Harvard Data Science Review • Issue 2.4, Fall 2020

Selective Inference: The Silent Killer of Replicability

Yoav Benjamini[1]

[1]**Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel**

Published on: Dec 16, 2020

DOI: https://doi.org/10.1162/99608f92.fc62b261

ABSTRACT

Replicability of results has been a gold standard in science and should remain so, but concerns about lack of it have increased in recent years. Transparency, good design, and reproducible computing and data analysis are prerequisites for replicability. Adopting appropriate statistical methodologies is another identified one, yet which methodologies can be used to enhance replicability of results from a single study remains controversial. Whereas the *p*-value and statistical significance are carrying most of the blame, this article argues that addressing selective inference is a missing statistical cornerstone of enhancing replicability. I review the manifestation of selective inference and the available ways to address it. I also discuss and demonstrate whether and how selective inference is addressed in many fields of science, including the attitude of leading scientific publications as expressed in their recent editorials. Most notably, selective inference is attended when the number of potential findings from which the selection takes place is in the thousands, but it is ignored when 'only' dozens and hundreds of potential discoveries are involved. As replicability, and its closely related concept of generalizability, can only be assessed by actual replication attempts, the question of how to make replication an integral part of the regular scientific work becomes crucial. I outline a way to ensure that some replication effort will be an inherent part of every study. This approach requires the efforts and cooperation of all parties involved: scientists, publishers, granting agencies, and academic leaders.

Keywords: *p*-value, multiple comparisons, false discovery rate, reproducibility

Visit the web version of this article to view interactive content.

Yoav Benjamini gives a talk that briefly presents the ideas in this article and its motivation.

1. The Reproducibility and Replicability Crisis

Experimental science has been based on the paradigm that a result obtained from a one-time experiment is insufficient to establish the validity of a discovery. The ultimate challenge that a discovery faces is being replicated by others. This paradigm goes back to the 17th century: Robert Boyle, the founder of modern experimental science, argued that a finding should be replicated a few times before being considered convincing by the scientific community. When the Dutch Christian Huygens noted a phenomenon related to vacuum in Amsterdam that Boyle could not replicate in his own lab, he took the trip to England to demonstrate the same phenomenon in Boyle's laboratory. The debate and demonstration have made replicability of results by others a gold standard in science (Shapin & Schaffer, 1985). How many times should a result be replicated, however, remained rather vague, until statistical science clarified the implications of the number of replications within a study with respect to the weight of evidence. The *p*-value was formally suggested by Karl Pearson

(1900) for that purpose, and was later popularized by Ronald A. Fisher as a way to measure the weight of evidence from those replications against the hypothetical explanation of mere chance. It was also Fisher (1935, p. 14) who made the connection between the *p*-value as a measure of evidence in the single study and the concern about replication across studies: "no isolated experiment, however significant by itself, can suffice for the experimental demonstration of any natural phenomenon." Offering statistical significance as formulated by the *p* value less than a threshold, Fisher further states: "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment, which will rarely fail to give us a statistically significant result." This rule for a replicated discovery has served science well for almost a century, despite the philosophical disputes surrounding it.

Serious concerns about published scientific discoveries that could not be replicated by other researchers started to surface some 25 years ago. Not surprisingly it was just as the industrialization of the scientific process took a sharp rise, with the invention of genomic tools, imaging tools, recording equipment, and the increasing ease of storing results and conducting analyses en masse. Soriç (1989, p. 608) analysed the use of regular significance testing at the 0.05 level in medical research in face of the increase in the number of tests being made, and warned that "a large part of science can thus be made untrue." Mann (1994, p. 1687) warned about the problem in behavioural genetics, where each discovery is "greeted unskeptically in the popular press." Lander and Kruglyak (1995, p. 241), at the beginning of the use of high throughput methodologies in genomics, warned that "a substantial proportion of claims cannot be replicated."

These warnings were targeted at their respective scientific audience only, and sometimes had a relevant impact. Ioannidis (2005) did much to popularize this concern by provocatively titling it "Why Most Published Research Findings Are False." He further claimed to have proved mathematically that this is indeed the case, using similar arguments to those of Soriç (1989). The paper attracted wide readership outside the scientific community and has been followed by many efforts to document and explain the problem. Major stories in the general media have used headlines such as 'Is there something wrong with the scientific method?' or 'Trouble at the lab.' In addition to these public concerns, some highly publicized cases of fraud also caught public attention. The retractions of Diederik Stapel's works because of falsifications and fabrications of data (Levelt et al., 2012,) in particular, served as the ultimate proof that indeed most of current science is false. Such concerns led to the Psychological Reproducibility Project, where the main result of each of 100 papers from three leading journals in the field was tested for replication, again by others. At the end of this effort, which started in 2011 and ended in 2015, only 34% of the studies' main results were replicated in the sense of getting a second statistically significant result (Open Science Collaboration, 2015).

Unfortunately, the terms reproducibility and replicability are often used interchangeably. Following editorials in *Biostatistics* (Diggle & Zeger, 2010; Peng, 2009), this terminology is adopted for making an important distinction: Reproducibility is a property of a study, reflecting the possibility for others to start from the original data, through the analysis, and get the same figures, tables, summaries, and conclusions as appear in

the published version. Replicability is a property of the result of a study, whereby others run the entire study: from enlisting subjects through collecting data and analyzing the results, in a similar but not necessarily identical way, yet essentially get the same results (see also *Nature*, 2013; National Academies of Sciences, 2019). An important conclusion from the distinction is that *while the reproducibility of a stand-alone study can be assured,we can only enhance its replicability.*

How can the replicability of a study be enhanced? Certainly, reproducible data analysis and computations are needed, as well as transparently designed and conducted experiments. These aspects have received much attention in editorials in *Nature* (2013) and in *Science* (McNutt, 2014) as well as an article by Nosek et al. (2015), among many others. Along with these recommendations, almost all commentaries further pointed at statistics and data analysis practices as possible causes of lack of replicability. However, it was not clear at all what the problems with current statistical practices are. This is well demonstrated by the journal *Science*, setting up at that time a statistical editorial board that has been chartered with examining submitted papers for potential statistical problems.

2. The Misguided Attack

This absence of clear identification of problems, and, hence, guidelines to their solution, was entered by the 'New Statistics' movement with its objection to statistical testing in general and the 'p values' in particular. An editorial in the journal *Psychological Science* seeks "to aid researchers in shifting from reliance on NHST" (Null Hypothesis Statistical Testing): "we have published a tutorial by Cumming (2014), a leader in the new-statistics movement." Among the latter's 20-plus principles one can find: "do not trust any p-value," "avoid using statistical significance or p-value; omit any mention of null hypothesis statistical testing," and "Routinely report 95% confidence intervals." Taking a more extreme stand, the *Journal of Basic and Applied Social Psychology* banned the use of *p*-values and discouraged the use of any statistical methods (Trafimow & Marks, 2015), taking us back to the 19th century when the results of studies were reported merely by tables and figures, with no quantitative assessment of the uncertainties involved. (For unfortunate implications of the ban on the results reported in that journal a year later, see Fricker et al., 2019.)

Unfortunately, these attacks had their impact. *Nature* published a review putting much of the blame on the practices involving the use of the *p*-value (Nuzzo, 2014). The board of the American Statistical Association (ASA) caved in to this pressure and started a yearlong discussion about the misinterpretation and misuses of the *p*-value, rather than the replicability problem in general (Wasserstein & Lazar, 2016). Indeed, it singled out the *p*-value as the cause for the replicability problems in science. It did state that the *p*-value "can be useful," but then came a list of warnings—all stated about the *p*-value though all the stated warnings are relevant to other statistical procedures as well. Opening the conclusion section with, "*In view of the prevalent misuses of and misconceptions concerning p-values some statisticians prefer to supplement or even replace p-values with other approaches,"* made it clear that the leadership of ASA backs the position of these psychology journals, as far as *p*-value and statistical significance go.

