The failure of null hypothesis significance testing when studying incremental changes, and what to do about it*

Andrew Gelman†

3 Jul 2017

Abstract

A standard mode of inference in social and behavioral science is to establish stylized facts using statistical significance in quantitative studies. However, in a world in which measurements are noisy and effects are small, this will not work: selection on statistical significance leads to effect sizes which are overestimated and often in the wrong direction. After a brief discussion of two examples, one in economics and one in social psychology, we consider the procedural solution of open post-publication review, the design solution of devoting more effort to accurate measurements and within-person comparisons, and the statistical analysis solution of multilevel modeling and reporting all results rather than selection on significance. We argue that the current replication crisis in science arises in part from the ill effects of null hypothesis significance testing being used to study small effects with noisy data. In such settings, apparent success comes easy but truly replicable results require a more serious connection between theory, measurement, and data.

1. The problem

A standard mode of inference in social and behavioral science is to establish stylized facts using statistical significance in quantitative studies. A "stylized fact"—the term is not intended to be pejorative—is a statement, presumed to be generally true, about some aspect of the world. For example, the experiments of Stroop and of Kahenman and Tversky established stylized facts about color perception and judgment and decision making. A stylized fact is assumed to be replicable, and indeed those aforementioned classic experiments have been replicated many times. At the same time, social science cannot be as exact as physics or chemistry, and we recognize that even the most general social and behavioral rules will occasionally fall. Indeed, one way we learn is by exploring the scenarios in which the usual laws of psychology, politics, economics, etc., fail.

The recent much-discussed replication crisis in science is associated with many prominent stylized facts that have turned out not to be facts at all (Open Science Collaboration, 2015, Jarrett, 2016, Gelman, 2016b). Prominent examples in social psychology include embodied cognition, mindfulness, and ego depletion, as well as sillier examples such as the claim that beautiful parents are more likely to have daughters, or that women are three times more likely to wear red at a certain time of the month.

These external validity problems reflect internal problems with research methods and the larger system of scientific communication. Stylized facts are supposed to be generally true, but they are typically studied in narrow laboratory-like environments on non-representative samples of people. Statistical significance and p-values, which are meant to screen out random patterns and assure near-certain conclusions, instead get used to draw inappropriate confidence from noisy data. Peer

*For *Personality and Social Psychology Bulletin*. Some of this material appeared in blog posts cited in the references. We thank Chris Crandall for helpful comments and the U.S. National Science Foundation, Institute of Education Sciences, Office of Naval Research, and Defense Advanced Research Projects Agency for partial support of this work.

†Department of Statistics and Department of Political Science, Columbia University, New York

review often seems merely to exacerbate these problems when placed in the hands of ambitious public relations entrepreneurs. And follow-up research has often failed too: until recently, it has been difficult to publish direct replications of published work, with researchers instead performing so-called conceptual replications which are subject to all the same replication problems as the original work. This is how a paper such as Bargh, Chen, and Burrows (1996) could be cited 4000 times and spawn an entire literature—and then turn out to fail under attempted independent replication.

The immediate puzzle of the replication crisis is: how have researchers been able to so regularly obtain statistical significance when studying such ephemeral phenomena? The answer, as pointed out by Simmons, Nelson, and Simonsohn (2011) and others, is that low p-values are easily obtained via "p-hacking," "harking," and other "questionable research practices" under which data-processing and data-analysis decisions are performed after the data have been seen.

At this point it is tempting to recommend that researchers just stop their p-hacking. But unfortunately this would not make the replication crisis go away! The problem is that if you are studying small, highly variable phenomena with noisy measurements, then summaries such as averages, comparisons, and regression coefficients will be noisy. If you report everthing you see you'll just have a pile of noise, and if you condition on statistical significance you'll drastically overestimate effects and often get their signs wrong (Gelman and Carlin, 2014). So eliminating p-hacking is not much of a solution if this is still happening in the context of noisy studies.

Null hypothesis significance testing (NHST) only works when you have enough accuracy that you can confidently reject the null hypothesis. You get this accuracy from a large sample of measurements with low bias and low variance. But you also need a large effect size. Or, at least, a large effect size, compared to the accuracy of your experiment.

