hum Behav 2*, 637–644 (2018). https://doi.org/10.1038/s41562-018-0399-z

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*, 115–144.

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science, 2*, 233–239. https://doi.org/10.1177/2515245919858072

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Corker, K. S.(in press). Strengths and weaknesses of meta-analysis. In L. Jussim, S. Stevens, & J. Krosnick (Eds.) *Research integrity in the behavioral sciences*. Draft preprint at https://psyarxiv.com/6gcnm/

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., … Schulte-Mecklenbeck, M. (2019). Seven easy steps to open science: An annotated reading list. *Zeitschrift für Psychologie*, *227,* 237-248. doi.org/10.1027/2151-2604/a000387.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

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

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019) Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS ONE 14*: e0216125. https://doi.org/10.1371/journal.pone.0216125

Eich, E. (2014). Business not as usual. *Psychological Science, 25*, 3–6.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191

Flake, J. K., & Fried, E. I. (2019, January 17). Measurement schmeasurement: Questionable measurement practices and how to avoid them. https://doi.org/10.31234/osf.io/hs7wm

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*(2), 151–156.

Frith, Uta. (2019). Fast lane to slow science. *Trends in Cognitive Sciences*, *24*, 1-2. 10.1016/j.tics.2019.10.007.

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly cited psychology and psychiatry articles. *PLoS ONE, 13*(8).

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Kidwell M. C., Lazarević, L. B., Baranski, E., Hardwicke, T.E., Piechowski, S., Falkenberg. L-S. et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biol 14*: e1002456. https://doi.org/10.1371/journal.pbio.1002456

Lindsay, D. S., Simons, D. J, & Lilienfeld, S. O. (2016). Research preregistration 101. APS Observer. https://www.psychologicalscience.org/observer/research-preregistration-101

Mellor, D., Vazire, S., & Lindsay, D. S. (2018). Transparent science: A more credible, reproducible, and publishable way to do science. Chapter to appear in R. J. Sternberg (Ed.), Guide to publishing in psychology journals (2nd ed). Cambridge University Press.

Makin, T. R., & Orban de Xivry, J. J. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. eLife 2019;8:e48175 DOI: 10.7554/eLife.48175

ManyBabies Consortium. (in press).). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*. Preprint at https://psyarxiv.com/s98ab/

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*, 131–144.

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

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534.

Nosek, B., Beck, E., Campbell, L., Flake, J., Hardwicke, T., Mellor, D., van 't Veer, A., & Vazire, S. (2019). Preregistration is hard, and worthwhile. https://doi.org/10.1016/j.tics.2019.07.009

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. (2015). *Science*, *349*(6251), 1–8.

Pexman, P., & Jamieson, R. (in press). Moving beyond 20 questions: We (still) need stronger psychological theory. *Canadian Psychology.*

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00813

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*, 559–569.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10*, 886-899.

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2019, October 31). Is preregistration worthwhile?. https://doi.org/10.1016/j.tics.2019.11.009

Vazire, S. (2017). Quality uncertainty erodes trust in science . **Collabra, 3**.

Wagenaar, E.J. (2019). A breakdown of Preregistration is Redudant at Best. www.bayesianspectacles.org/a-breakdown-of-preregistration-is-redundant-at-best/

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728.

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*.