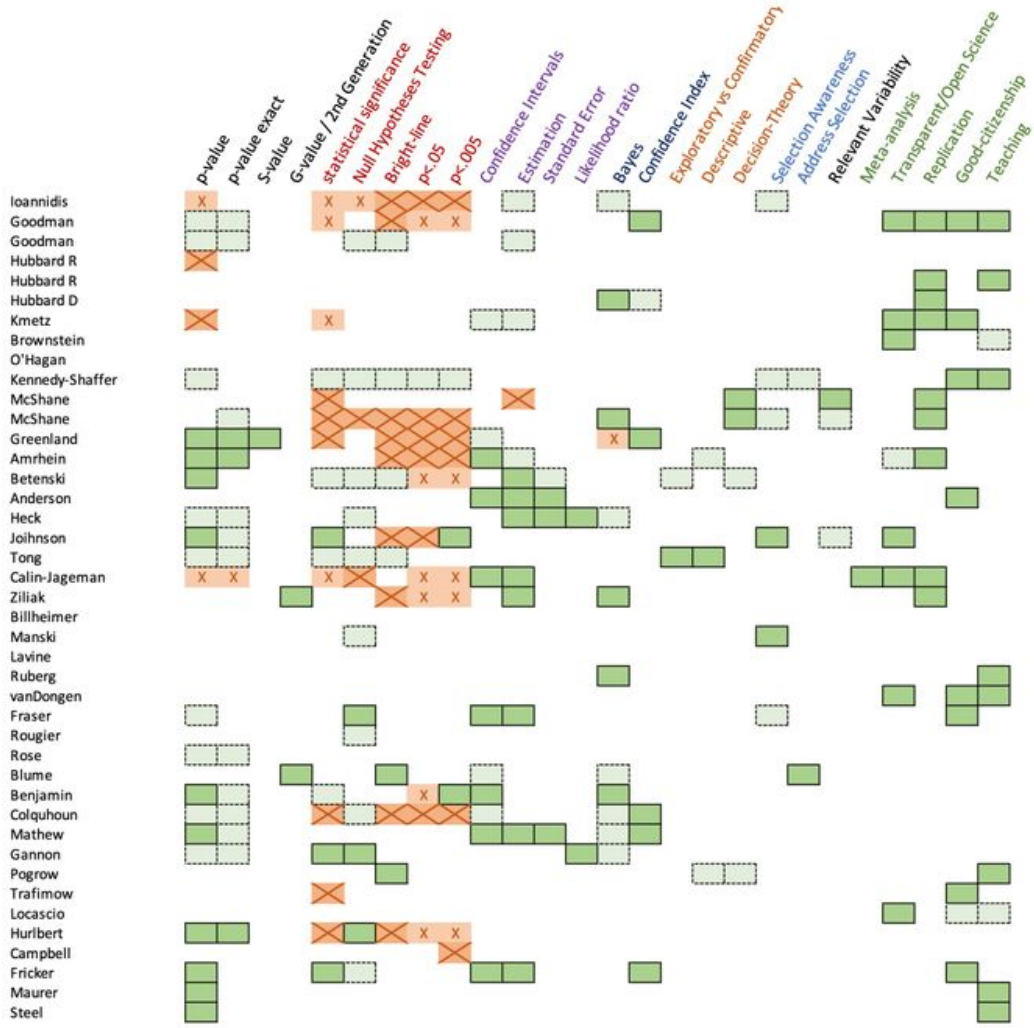
The other approaches offered in the ASA statement included the use of confidence intervals, prediction intervals, estimation, likelihood ratios, and Bayesian methods such as Bayes factors and credibility intervals. Yet all of these other approaches, as well as most statistical tools, may suffer from many of the same problems as the *p*-value does (see Benjamini, 2016). Consider 95% confidence intervals, for example. We do have a field experiment demonstrating the irrelevance of merely replacing *p*-values with confidence intervals: Research in epidemiology has been relatively a '*p*-value free zone' since some influential journal editors objected to the use of *p*-values (Rothman, 1998), and 95% confidence intervals are used instead. Obviously, because of the equivalence between a decision based on 95% confidence intervals not covering 1, and '$p < 0.05$' for testing the hypothesis of no difference in the relative risk, highlighting such intervals is identical to highlighting statistically significant results at 0.05. Nevertheless, selecting such results into the abstract is a common practice in epidemiology.

Reporting confidence intervals only did not save from over-publicity a study by Zerbo et al. (2017) about association between autism of the born child and influenza vaccinations during the pregnancy: eight different associations were assessed (16 by a different count) and it was reported in the abstract that "first-trimester influenza vaccination was the only period associated with increased autism spectrum disease risk," a finding supported by 95% confidence interval not covering 1. A warning that these results might be due to random variation plus selective inference came in the form of reporting the *p*-values adjusted for selection, using the Bonferroni procedure (see Box 1), which was well above 5%. In fact, reporting only 95% confidence intervals makes it very difficult to assess the effect of selection by a casual reader.

The ASA initiated a follow-up symposium in October 2017 on "Scientific Method for the 21st Century: A World Beyond '$p < 0.05$'," which proceeded in the direction that the ASA statement took. Its contributions were published as 43 papers in *The American Statistician* issue devoted to the topic, and summarized into recommendations by its editors with a very strong statement against statistical significant (Wasserstein et al., 2019). It was perceived as a second formal statement by ASA, even though it was not. My interpretation and compilation of the verbal recommendations into a visual display is presented in Figure 1, where each paper is assigned a row, and each column represents one or more statistical methods or actions that were recommended for the world beyond. The diversity is large: there is no statistical approach that has not received a recommendation from someone, including the *p*-value. Indeed, the *p*-value seems to have been somewhat rehabilitated, both by the participants and by the editors, shifting their emphasize in terms of Don'ts: Do not use bright-lines such as '$p < .05$,' and in particular do not use the wording 'statistical significance,' and more generally do not use any 'bright-line.' Still, notice that even '$p < .05$' or '$p < .005$' have supporters, in spite of the title of the conference, and rightfully so: How can you design a confirmatory clinical trial without justifying the number of patients with power calculations that rely on a bright-line? This demonstrates that when actions are involved, bright-lines are essential. Moreover, the use of bright-lines and their associated error probabilities as a philosophical foundation in science has often been argued, with a recent persuasive contribution by Debora Mayo (2018). The issues that received unanimous support appear in the four right-most columns,

which represent a code of conduct, rather than choices of statistical methodologies. Wasserstein et al. (2019) summarized them in seven words: "Accept uncertainty. Be Transparent, Open and Modest."

Table 1. A semi-graphical presentation of the do's list for statistics in the 21st century beyond 'p < .05'. Adapted from *The American Statistician* (Wasserstein et al., 2019). Each row presents the recommendation of a paper (as I interpret it,) and each column represents a statistical method(s) or action that should be used in the world beyond. Crossed bold orange is 'never use,' crossed pink is a weaker 'avoid using if possible,' framed light green is 'can be used,' and framed deep green is a strong 'use it.'



The ASA statement emphasized the problems of the *p*-value when it is used as the only statistical summary. It also has advantages, not the least serving as a common language to express the uncertainty in the experimental evidence across branches of science. Moreover, it is the least demanding in terms of its underlying assumption, assumptions that can sometimes be guaranteed by a well-designed and executed experiment. Hence in some

emerging scientific areas it is the only relevant statistical method. But in any case, causing replicability problems is not one of its limitations, and getting rid of it will do little or nothing to help solve the replicability crisis. As Table 1 seems to indicate, no particular other statistical method should be blamed as well. Instead, I argue that two statistical obstacles to replicability are relevant across all statistical methodologies yet are too often ignored. Both are well known, but their full scope is not appreciated by many scientists and even some statisticians in the new world of 'statistics in the 21st century.'

The first is incorporating in the statistical analysis the level of variability that is relevant for replicability, or "hunting out the real uncertainty" in the language of Mosteller and Tukey (1977). The second is selective inference. The effect of the first issue has not changed over the years, but selective inference is increasingly challenging in a world of industrialized science, and this may explain why the replicability problem has gained increased visibility in the last two decades. Therefore, most of the following discussion will be devoted to selective inference, with just a short note on the choice of the relevant variability underlying my recommendations.