But we've grabbed all the low-hanging fruit. In medicine, public health, social science, and policy analysis we are studying smaller and smaller effects. These effects can still be important in aggregate, but each individual effect is small.

The NHST approach as currently practiced has (at least) four serious problems:

1. Overestimation of effect sizes. The "statistical significance filter," by which estimates are much more likely to be reported if they are two standard errors from zero, introduces a bias which can be huge.

2. Estimates in the wrong direction.

3. Extraction of statistically significant patterns from noise.

4. Incentives for small samples and noisy measures.

Null hypothesis significance testing in psychology has been criticized for a long time (for example, Meehl, 1967, Krantz, 1999), but in recent years the crisis has been taken more seriously with recognition of the above four issues.

2. Two examples

We demonstrate the problems with standard practice in two recent articles: a economic policy analysis in education, and an experimental paper in social psychology, in both cases published in leading scientific journals. We chose these two papers not as a representative sample of published work in economics and psychology but rather to indicate the fundamental misunderstandings held even by respected scholars in their fields. As McShane and Gal (2017) demonstrate, similar errors are held by statisticians as well.

2.1. Biased estimation in policy analysis

Summarizing the result of an experiment on early-childhood intervention, Gertler et al. (2014) wrote:

> "We report substantial effects on the earnings of participants in a randomized intervention conducted in 1986–1987 that gave psychosocial stimulation to growth-stunted Jamaican toddlers. . . . the intervention had a large and statistically significant effect on earnings. . . . The estimated impacts are substantially larger than the impacts reported for the US–based interventions, suggesting that ECD interventions may be an especially effective strategy for improving long-term outcomes of disadvantaged children in developing countries."

The problem here can be seen in the phrase "large and statistically significant effect." This particular study was performed on only 129 children; the outcome is inherently variable; hence standard errors will be high; there are many researcher degrees of freedom by which statistical significance can be obtained comparisons are much more likely to be published when statistically significant; hence published results are biased from the statistical significance filter, and this bias can be huge.

The bias is also potentially consequential for policy recommendations. Consider the last sentence of the above quote, which is making a substantive claim based on the point estimate being large, without any correction for the bias of that estimate.

Such adjustment is not trivial, as the size of the bias depends on the magnitude of the underlying effect as well as on the uncertainty in the point estimates. Consider the example of a normally-distributed effect-size estimate with standard error of 12% (the approximate value from the Gertler et al. paper, whose 25% estimate just reached the statistical significance threshold). Following Gelman and Carlin (2014) we can use the properties of the normal distribution to determine the expected estimate of the magnitude of the effect, conditional on statistical significance.

The calculation goes like this: We start with a hypothesized true effect size θ, then take the normal distribution of point estimates $\hat{\theta}$ (in this case, assumed to have mean θ and standard deviation 0.12), and then consider the subset of these estimates that are at least two standard errors from zero (thus, those cases where $|\hat{\theta}| > 0.24$). From this conditional distribution we can compute $\mathrm{E}(|\hat{\theta}|, \text{ given } |\hat{\theta}| > 0.24)$, the expected magnitude of the point estimate under selection for statistical significance. This value depends on the true effect size, so we can graph it as a function of θ.

Figure 1 shows the results. For any reasonable underlying effect size estimate, the bias is huge. For example, if the true benefit of early childhood intervention in this population is 5% (not a trivial bump in earnings), then the expected magnitude of the effect-size estimate is 29%, for a bias of 24%. If the true benefit is 10%, the bias is 19%. Even if the true benefit is an (a priori implausibly large) 25%, the estimate, conditional on statistical significance, has a 9% bias. As Figure 1 demonstrates, the bias approaches zero only when the true effect size exceeds an implausible 50% in adult earnings from that intervention on four-year-olds.

The size of the bias depends on the unknown parameter value, so any bias correction would rely on assumptions. But making such an assumption would be better than implicitly assuming a bias of zero. The trouble arises from the attitude that statistical significance assures one of a kind of safety which then allows point estimates and confidence intervals to be taken at face value. In fact, selection bias is relevant whatever the outcome of the selection.