Pointing at selective inference as a chief culprit might be surprising, as much of the currently proposed and even enforced means to improve replicability, such as transparency and preregistered protocol, are designed to address the danger from unobservable selection. As explained in Section 3, I argue that it is just as dangerous to replicability when the selection is from the many results that are evident in the published work, yet the statistical inference is not adjusted for the selection. Such selection is a common situation that is typically not addressed by researchers. Only when the number of potential discoveries is high enough does the need to address selection forces itself upon the researchers.

3. Selective Inference

Selective inference is focusing statistical inference on some findings that turned out to be of interest only after viewing the data. Without taking into consideration how selection affects the inference, the usual statistical guarantees offered by all statistical methods deteriorate. Since selection can take place only when facing many opportunities, the problem is sometimes called the multiplicity problem.

Selection is manifested in many different ways in current research, that can be divided into two major types according to whether they are evident in the published work or not. The s*election which is not evident in the published work* has been widely discussed. It includes the publication bias where only results that are statistically significant are published, either because of editors' and reviewers' attitudes, or, as result of these same attitudes, researchers not submitting work that did not produce such results (the file-drawer problem, Rosenthal, 1979). Selection is not evident when it involves the flexibility that a researcher has when analyzing the data before framing the final analysis: choosing to include or omit some data points, including or omitting features, selecting the statistical methods and their tuning parameters, and so on—all the while looking at their implications for the results. Such practices have earned numerous derogatory labels, and addressing such

practices has been advocated by all (see National Academies of Sciences report, *Reproducibility and Replicability*, 2019, for a comprehensive review).

Surprisingly, selection is a problem even when *evident in the published work* where it takes more indirect and subtle forms. I shall hereafter refer to it as 'evident selective inference':

(a) Selection into the abstract (as demonstrated above in Zerbo et al., 2017), which is probably the most common practice for selecting and highlighting; As another example, consider the abstract in Giovannucci et al. (1995), who look for relationships between more than 100 types of foods and the risk of prostate cancer, reports in its abstract only three food intakes, where the association was statistically significant at 0.05: tomatoes, tomatoes sauce, and pizza. Given the food involved, it garnered wide press and media coverage. The paper has been cited close to 2,000 times, and much research followed over the years, including six meta-analyses, four ending with inconclusive results;

(b) Selection by a table, where out of the many items investigated only a few selected ones appear in the table. For example, Zeggini et al. (2007) conducted a genome-wise scan for possible associations of each single nucleotide polymorphism (SNP) on the genome with Type II diabetics. The only results given in their table 1 are the 11 top results out of some 400,000 potential associations;

(c) Selection by highlighting the results for one analysis, even though it is mentioned that parallel analyses were done as well and yielded 'almost similar results';

(d) Selection by a figure, where only the top few findings are presented within a figure. In Stein et al. (2010), where association of the SNPs with volume in the brain across Alzheimer's and healthy patients was analyzed, only the results from five top SNPs out of the 350,000 and only the slices of the brain where some association was present are displayed in a figure;

(e) Selection by highlighting those passing a threshold regarding 'statistical significance' at '$p < .05$', '$p < .005$' or at '$p < 5*10^{-8}$' for genome-wise association studies, threshold on effect size, and thresholding on confidence or credibility intervals not covering zero. These often constitute the findings selected into the abstract (Benjamini & Hechtlinger, 2013), and the finding prone to be selected by readers of the paper;

(f) Model selection, using any of the many forms of selecting variables/features into a model and then ignoring their being selected while calculating *p* values and confidence intervals.

Box 1

The Bonferroni method for simultaneous testing

Controls the probability at α of making even one false discovery out of m, by:

(i) Lowering the level by using $p \leq \alpha^{Bon} = \alpha/m$

(ii) Alternatively, adjusting each p-value p_i to $p_i^{Bon} = p_i m$, and rejecting the corresponding hypothesis if $p_i^{Bon} \leq \alpha$. For reporting while addressing the selection, p_i^{Bon} can be used the way p-value is being used.

The Bonferroni method for simultaneous confidence intervals

Controls the probability of making even one noncovering confidence intervals by constructing marginal ones at $(1- \alpha^{Bon}) \cdot 100\%$.

These cases of selective inference are as damaging as those that cannot be observed from the published work, yet can be easily addressed.

There is no reason to condemn selection and selective inference when the selection is evident. On the contrary: selection for highlighting in its many forms is not only unavoidable but even an important practice in science. It is an essential part of the research in large complex studies such as genomics, brain imaging, or their interface. But it should also be required in small to medium size studies such as the pizza study or the influenza vaccination one. Journalists and the public following them, but also physicians and scientists, mainly read the abstracts of publications and want to get quickly to their major findings. There should be no problem with that, even though any such selection defines a bright-line in practice.

However, such selection has to be adjusted for in order to yield valid *p*-values, confidence intervals, credible intervals, and other forms of statistical inference. Otherwise, these lose their meaning as credible quantitative evaluations of uncertainty. A brief review of approaches to address selective inference follows, after which the current status of their use will be discussed.

4. Addressing Selective Inference

There is a large set of statistical tools for this purpose, which can be categorized according to their stated goal.

4.1. Simultaneous Inference

Simultaneous inference is the most conservative approach, adjusting *p* values and confidence intervals so that the probability of making even one error among the selected ones is controlled at a desired level. It was first proposed in the 1950s in the context of pairwise comparisons between groups, where comparing seven groups with its 21 comparisons using the unadjusted '$p < 0.05$' rule one might expect to find one type I error. That is, the unadjusted '$p < 0.05$' rule has a much higher false positive rate than the declared 5% when it is used to compare the group with the largest observed average to the group with the smallest one (Tukey, 1953). More

generally, consider a family of *m* potential inferences, each about a different parameter μ_i. The inferences can be tests of the hypotheses, where a rejected hypothesis is a statistical discovery, and if rejected in error (type I) a false discovery is made. The inference can be the construction of a confidence interval for a parameter μ_i, and an error is made if the confidence interval fails to cover. Setting the goal for the adjusted inference so that whatever the configuration of true and false hypotheses, or whatever are the parameters' values, counting by *V* the unknown random number of errors made,

$$\Pr\left(\text{making even one error over the entire family of inferences}\right) = \Pr(\ \mathrm{V}\ > 0) \leq \mathrm{a}.$$

Hence this assurance holds for any selected subset of inferences, offering simultaneous error control. The most widely used general method to achieve such control is the Bonferroni method, which actually achieves $E(V) \leq \alpha$, thus implying the above. Other methods, limiting or enlarging the family of potential discoveries over which simultaneous error control is desired, as well as addressing statistical and logical dependencies among the hypotheses, are in constant development (Berk et al., 2013; Goeman & Sollari, 2014). Many of these developments are in response to the stringent requirements of simultaneous error control that are set by drug-regulating agencies across the globe to address the selective inference problem in Phase III studies for drug approval (e.g. Dmitrienko et al., 2008, and Industry Guidance on Multiple Endpoints in Clinical Trials, US Food and Drug Administration, 2017.)

4.2. On the Average Over the Selected

On the average over the selected is a much more lenient approach, ensuring that any property a statistical method has for a single inference will remain true in expectation *on the average over the selected*. This approach appeared first as the false discovery rate (FDR) control in testing. Identifying a discovery with a rejected hypothesis, and a false discovery with a type I error, the goal is to select the largest set of discoveries while assuring that the false discovery proportion is controlled at the desired level *q* in expectation. Namely,

$$\mathrm{FDR} = \mathrm{E}\left(\frac{\#\ \textbf{False Discoveries}}{\#\ \textbf{of Discoveries}}\right) \leq \mathrm{q},$$

with the proportion set to zero if no discovery is made. The FDR approach was introduced in Benjamini and Hochberg (1995), along with a general-purpose FDR controlling procedure, now called the Benjamini-Hochberg (BH) procedure (see Box 2). Both the approach and the associated methodology has seen many variations and developments in the last two decades: adaptive methods (Benjamini & Hochberg, 2000; Storey et al., 2006); non-parametric methods (Dudoit, 2002); hierarchical methods (Yekutieli, 2008; Benjamini & Bogomolov, 2014; Ramdas et al., 2019,); knockoff methods (Barber & Candes, 2015); and online methods (Foster & Stein, 2008; Javanmard & Montanari, 2018) to specify just a few. Developments in Bayes and empirical Bayes approaches will be discussed in Section 5.

Box 2

The false discovery rate (FDR) controlling testing procedure (BH)
Controls the expected average type I errors over the rejections made at level q.
Sort the p-values of the m hypotheses $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)} \leq \ldots \leq p_{(m)}$;

(i) Calculate the largest k for which $p_{(k)} \leq qm/k$, and reject the k hypotheses corresponding to $p_{(1)},\ldots,p_{(k)}$, rejecting none if no such j exists;

(ii) Equivalently, adjust each $p_{(i)}$ to $p_{(i)}^{BH} = min_{j \geq i} \; (\; p_{(j)}m/j \;)$; rejecting the corresponding hypothesis if $p_{(i)}^{BH} \leq q$. (called FDR-adjusted p value and also q value, in both cases using BH). For reporting while addressing the selection $p_{(i)}^{BH}$ can be used the way the p value is.

The false coverage-statement rate (FCR) controlling confidence intervals
Control at level q the expected average number of confidence intervals failing to cover their parameters over the selected confidence intervals.
$S(X)$ (of size $|S(X)|$) is the (simply) selected set of parameters after looking at the data;
Construct for these regular (marginal) confidence intervals at $(1 - q/|S(X)|) \cdot 100\%$ level.

When it comes to confidence intervals, the same concern amounts to controlling the false coverage-statement rate (FCR):

$$\mathrm{FCR} = \mathrm{E}\left(\frac{\#\ \textbf{Intervals failing to cover their respective parameters}}{\#\ \textbf{Intervals constructed}}\right) \leq \mathrm{q}.$$

The intervals selected for construction can be obtained by testing or by other simple selection rule $S(X)$ from the data X, say by selecting the four parameters estimated to be the largest. The fact that regular unadjusted confidence intervals do not offer such protection is demonstrated in Figure 2 and in Column 1 of Figure 3, where the rule is to select those parameters whose unadjusted intervals do not cover zero. A general method for controlling the false coverage-statement rate controlling method that addresses this concern, making use of the size of the selected set $|S(X)|$, has been offered in Benjamini and Yekutieli (2005) (see Box 2). If the BH procedure is used to test whether $\mu_I = 0$, the resulting FCR controlling interval does not cross zero if and only if the corresponding hypothesis is rejected. It is interesting to note that only when targeting confidence intervals, the distinction between the three goals when facing inference following data-dependent selection crystalized: (i) making few errors on the average over all (regular confidence intervals), (ii) making few errors on the average over the selected (FDR and FCR), and (iii) making few errors over all (simultaneous inference). The second goal, assuring the property that a single regular confidence interval offers, to hold 'on the average over the selected' was referred to as the 'selective inference' goal.

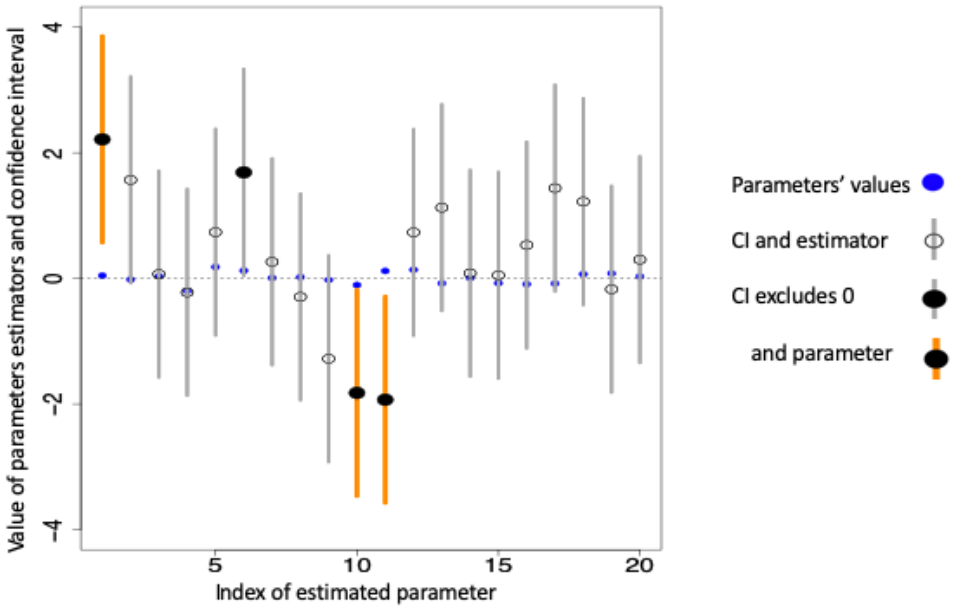


Figure 1. Confidence intervals (CI) coverage: On the average over all vs. over the selected. Twenty means of Gaussian distributions are denoted by blue dots near zero (but none of them at zero). The circles (open or closed) denote their estimators, with the lines indicating the corresponding 90% CIs. Three of the CIs (orange) do not cover their respective means, an error of 3/20 = 15%, which is close to the expected 10% noncoverage (with more simulations the noncoverage rates would approach 10%). Four of the 20 CIs do not cover zero, indicated by full circles at their centers. Therefore, when we choose intervals that do not contain zero, 3/4 = 75% of them will not cover their means. That is, they will have 25% correct coverages, far from the declared 90% of coverage. These selected CIs are too optimistic, and their counterparts in replication efforts will tend to dwindle toward zero, leading to nonreplicability.

4.3. Conditional Inference

Conditional inference is a different approach for addressing selection by constructing for each parameter the inference conditional on it being selected by a specific data-dependent rule. Namely, while regular (marginal) $(1-\propto)\cdot 100\%$ confidence interval (CI) offers Pr(CI_i fails to cover μ_i) ≤ α, for the conditional confidence interval, we require that

$$\begin{aligned}&\Pr\left(\text{CI fails to cover its parameter } \textbf{given} \text{ it is selected}\right)\\ &= \Pr\left(\mu_i \notin \mathrm{CI}_i \middle| i \in S(X)\right) \leq \alpha.\end{aligned}$$

Using conditional inference on each of the selected parameters controls the FCR. The constructed intervals can be nevertheless much wider for some parameters than those offered by simultaneous and FCR approaches, and sometimes can be as narrow as the regular ones. The implications of the publication bias can be assessed by

constructing confidence intervals conditional on passing the '$p < 0.05$' threshold, This is an especially informative interval in that it can changes its size and degree of symmetry depending on the distance from the threshold: The confidence interval is wider toward zero than away from it, and crosses zero when the result is close to the threshold; it converges slowly to the marginal symmetric 95% confidence interval when the corresponding point estimate moves away from the threshold (Weinstein et al., 2013.) This way, it allows reviewers to assess the impact of a generally accepted bright-line such as '$p < .05$'or '$p < .005$' on a particular study. It can similarly be used to reduce the over-optimism about the actual size of the selected large (Zhong & Prentice, 2008.) The conditional approach has seen a surge of interest in the last six years, addressing many practical selection problems such as inference on coefficients in a selected model. In particular, coping with dependency among the estimators requires ingenious methods and computations-intensive solutions (e.g., Lee et al., 2016).

4.4. Sample-Splitting

Sample-splitting is yet another approach, feasible when the number of cases (or subjects) is large. The data can be partitioned into a training or learning set and a holdout set, where the selection is made from data in the first set. Inference in whatever form takes place in the holdout data set, possibly with further selection therein adjusted for (Wasserman & Roeder, 2009.) Multi-sample splitting repeats this process to generate stable *p*-values. This idea has been rigorously treated in the context of post–model selection inference (see Dezeure, 2015, for a review). Care must be taken to assure that further selection does not take place at the second stage, but if it does, this second-stage selection also needs to be adjusted for, using one of the above criteria.