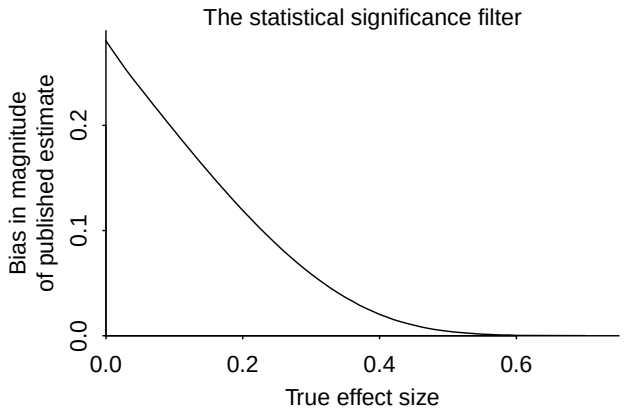

Figure 1: *Bias in expected magnitude of effect size estimate, conditional on statistical significance, as a function of actual effect size, for the early-childhood intervention study of Gertler et al. (2014). The raw estimate, before selection is assumed to be normally distributed with mean equal to the true effect and standard error 0.12.*

2.2. Forking paths in social psychology

In the article, "Caught Red-Minded: Evidence-Induced Denial of Mental Transgressions," Burum, Gilbert, and Wilson (2016) follow a standard practice in psychology research by making general claims based on three lab experiments conducted on a couple hundred college students. The p-values of the main comparisons are 0.04, 0.03, and 0.06, but these were chosen out of a large number of potential comparions and in the presence of many open-ended data exclusion rules and other researcher degrees of freedom.

There is no reason to expect the true effect sizes in such experiments to be large, and even less do we expect arbitrary comparisons within these experiments to represent large underlying effects. As a result, just for purely mathematical reasons an attempt to learn from such data via p-values is doomed to fail.

Consider, for example, this passage from Burum et al. (2016), which demonstrates the general mode of reasoning based on statistical significance:

> "The only other measure that differed between conditions was the report of the victim's attractiveness. Although the immediate evidence and delayed evidence conditions did not differ on this measure, $t(37) = 0.88$, $p = .385$, participants in the no evidence condition did find the victim more attractive than did participants in either the immediate evidence condition, $t(37) = 3.29$, $p = .002$, or the delayed evidence condition, $t(34) = 2.01$, $p = .053$. This may be because participants in the no evidence condition answered this question closer to the time that they last saw the victim than did participants in any other condition. The magnitude of the means on the remaining measures suggests that participants found the victim attractive, considered the crime very serious, felt uneasy about watching the video, and believed that pupillary dilation and eyeblink rate provide suggestive but not conclusive evidence of sexual arousal."

This sort of argument is little more than chasing of noise. With small effects and high variation, one can easily have real differences appear to be zero ($p = 0.385$), just as, with selection, one can see many null differences that are significant at the $p = 0.05$ level. And the claim of "suggestive

but not conclusive evidence" is just bizarre as this was not addressed in any way by the questions in that study. The real lesson to be learned here is that a well-ordered pile of numbers offers nearly unlimited scope for storytelling (Coyne, 2017).

At this point, one might argue that this work has value as speculation, and that might be. We don't see much interest in a claim such as, "under some circumstances, confronting people with public evidence of their private shortcomings can be counterproductive," but perhaps it has value in the context of the literature in its subfield. But, if so, this would be just as legitimately interesting if presented as theory + speculation + qualitative observation, without the random numbers that are the quantitative results, the meaningless p-values and all the rest. The work should stand or fall based on its qualitative contributions, with the experimental data explicitly recognized as being little more than a spur to theorizing.

One of the authors of that paper was associated with a press release that claimed that "the replication rate in psychology is quite high—indeed, it is statistically indistinguishable from 100%" (Ruell, 2016). The paper discussed above featured a sort of replication (study 3), but it was not preregistered and the resulting p-value still exceeded 0.05, thus demonstrating yet another forking path in that had non-significance been the desired result, that could have been claimed too. One way to get a replication rate of 100% is to have freedom to decide what is or is not a replication.

3. Potential solutions

To study smaller and smaller effects using NHST, you need better measurements and larger sample sizes. The strategy of run-a-crappy-study, get p less than 0.05, come up with a cute story based on evolutionary psychology, and PROFIT . . . well, it doesn't work anymore. OK, maybe it still can work if your goal is to get published in PPNAS, get tenure, give Ted talks, and make boatloads of money in speaking fees. But it won't work in the real sense, the important sense of learning about the world.