Numerous in-between approaches are being developed, especially for addressing complex hierarchical studies where families or clusters of potential discoveries are selected at the first level, and selecting promising members within the selected families in the second level. Examples include contiguous clusters of activity in the brain and localized voxels within, associations of gene expression (first level) and locations on the genome (second level) in multiple tissues (third level) (see Bogomolov et al., 2020, and Heller et al., 2018).

Addressing selective inference in general is a very active area of statistical research, and better methodologies are constantly under development. However, it is important to emphasize that the field is mature enough to already offer practical means to address selective inference evident in the study for most—if not all—such problems encountered in data science.

5. The Status of Addressing Selective Inference

I review the status of selective inference from the multiple results that appear in the published work—the evident selective inference. At one extreme comes the care given to selective inference in experiments for drug approval and registration. There is a long tradition of careful adherence to addressing selective inference, both out of and in the reported study, by the regulating bodies for drug registration. Whenever there is more than a single primary endpoint that may indicate the success of a treatment, simultaneous inference is evoked, and if

failed—so fails the experiment. Hence at these Phase III trials an effort is made by the drug companies to keep the number of primary endpoints small. There are variations in the way selective inference is addressed when dealing with the larger number of secondary endpoints (used for labeling) and with safety outcomes, but with too little attention to the latter.

5.1. Large-Scale Research

As argued before, selection is unavoidable in large-scale research where a large number of potential discoveries are screened. It is therefore widely used and even mandated when it comes to studies with thousands, millions, and even billions of potential discoveries. Researchers nowadays use lower significance levels as thresholds to control some error rates, like the '$p \leq 5*10^{-8}$' used in searching for associations between a single disease and locations on the genome. This level was chosen to assure that the (approximate) simultaneous error will be less than .05. The control of FDR at the .05 or 0.1 level is widely used for testing in these genomwise association scans, as well as in other complex ones, in genomics, proteomics, and functional imaging. The result is again a much lower *p*-value threshold requirement for each single discovery.

When confidence intervals are constructed for the selected few, the fact that they were selected based on the same data is almost totally ignored. For example, in the Zeggini et al. (2007) study their table 1 includes the 11 significant SNPs at the '$p \leq .05$' level *after adjusting for selection*, yet the confidence intervals for the relative risks of diabetics to carriers of the rarer SNP versus noncarriers are presented as usual, with no adjustment for their being the most promising 11 selected out of 400,000. These confidence intervals are surely overly optimistic and the estimated effects are likely to shrink substantially upon replication, yet the practice has not changed over the years. The recent genomwise association study of severe respiratory failures in Covid-19 patients reports in its abstract the two associated loci with selection adjusted significance, together with regular (unadjusted) confidence intervals for their effects (Ellinghous et al., 2020).

It is illuminating to observe the stage at which researchers in a field realize they have a selective inference problem. In quantitative traits loci analysis, experimentalists mapped associations of a trait with 700 different locations on the genome before a warning was issued by leaders in the field that "there is danger that a large proportion of findings be false" (Lander & Kruglyak, 1995). In the analysis of gene expression data using micro-arrays, the number of different genes screened in a single experiment reached ~8000 while $p < .01$ was still in use, when attention was first devoted to the multiplicity problem. With millions of potential findings being screened these days in a study, there is no doubt that selective inference should be addressed, in one way or the other, and indeed, it has become a required standard for reporting results (see more in the final section).

5.2. Medical Research

Unfortunately, this is not the case in most of medical research whether preclinical, clinical, or epidemiological. In an in-depth analysis by Benjamini and Cohen (2017) of a random sample of 100 papers from the *New England Journal of Medicine* (*NEJM*; 2002–2010), the number of efficacy endpoints in a paper ranged from 4

to 167 and averaged 27. Only in 20 of them was the issue of multiplicity addressed, in none fully so. The .05 threshold for each individual discovery was used throughout. Though all studies designated primary endpoints, positive conclusions were derived even when the primary endpoints did not support such a conclusion, and the study was nevertheless declared a success.

A greater awareness to the problem of selective inference seem to have taken place recently. In July 2019 the *NEJM* issued new statistical guidelines. As explained in Harrington et al. (2019), among them is "a requirement to replace *p*-values with estimates of effects or association and 95% confidence intervals when neither the protocol nor the statistical analysis plan has specified methods used to adjust for multiplicity." Progressing in the right direction, the new guidelines acknowledge the problem of multiple comparisons, but their solution is to ignore the problem rather than address it. The editorial describes a study published a few months earlier in the *NEJM*: "fatty acids did not significantly reduce the rate of either the primary cardiovascular outcome or the cancer outcome. If reported as independent findings, the *p*-values for two of the secondary outcomes would have been less than 0.05; However, the article reported only the hazard ratios and confidence intervals for the intervention effects for those secondary outcomes, consistent with recently implemented Journal guidelines limiting the use of *p*-values for secondary and other comparisons." The *p*-values are not shown, the confidence intervals are left unadjusted, and the problem of selective inference in face of the multiple comparisons is mopped under the rug.

5.3. Experimental Psychology

Evident selection is also not addressed in experimental psychology. Take, for example, the 100 papers of the Psychological Reproducibility Project described in the introduction: the number of inferences per paper ranged from 4 to 740 and averaged 72 (by our approximate count that included inferences not restricted to '*p*=' statements). Only 10 of the 98 studies, where the main finding was to show an effect, partially adjusted their results for selection. In 90 of them further adjustment was needed. Using the hierarchical FDR method of Benjamini and Bogomolov (2014) to adjust for the selection, a method rather lenient in that it utilizes the special structure of a complex study, 22 previously statistically significant results turned nonsignificant after FDR adjustment. Out of these, 21 were results that failed the replication efforts by the project's members, so addressing selective inference in the original analyses could have prevented these 21 from being published as statistically significant discoveries, missing merely the single actually replicated result. Such a change should substantially increase the replicability of results in experimental psychology (Zeevi et al., 2020).

5.4. Open Science Framework

Evident selection need not be addressed according to the Open Science Framework (https://osf.io/)—it is enough to be open about the multiplicity of inferences being made and report them all, an attitude shared with the ASA statement (Wasserstein & Lazar, 2016.) In their illustrative study there is one main question regarding a score constructed from accumulating scores of 29 categories to be compared between two types of

registration. The preregistration discusses at length every detail of the analysis, both of the accumulated score and of the separate ones. It also specifies that each one of the 29 tests of the categories will be conducted at marginal 1/20 level. This protocol, which the authors clearly regard as exemplary, allows no flexibility on the inference criteria for these follow-up analyses. As no adjustment for selection is allowed, we expect to find at least one confidence interval not covering zero, whether rightfully or not.

5.5. Bayesian Approach

Evident selection need not be addressed according to many statisticians adhering to the Bayesian approach. They dismiss the concern about multiplicity and selective inference relying on the theoretical argument that conditioning on the data generated via the prior nullifies the effect of selection (e.g., Gelman et al, 2012), Nevertheless, posterior credibility intervals are not immune from such selection effect, if the prior used does not reflect exactly the distribution from which the effects to be selected are screened. A vivid demonstration of the extent of the problem in relying only on Bayesian inference to avoid addressing selective inference is given in Figure 3, where the assumed model for the prior deviates slightly from the real prior generating the parameters. The model and software used for the example are a modification of that used by Gelman in his blog. Gelman used it to demonstrate how the standard unadjusted confidence intervals suffer from the selection of those that do not cover zero, while Bayesian credibility intervals do not need any adjustment. His results are reproduced in the two leftmost columns. The third column shows how false coverage-statement intervals do have appropriate coverage for the selected. The parameters were then generated from a slight modification of the specified prior, mixing it with a wider normal distribution with probability of 1/1000. This had a small effect on the performance of posterior credibility intervals *on the average over all parameters*, but had a devastating effect on the coverage error property of the selected few. Seventy-three percent of the intervals that were selected for not including zero did not cover their respective parameter—instead of the desired 5%.

Table 2. Performance of frequentists and Bayesian intervals under selection. On the left three columns, the assumed prior and the model generating the parameters are the same. On the right three columns the assumed prior remains as on the left, but the parameters are generated from a slightly different distribution, the one on the left mixed by a small probability of 1/1000 of another distribution. The BH false discovery rate (FDR) controlling procedure and the FCR adjusted confidence intervals are described in Box 2.

Parameters generating Prior		$\mu_i \sim N(0,0.5^2)$			$\mu_i \sim N(0,0.5^2)$ *w.p. 0.999* $\sim N(0,3^2)$ *w.p. 0.001*	
Type of 95% confidence/credibility intervals	**Standard (Marginal)**	**Bayesian Credibility**	**Regular FCR-Adjusted**	**Standard (Marginal)**	**Bayesian Credibility**	**BH-Selected FCR-Adjusted**

Intervals not covering their parameter	5.0%	5.0%		5.0%	5.1%	5.6%
Intervals not covering 0: The Selected	8.0%	0.01%		8.0%	0.03%	0.03%
Intervals not covering their parameter – Out of the Selected	45.9%	5.5%	5.0%	**45.5**	**72.9%**	5.6%

The disagreement between the model for the prior and the true generating mechanism is somewhat circumvented in high-dimensional problems by the empirical Bayes approach, where the prior is estimated from the data. Furthermore, since in such problems the FDR has a Bayesian interpretation (Genovese & Wasserman, 2002), selective inference using FDR-motivated error rates have been developed, first for microarray analysis (Efron et al, 2001) and more generally culminating in Efron (2012). Yekutieli (2012) sets up a different Bayesian framework, similar to the conditional approach, where addressing selection is needed, and provides an empirical Bayes valid inference for the selected parameters. Again, these are mostly relevant to large-size problems and have seen less use in medium-size problems.

5.6. *Nature*

Evident selection need not be addressed, also according to a recent comment in *Nature* (Amrhein et al., 2019), which called to retire statistical significance. Backed by signatures of some 800 followers, the commentary argues against the use of statistical significance for any purpose, and in general of *any bright-line*. For support, they rely on the editorial of Wasserstein et al. (2019) and on three of the 43 papers therein. One of them, Hurlbert et al. (2019), argues vividly against statistical significance, and discusses the questions that might be raised against this stand. Considering the hypothetical question of how we can address multiple comparisons without a threshold, they answer: “You can’t. And shouldn’t try!” (p. 354 therein.) The first claim is wrong, since using adjusted *p*-values can address multiplicity without relying on prespecified thresholds. The second claim clarifies their attitude toward addressing selective inference. It is even clearer when viewing the paper that they give as an example to nuanced reporting that needs no bright-lines (Reifel et al., 2007) and addresses multiple results. In that paper, a table displays the results of 41 regression analyses for log abundance of species of plankton as a function of distance downcurrent of rivers’ inflow—all unadjusted for the selection of the smaller (more significant) *p* values. But the abstract identifies eight results from their table: five with $p \leq .01$ and three more with $p \leq .1$, so only results statistically significant at .1 were selected into the abstract.

Without adjustment for their selection, they give too optimistic an impression about the uncertainty in these results than is warranted.

5.7. General

I therefore argue that unattended selective inference—even though it is evident in the published paper—is a silent killer of replicability in middle-size studies, where the number of inferences from which selection is being made is above a handful, but not yet in the thousands. This also explains why the replicability problem surfaced in some areas of science 20 to 30 years ago, just as the industrialization of science increased the pool of potential discoveries from which only a few have to be selected and followed. When the size of the pool of potential discoveries increases slowly there is the tendency to ignore the available solutions to address the selection bias, as all imply a price in terms of reduced probability of making a true discovery. When the size of the problems encountered becomes even larger it also becomes clear that it is not enough to be aware of selective inference. A decision to address the evident problem is unavoidable, and the researchers themselves reach this conclusion (see examples in Section 2). But we clearly cannot wait years for this to happen in important areas such as medical research.

6. Replication as a Way of Life in Scientific Work

Let me emphasize again that the replicability of the stand-alone study cannot be assured, only enhanced, and I argued that addressing evident selective inference is essential for that. I mentioned only briefly that selective inference that is not evident might also have an important impact on replicability. Nevertheless, this problem attracted most of the discussions so far, starting with Ioannidis (2005) through committee reports and editorials. An emphasis is given by all to the reproducibility aspects of the study, or its pre-reproducibility, as discussed in Stark and Saltelli (2018). My interest in this last section is in the next stage, in assessing the replicability of results from further studies.

Replicability by others is essential, and is recognized by scientists as such. Interestingly, in areas where replication seems impossible, great efforts are being made to get as close as possible to this ideal (Junk & Lyons, 2020). Only one particles collider in the world (at CERN) can accelerate particles to the high speeds needed, so that upon their collision the Higgs boson, if exists, can be detected. The search for this particle was conducted by two teams using two different experimental setting and different machines where the results of collisions were measured. Only then, in 2012, did they publish the discovery of the Higgs boson, in two back-to-back papers in *Physics Letters B* (ATLAS collaboration, 2012; CMS Collaboration, 2012). Earth scientists study our unique earth and often study a unique phenomenon. For replicability of discoveries on the global scale they rely on different teams of researchers utilizing different models and methods of analysis as a check for replicability and incorporate guidelines that will make such checks feasible (McNutt et al., 2016). In both fields there is further reliance on blinded analysis where the data analysts are not aware of which result supports the research hypothesis (e.g., Klein & Roodman, 2005).

In other areas of science replicability is essential not only as a validation of previous results, but as first step toward their generalization. Indeed, the borders between replicability and generalizability are not sharp, and need not be so. If a result is applicable to a single sharply specified situation, in a particular population, even if it is replicated under a very similar setting, but only then, is it still of interest? Sometimes it might be a hint of a difference of interest, but more often it may represent the peculiarities of the setting of the study.

That is the reason why capturing the relevant variability in the assessment of the uncertainty is joining the addressing of selective inference as cornerstones of replicability. Ideally the relevant variability should resemble the changes expected in the result when another experimenter, in another lab, is making an honest and unbiased effort to achieve the same goal. While simple to state, capturing the relevant variability is not a minor challenge. For a concrete example consider the simplest mice phenotyping experiments, where two or more genetically pure lines of mice are compared on some quantitative traits. Crabbe et al. (1999) found in a well-coordinated and highly transparent study across three laboratories genotype-by-laboratory interaction, which led them to conclude in their abstract that "experiments characterizing mutants may yield results that are idiosyncratic to a particular laboratory." In another multi-lab study (Kafkafi et al., 2005) we demonstrated that when this interaction is incorporated into the variability estimates on top of the usual animal-by-animal variability (technically by using a random-lab model), an adjusted yardstick is derived against which genotype differences can be judged. Consequently, the random-lab model raises the benchmark for finding a significant genotype effect and widens the confidence interval of the estimated effects sizes. Some differences remained statistically significant, despite the interaction, while some did not—trading some statistical power for greater likelihood of replication. Still, should every mouse phenotyping experiment be conducted in multiple labs? That is quite unlikely, so alternative approximating solutions are being studied (Kafkafi et al., 2018).

A similar conclusion was reached in experimental psychology by Schooler (2011). Summarizing an extensive replication effort of his work on 'verbal shadowing,' he realizes that results should be replicated as a way of life and states that he came to agree with three colleagues to run each other's experiments. However, he also admits that this cannot be the solution for all studies, and single-lab experiments will prevail. Indeed, in one of his latest papers (Protzko et al., 2019) they resorted to replicating their own study in full (1,500 participants), rather than leaving it to others, as a partial step for the single study.