One of the saddest aspects of all this is seeing researchers jerked around like puppets on a string based on random patterns from little experiments. One could spend an entire career doing this. It is the duty of statisticians not just to criticize but to explain how these patterns of behavior can arise and perpetuate themselves, so future researchers don't make the same mistakes.

How can we do better?

3.1. Procedural solutions

The current system of scientific publication encourages the publication of speculative papers making dramatic claims based on small, noisy experiments. Why is this? To start with, the most prestigious general-interest journals—Science, Nature, and PNAS—require papers to be short, and they strongly favor claims of originality and grand importance, thus favoring "high concept" studies (for example, that election outcomes are determined by college football games, a claim that was disputed by Fowler and Mongagnes, 2015). Second, given that the combination of statistical significance and a good story is sufficient for publication, and given that statistical significance is easy to come by with small samples and noisy data (Simmons, Nelson, and Simonsohn, 2011), there is every motivation to perform the quickest, cheapest study and move directly to the writeup, publication, and promotion phases of the project.

How, then, to change the incentives? One approach is to move from a flagship-journal model of closed pre-publication review to an Arxiv and PubPeer model of open post-publication review. A continuing and well-publicized ongoing review process creates a negative incentive for sloppy work,

because if your unfounded claims get published in Science or Nature, and then later fail to replicate or are otherwise shown to have serious flaws, you pay the reputational cost. This is standard economics, to move from a one-shot to a multiple-round game.

Another advantage of post-publication review is that its resources are channeled to articles on important topics (such as education policy) or articles that get lots of publicity (such as the flawed recent claim that there is a maximum limite on human ages; see, for example, Devlin, 2017). In contrast, with regular journal submission, every paper gets reviewed, and it would be a huge waste of effort for all these papers to be carefully scrutinized. We have better things to do. This is an efficiency argument. Reviewing resources are limited (recall that millions of scientific papers are published each year) so it makes sense to devote them to work that people care about.

Another set of procedural reviews are focused more closely on replication and might be particularly appropriate to fields such as experimental psychology where replication is relatively inexpensive. From one direction, there has been a move to encourage authors to preregister their plans for data collection, data processing, and data analysis plans (Simmons, Nelson, and Simonsohn, 2012), or to accept papers based on their research design so there is no longer an implicit requirement that a study be a "success" to be published.

From the other direction, one can make it easier for outsiders to publish replications and criticisms; publish these in the same journal as published the original paper. Online there should be no problem with space restrictions, and if publication credit is an issue, one can give the citation some distinguishing name such as Replications in Psychology so that replications will not be confused with original research.

And not all criticism should be thought of as "debunking." Consider the two papers discussed above. Gertler et al. present a biased estimate, and we should think of the work is strengthened, not weakened, by acknowledging and attempting to correct this bias. The work of Burum et al. is exploratory, and its contributions would be much clearer if the raw data were presented and if results were not characterized based on p-values. The authors of these and similar papers should appreciate this statistical guidance, and the expectation of open post-publication review should motivate future researchers to anticipate such criticisms and avoid the errors arising from selecting on statistical significance.

3.2. Solutions based on design and data collection

The next set of solutions arise within an individual study or experiment. A standard recommendation is larger sample size, but we think that an even better focus would be on quality of measurement. Part of this is simple mathematics: a reduction of measurement error by a factor of 2 is as good as multiplying sample size by 4. Even more, though, we should be concerned with bias as well as variance. A notorious recent example came from researchers studying ovulation, who characterized days 6–14 as the most fertile time of a woman's period, even though the standard recommendation from public heath officials is days 10–17 (Gelman, 2014a). Our point here is that (1) a more careful concern with measurement could have led to a more careful literature review and the use of the more accurate interval, and (2) no increase in sample size would correct for this bias.

The data in this particular example were collected in a one-shot survey; presumably the data would've been more accurate had they been collected in diary form, but that would have required more effort, both for the participants and for the researchers. That's the way it goes: getting better data can take work. Once the incentives have been changed to motivate higher-quality data collection, it can make sense to think about how to do this.

A related step, given that we're already talking about getting more and better information

from individual participants in a study, is to move from between-person to within-person designs (Gelman, 2016a).

When studying the effects of interventions on individual behavior, the experimental research template is typically: Gather a bunch of people who are willing to participate in an experiment, randomly divide them into two groups, assign one treatment to group A and the other to group B, then measure the outcomes. If you want to increase precision, do a pre-test measurement on everyone and use that as a control variable in your regression. But here we argue for an alternative approach: study individual subjects using repeated measures of performance, with each one serving as his or her own control.

As long as your design is not constrained by ethics, cost, realism, or a high drop-out rate, the standard randomized experiment approach gives you clean identification. And, by ramping up your sample size, you can get all the precision you might need to estimate treatment effects and test hypotheses. Hence, this sort of experiment is standard in psychology research and has been increasingly popular in political science and economics with lab and field experiments.

However, the clean simplicity of such designs has led researchers to neglect important issues of measurement, as pointed out by Normand (2016):

> "Psychology has been embroiled in a professional crisis as of late. . . . one problem has received little or no attention: the reliance on between-subjects research designs. The reliance on group comparisons is arguably the most fundamental problem at hand . . .
>
> But there is an alternative. Single-case designs involve the intensive study of individual subjects using repeated measures of performance, with each subject exposed to the independent variable(s) and each subject serving as their own control."

Why would researchers ever use between-subject designs for studying within-subject phenomena? We see several reasons:

- The between-subject design is easier, both for the experimenter and for any participant in the study. You just perform one measurement per person. No need to ask people a question twice, or follow them up, or ask them to keep a diary.

- Analysis is simpler for the between-subject design. No need to worry about longitudinal data analysis or within-subject correlation or anything like that.

- Concerns about poisoning the well. Ask the same question twice and you might be concerned that people are remembering their earlier responses. This can be an issue, and it's worth testing for such possibilities and doing your measurements in a way to limit these concerns. But it should not be the deciding factor. Better a within-subject study with some measurement issues than a between-subject study that's basically pure noise.

- The confirmation fallacy. Lots of researchers think that if they've rejected a null hypothesis at a 5% level with some data, that they've proved the truth of their preferred alternative hypothesis. Statistically significant, so case closed, is the thinking. Then all concerns about measurements get swept aside: After all, who cares if the measurements are noisy, if you got significance? Such reasoning is wrong but one can see its appeal.

One motivation for between-subject design is an admirable desire to reduce bias. But we shouldn't let the apparent purity of randomized experiments distract us from the importance of careful measurement. Real-world experiments are imperfect—they do have issues with ethics, cost, realism, and dropout, and the strategy of doing an experiment and then grabbing statistically-significant comparisons can leave a researcher with nothing but a pile of noisy, unreplicable findings.

3.3. Improved statistical analysis

Finally, once the data have been collected, in whatever form, they can be analyzed better. We have already railed against null hypotheis significance testing, which creates incentives to distort data in order to reach magic thresholds, and, even in the absence of any over p-hacking, leads to biased estimates and overconfidence.

The standard approach to multiple comparisons is to report the largest or most significant comparison and then adjust for multiplicity, or to rank all comparisons by statistical significance and then report the ones that exceed such threshold. Such procedures can make sense in so-called needle-in-haystack problems where there is some small number of very large effects surrounded by a bunch of nulls, a situation that could arise in genetics, for example. But in psychology or political science or economics, we think it generally makes much more sense to think of there being a continuous distribution of effects, in which case it is highly wasteful of information to focus on the largest or the few largest of some set of noisy comparisons.

What, then, should be done instead? To start with, display as much of the data as possible. And if there are concerns about researcher degrees of freedom, perform *all* reasonable possible analyses, what Steegen et al. (2016) call the "multiverse." The point of the multiverse analysis is not to get better p-values but rather to recognize that all these possible analyses are legitimately of interest.

Rather than pulling out individual comparisons, we recommend multilevel modeling (Gelman, Hill, and Yajima, 2011). Take advantage of the fact that lots and lots of these studies are being done. Forget about getting definitive results from a single experiment; instead embrace variation, accept uncertainty, and learn what you can. In a multilevel model, parameters are estimated in groups—for example, instead of independently estimating many different possible interactions and checking the statistical significance of each, you would estimate the distribution of interaction effects, and then the estimate of any particular interaction would be partially pooled toward the larger model.