At this point a distinction is needed between a study that results in an immediate action, which can be as simple as marketing decisions or as difficult as in drug approval processes, and a scientific discovery. For actions, relying on some threshold is fine. It can rely, say, on $p < .05$, $p < .05$ in two studies, or thresholds involving more than one statistical summary, such as $p < .005$, and standardized effect size bigger than half. Reaching a conclusion about a scientific discovery that does not require an immediate action is different. Results supporting it or contradicting it can accumulate, and a conclusion can be taken in view of their accumulation. With multiple replication efforts in sight, the result of a replication study for the second purpose should also not be expected to be a clear yes or no answer, just as the original study should rarely be seen as such.

Replication results can and should accumulate, and for that purpose many small studies are generally better than a single large one of the same total size, even if each small study is underpowered, because they reflect the relevant study to study variability. This last statement goes contrary to the current dogma, which states that studies, and especially replication ones, should not be underpowered. There are now statistically rigorous ways to combine the results of many small studies, say *k* of them, and get an assessment of replicability of evidence in *u*-out-of-*k* studies (see Box 3 and Jaljuli et al., 2019). Returning to the pizza and prostate cancer example (Giovannucci et al., 1995), the most recent meta-analysis of Rowles et al. (2017) included 21 dietary studies. The highlighted relevant finding is the negative association between dietary intake of lycopene (e.g. tomatoes and its products) and prostate cancer in the combined analysis, with $p=0.017 < .05$ and 95% CI of .78 to .98. Nevertheless, analyzing it for replicability as above showed that if a single study is dropped from a jointly analyzed group of studies, the result is no longer statistically significant at .05, indicating the lack of replicability of this result. Taking a positive point of view, the evidence in three studies each with $p = .06$, is not only enough to reject the hypothesis that there is no effect in all studies (via the Fisher combination test), but also enough to reject the hypothesis that there is at most one study with no effect—yielding a claim of replicability at the .05 level.

Box 3

(i) For quantifying the claim that there is an effect in at least u studies out of k: Define $r_{u|k}$ to be the smallest level at which the hypothesis that there is an effect in no more than (*u-1)* out of *k* in the same direction can be rejected.

(ii) The minimal requirement for replicability is that there are at least two studies with effects in the same direction. Henceforth, the $r_{2|k}$ will be referred to simply as the r-value.

(iii) Set u_{max} to be the largest *u* for which $r_{u|k} \leq \alpha$, u_{max} is the 1- α lower bound on the number of studies with effect in the same direction.

The methodology for testing the (*u-1*)-out-of-*k* hypothesis is by dropping the most convincing u-1 results and testing the global null hypothesis that there is no effect in the remaining, using combination p-values.

Since the replicability of a result *by others* is the key, we should identify and establish policies that make replicating others' work an integral part of the regular way of conducting science. Running 'reproducibility projects,' in which teams of scientists join to try and replicate all, or a random sample from all, studies of one type or the other is not the right way for two reasons. Once the first hype about such projects is gone, scientists will be reluctant to try and replicate others' work just for the purpose of their replication, with no relevance for their own work. Indeed, the cancer biology reproducibility research (Nosek & Errington, 2017) replicating

studies from 2000 to 2002 was slow to develop. A second objection is about which results replication attempts should be targeted. Just taking random results from leading journals and investing time and money in replicating them all cannot become a way of life. The suggestion of publishing more papers reporting the results of individual replication efforts is fine, but can journals devoted entirely to publishing negative results, be sustainable? I doubt it. This should be contrasted with the recent surge in the venues for publishing meta-analysis reviews indicating the interest readers have in studies that compile and critically integrate and evaluate the results of multiple studies, positive, negative, and inconclusive.

I suggest a way that looks quite different, but that is not too far from the traditional way scientists work. It does require the participation and cooperation of the scientists, journal editors, granting agencies, and academic leaders, as well as institutions such as the open science framework. The proposal has the following five key components.

1. Every research proposal and every submitted paper should have a replicability-check component in the study. In that part, a result considered by the authors important for their proposed research will be up for replication. Just as the introductory section of a paper presents and discusses the literature on which the reported work is based, so will one aspect of this preceding work be up for replication. Naturally, it will be a part that relies on the same setup and tools as in the proposed study. This way, only results that are considered worthy enough by other researchers will attract replication attempts.
2. The replicability study design will be preregistered with institutions such as the Open Science Collaboration. The results of the replicability components will be reported whatever the outcome is, in the extended-abstract/main-body, in two to three searchable sentences. The details will be given in the supplementary material, but space has to be devoted by publishers for the basic information about the replication effort. Results should not be reported as 'replicated / not replicated,' but rather by presenting, for example, exact *p* value, effect size estimate, confidence interval, or their respective analogous Bayesian quantities.
3. Granting agencies will review a proposed replication effort and its adequacy as part of the regular submission of a novel research proposal, and will allocate funds appropriately. Inconclusive or negative replication results should not be a reason to cancel a grant, and this can be avoided by including in the proposal a bailout direction of research. The proposed replication of an already well-established result may be discouraged by the reviewers, or be modified toward generalizability.
4. Meta-analysis reviews will utilize the results of the replication components to get estimates of the level of replicability, together with better effect size estimates and their standard errors. The fact that these components will be entirely preregistered and the results will be published (preferably in the same format) whatever the novel results are, will eliminate from the replication efforts the publication bias and other selective inference practices that are not evident in the published work. Generally, such meta-analyses will be easier to perform and be much more informative than current ones.
5. Special recognition is appropriate to authors of papers for which replications of results were attempted and supported, because their result was considered important enough by other investigators in the field to invest

> the efforts in trying to replicate it. Counting such research means much more than merely counting citations, especially those like '(see also [7]–[16]).' Therefore, academic leaders should give special credit to such authors in promotion decisions, prizes, grant awards, and such.

Summing up, enhancing the replicability of the single study by addressing selective inference and creating an environment where replication by others becomes a way of life to scientists, as it was at the beginning of modern science, will greatly reduce the replicability problems we encounter today. Such efforts should also revive the public trust in the scientific method and the discoveries made, a goal set by the U.S. Congress to the reproducibility and replicability effort conducted by the U.S. National Academies.

Acknowledgment

This work partly reflects my Rietz Lecture presented at the Joint Statistical Meeting of 2019. I am indebted to Neri Kafkafi, Sofi Estashenko, Amit Meir, Yoav Zeevi, Daniel Yekutieli, and Ruth Heller, as I rely in this review on our joint work on replicability. My collaborative work on selective inference involves too many researchers to mention by name, but they were all important in shaping many of the suggestions reviewed. Finally, I am thankful to Henry Braun and Abba Krieger for friendly discussions on the manuscript, and to the reviewers and the editor for helpful suggestions. This work was supported over the years by a European Research Council grant PSARPS 294519, US National Institute of Health grant HG006695, and NSF-US-Israel Binational Science Foundation grant 2016746.

Disclosure Statement

Yoav Benjamini has no financial or non-financial disclosures to share for this article.

References

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

ATLAS Collaboration. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, *716*(1), 1–29. https://doi.org/10.1016/j.physletb.2012.08.020

Barber, R. F., & Candes, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics*, *43*(5), 2055–2085. https://doi.org/10.1214/15-AOS1337

Benjamini, Y. (2016). It's not the *p*-values' fault. *The American Statistician*, *70*(2) (supplemental material to Wasserstein, R. L., & Lazar, N. A., The ASA's statement on *p*-values: Context, process, and purpose). https://doi.org/10.1080/00031305.2016.1154108

Benjamini, Y., & Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B, 76*(1), 297–318. https://doi.org/10.1111/rssb.12028

Benjamini, Y., & Cohen, R. (2017). Weighted false discovery rate controlling procedure for clinical trials. *Biostatistics, 18*(1), 91–104. https://doi.org/10.1093/biostatistics/kxw030

Benjamini, Y., & Hechtlinger, Y. (2013). Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek. *Biostatistics, 15*(1), 13–16. https://doi.org/10.1093/biostatistics/kxt032

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*(1), 289 –300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., & Hochberg Y. (2000). The adaptive control of the false discovery rate in multiple comparison problems. *The Journal of Educational and Behavioral Statistics, 25*(1), 60–83. https://doi.org/10.3102/10769986025001060

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association, 100*(469), 71–81. https://doi.org/10.1198/016214504000001907

Berk, R., Brown, L., Buja, A. Zhang, K., Zhao, I. (2013). Valid post-selection inference. *The Annals of Statistics, 4*1(2), 802–837. https://doi.org/10.1214/12-AOS1077

Bogomolov, M., Peterson, C.B., Benjamini, Y. Sabatti C. (2020). Hypotheses on a tree: New error rates and testing strategy. *Biometrika* (in press).