Another way forward uses Bayesian inference. This is controversial because of the need for a prior distribution, and many researchers will prefer to use purely data-based estimates. However, in settings with weak data and strong prior information, an unadjusted data summary can be little more than noise. For example, consider a much-publicized study finding that more attractive parents were more likely to have girl babies, a result obtained from a survey of 3000 Americans. From such a study one can estimate a difference in proportions to an accuracy of approximately 2 percentage points (from the formula for the variance of the binomial distribution, a proportion from a sample of 1500 can be estimated with standard deviation $\sqrt{0.5 * 0.5/1500} = 0.013$, hence the difference between two independent proportions each of size 1500 has standard deviation $\sqrt{0.013^2 + 0.013^2} = 0.018$). But from the literature we could expect the sex ratios for two populations divided by a crude measure of attractiveness to differ by less than one-tenth of one percentage point (Gelman and Weakliem, 2009). In this example, the prior information is something like 20 times stronger than the data. Trying to use a survey of 3000 people to estimate tiny differences in sex ratios: this makes about as much sense as using a bathroom scale to weigh a feather, when that feather is resting loosely in the pouch of a kangaroo that is vigorously jumping up and down.

Any reasonable Bayesian analysis of the sex-ratio example would discount the data from the small sample so much that it would be clear that essentially nothing can be learned from these data; as a practical matter, 95% posterior intervals for the population difference would almost certainly comfortably contain zero, even if the raw data were to show a difference that happened to be more than two standard errors from zero. In other settings the result would be more ambiguous, and Bayesian analysis shades toward multilevel modeling and reproducibility, in the following sense: the prior represents the distribution of true effects across a range of conditions, which can be identified

as the set of hypothetical ideal experiments over which the phenomenon of interest would be studied. Thus we appreciate the Bayesian formulation both for its practical benefits (see, for example, Ghitza and Gelman, 2013) and because of the mapping from questions about an appropriate prior distribution, to discussion of the range of applicability of a study.

4. Discussion

Ironically, classical statistical procedures are often thought of as safe choices, with null-hypothesis significance testing offering protection from drawing conclusions from noise, and classical point estimation offering unbiased inference. In practice, though, null hypothesis significance testing is used in a confirmatory fashion (Gelman, 2014b), yielding overestimation of effect sizes and overconfidence in replicability.

The result is some mix of misplaced trust in noisy claims, promulgation of exaggerated estimates in policy advocacy, and as a natural reaction, a general attitude of distrust in quantitative social research. For example, in the past few years, traditionally prestigious journals such as the Journal of Personality and Social Psychology, Psychological Science, and the Proceedings of the National Academy of Sciences have published a series of papers that can fairly be described as junk science, leveraging statistically significant p-values to make scientifically dubious assertions.

What is striking about the two examples discussed in Section 2 is how *avoidable* these errors have been. Gertler et al. (2014) presented, without comment, severely biased estimates of treatment effects—even though, as economists, the authors of this paper had the training to recognize and attempt to correct for this bias. In the field of psychology, Burum et al. (2016) followed nearly every step of a previously published satirical blog post (Zwaan, 2013), following an already discredited model of publishing statistically significant comparisons obtained by sifting through small samples of noisy data.

To move forward, we must operate on several fronts. At the systemic level, we should make it easier to publish solid work, and facilitate the publication of criticisms and replications, which should in turn reduce the incentives for overinterpretation of noise. There should be a stronger focus on collecting accurate and relevant data, and an openness to designs that allow within-person comparisons, even if this represents more effort in data collection. Finally, raw data should be shared as much as possible, and analyses should use all the data rather than jumping from one p-value to another. There should be more acceptance of uncertainty rather than a jockeying to present conclusions as more solid than they actually are.

All these lessons are generic and could've been made at any time during the past hundred years. But they are particularly relevant now, in part because of the explosion of scientific publication in recent years and in part because most science is incremental. Null hypothesis significance testing has perhaps never been a very good idea, when studying small effects and complex interactions it is particularly useless. Or, one might say, it has been a useful way for some people to get publications and publicity but not a useful way of performing replicable science. In this paper we have argued that the current replication crisis in science arises in part from the ill effects of NHST being used to study small effects with noisy data. In such settings, apparent success comes easy but truly replicable results require a more serious connection between theory, measurement, and data.

We also emphasize that these solutions are technical as much as they are moral: if data and analysis are not well suited for the questions being asked, then honesty and transparency will not translate into useful scientific results (Gelman, 2017). In this sense, a focus on procedural innovations or the avoidance of p-hacking can be counterproductive in that it will lead to disappointment

if not accompanied by improvements in data collection and data analysis that, in turn, require real investments in time and effort.

References

Bargh, J. A., Chen, M., and Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personal and Social Psychology* **71**, 233–244.

Burum, B. A., Gilbert, D. T., and Wilson, T. D. (2016). Caught red-minded: Evidence-induced denial of mental transgressions. *Journal of Experimental Psychology: General* **145**, 844–852.

Coyne, J. C. (2017). A bad abstract is good enough to be published in Journal of Experimental Psychology: General. Quick Thoughts, 16 Mar. https://jcoynester.wordpress.com/2017/03/16/a-bad-abstract-is-good-enough-to-be-published-in-journal-of-experimental-psychology-general/

Devlin, H. (2017). Maximum human lifespan could far exceed 115 years – new research. *Guardian*, 28 Jun. https://www.theguardian.com/science/2017/jun/28/maximum-human-lifespan-new-research-mortality

Fowler, A., and Montagnes, B. P. (2015). College football, elections, and false-positive results in observational research. *Proceedings of the National Academy of Sciences* **112**, 13800–13804.

Gelman, A. (2014a). Jessica Tracy and Alec Beall (authors of the fertile-women-wear-pink study) comment on our Garden of Forking Paths paper, and I comment on their comments. Statistical Modeling, Causal Inference, and Social Science blog, 31 May. http://andrewgelman.com/2014/05/31/jessica-tracy-alec-beall-authors-fertile-women-wear-pink-study-comment-garden-forking-paths-paper-comment-comments/

Gelman, A. (2014b). Confirmationist and falsificationist paradigms of science. Statistical Modeling, Causal Inference, and Social Science blog, 5 Sept. http://andrewgelman.com/2014/09/05/confirmationist-falsificationist-paradigms-science/

Gelman, A. (2016a). Balancing bias and variance in the design of behavioral studies: The importance of careful measurement in randomized experiments. Bank Underground blog, 24 Aug. https://bankunderground.co.uk/2016/08/24/balancing-bias-and-variance-in-the-design-of-behavioral-studies-the-importance-of-careful-measurement-in-randomized-experiments/

Gelman, A. (2016b). What has happened down here is the winds have changed. Statistical Modeling, Causal Inference, and Social Science blog, 21 Sept. http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/

Gelman, A. (2017). Honesty and transparency are not enough. *Chance* **30** (1), 37–39.

Gelman, A., and Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651.

Gelman, A., and Weakliem, D. (2009). Of beauty, sex, and power: Statistical challenges in estimating small effects. *American Scientist* **97**, 310–316.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* **5**, 189–211.

Gertler P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S. M., and Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation inter-

vention in Jamaica. *Science* **344**, 998–1001.

Ghitza, Y., and Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science* **57**, 762–776.

Jarrett, C. (2016). Ten famous psychology findings that it's been difficult to replicate. British Psychological Society Research Digest, 16 Sept. https://digest.bps.org.uk/2016/09/16/ten-famous-psychology-findings-that-its-been-difficult-to-replicate/

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association* **94**, 1372–1381.

McShane, B., and Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association.*

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science* **34**, 103–115.

Normand, M. P. (2016). Less is more: Psychologists can learn more by studying fewer people. *Frontiers in Psychology* **7**: 934.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* **349**, aac4716.

Ruell, P. (2016). Researchers overturn landmark study on the replicability of psychological science. Press release, Harvard University. http://projects.iq.harvard.edu/files/psychology-replications/files/harvard_press_release.pdf?m=1456973687

Simmons, J., Nelson, L., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* **22**, 1359–1366.

Simmons, J., Nelson, L., and Simonsohn, U. (2012). A 21 word solution. *Dialogue* **26** (2), 4–7.

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**, 702–712.

Zwaan, R. (2013). How to cook up your own social priming article. Rolf Zwaan blog, 17 Sept. https://rolfzwaan.blogspot.fr/2013/09/how-to-cook-up-your-own-social-priming.html