CMS Collaboration. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B, 716*(1), 30–61. https://doi.org/10.1016/j.physletb.2012.08.021

Crabbe, J. C., Wahlsten, D., & Dudek, B.C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science, 284*(5420), 1670–1672. https://doi.org/10.1126/science.284.5420.1670

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dezeure, R., Bühlmann, P., Meier, L., & Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, *p*-values and R-software hdi. *Statistical Science, 30*(4), 533–558. https://doi.org/10.1214/15-STS527

Diggle, P. J., & Zeger, S. L. (2010). Embracing the concept of reproducible research. *Biostatistics, 11*(3), 375. https://doi.org/10.1093/biostatistics/kxq029

Dmitrienko, A., Tamhane, A. C., & Bretz, F. (2008). *Multiple testing problems in pharmaceutical statistics*. Chapman and Hall /CRC Press. https://doi.org/10.1201/9781584889854

Dudoit, S. Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica, 12*(1), 111–139.

Efron, B. (2012). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction* (Vol. 1). Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association, 96*(456), 1151–1160. https://doi.org/10.1198/016214501753382129

Ellinghous, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P. Fernandez, J., Prati, D., Baselli, G., Asselta, R., Grimsrud, M. M., Milani, C., Aziz, F., Kassens, J., May, S., Wendorff, M., Wienbrandt, L., Uellendahl-Werth, F., Zheng, T., Yi, X., … & Karlsen, T. D. (2020). Genomewide association study of severe Covid-19 with respiratory failure. *New England Journal of Medicine, 383*(16), 1522–1534. https://doi.org/10.1056/NEJMoa2020283

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Foster, D. P., & Stine, R. A. (2008). α-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B, 70*(2), 429–444. https://doi.org/10.1111/j.1467-9868.2007.00643.x

Fricker, R. D., Jr., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in *Basic and Applied Social Psychology* after their *p*-value ban. *The American Statistician, 73*(Supplement 1), 374–384. https://doi.org/10.1080/00031305.2018.1537892

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Genovese, C., & Wasserman, L. (2002). Bayesian frequentist multiple testing. *Bayesian Statistics, 7*, 145–161.

Giovannucci, E., Ascherio, A., Rimm, E. B., Stampfer, M. J., Colditz, G. A., & Willett, W. C. (1995). Intake of carotenoids and retino in relation to risk of prostate cancer. *Journal of the National Cancer Institute, 87*(23), 1767–1776. https://doi.org/10.1093/jnci/87.23.1767

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine, 33*(11), 1946–1978. https://doi.org/10.1002/sim.6082

Harrington, D., D'Agostino, R.B., Gatsonis, C., Hogan, J. W., Hunter, D. J., Normand, S.L.T., Drazen, J. M., & Hamel, M. B. (2019). New guidelines for statistical reporting in the *Journal*. *New England Journal of Medicine, 381*(3), 285–286. https://doi.org/10.1056/nejme1906559

Heller, R., Chatterjee, N., Krieger, A., & Shi, J. (2018). Post-selection inference following aggregate level hypothesis testing in large scale genomic data. *Journal of the American Statistical Association, 113*(524),1770–1783.
https://doi.org/10.1080/01621459.2017.1375933

Hurlbert, S. H., Levine, R. A., & Utts, J. (2019). Coup de Grâce for a tough old bull: "Statistically Significant" expires. *The American Statistician, 73*(Suppl. 1), 352–357. https://doi.org/10.1080/00031305.2018.1543616

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Jaljuli, I., Benjamini, Y., Shehav, L., Panagloton, O., & Heller, R. (2019). Quantifying replicability and consistency in systematic reviews. *arXiv.* https://doi.org/10.48550/arXiv.1907.06856

Javanmard, A., & Montanari, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics, 46*(2), 526-554. https://doi.org/10.1214/17-AOS1559

Junk, T. R., & Lyons, L. (2020). Reproducibility and replication of experimental particle physics results. *Harvard Data Science Review 2(4).* https://doi.org/10.1162/99608f92.250f995b

Kafkafi, N., Benjamini, Y., Sakov, A., Elmer, G. I., & Golani, I. (2005). Genotype-environment interactions in mouse behavior: A way out of the problem. *PNAS, 102*(12), 4619–4624.
https://doi.org/10.1073/pnas.0409554102

Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Science*, *55*, 141–163.
https://doi.org/10.1146/annurev.nucl.55.090704.151521

Lander, E., & Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, *11*(3), 241–247. https://doi.org/10.1038/ng1195-241

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics, 44*(3), 907–927.
https://doi.org/10.1214/15-AOS1371

Levelt, W. J., Drenth, P. J. D., & Noort, E. (Eds.) (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Tilburg University, University of Amsterdam, and the University of Groningen. http://hdl.handle.net/11858/00-001M-0000-0010-258A-9

Mann C. C. (1994). Behavioral genetics in transition. *Science, 264*(5166), 1686–1689. https://doi.org/10.1126/science.8209246

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press. https://doi.org/10.1017/9781107286184

McNutt, M. (2014). Reproducibility. *Science, 343*(6168), 229. https://doi.org/10.1126/science.1250475

McNutt, M., Lehnert, K., Hanson, B., Nosek, B., Ellison, A., M., & King, J. L. (2016). Liberating field science samples and data. *Science, 351*(6277), 1024–1026. https://doi.org/10.1126/science.aad7048

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Addison-Wesley.

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

Nature. (2013). Announcement: Reducing our irreproducibility. *Nature, 496*(7446), 398. https://doi.org/10.1038/496398a

Nuzzo, R. (2014). Statistical errors. *Nature, 506*(7487), 150–152. https://doi.org/10.1038/506150a

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. J., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestable, M., Dafoe, A. Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., ... & Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *eLife, 2017*(6), Article e23383 https://doi.org/10.7554/eLife.23383

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 5, 50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Peng, R. D. (2009). Reproducible research and *Biostatistics*. *Biostatistics*, *10*(3), 405–408. https://doi.org/10.1093/biostatistics/kxp014

Protzko, J., Zedelius, C. M., & Schooler, J. W. (2019). Rushing to appear virtuous: Time pressure increases socially desirable responding. *Psychological Science*, *30*(11), 1584–1591. https://doi.org/10.1177/0956797619867939

Ramdas, A., Foygel Barber, R., Wainwright, M. J., & Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, *47*(5), 2790–2821. https://doi.org/10.1214/18-AOS1765

Reifel, K. M., Trees, C. C., Olivo, E., Swan, B. K., Watts, J. M., & Hurlbert, S. H. (2007). Influence of river inflows on spatial variation of phytoplankton around the southern end of the Salton Sea, California. *Hydrobiologia*, *576*(1), 167–183. https://doi.org/10.1007/s10750-006-0300-3

Rothman, K. J. (1998). Writing for epidemiology. *Epidemiology*, *9*(3), 333–337. https://doi.org/10.1097/00001648-199805000-00019

Rosenthal, R. (1979). File drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rowles III, J. L., Ranard, K. M., Smith, J. W., Erdman Jr, J. W. (2017) Increased dietary and circulating lycopene are associated with reduced prostate cancer risk: a systematic review and meta-analysis. *Prostate Cancer and Prostatic Diseases*, *20*(4), 361–377. https://doi.org/10.1038/pcan.2017.25

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), Article 437. https://doi.org/10.1038/470437a

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton University Press.

Soric, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, *84*(406), 608–610. https://doi.org/10.1080/01621459.1989.10478811

Stark, P. B., & Saltelli, A. (2018). Cargo-cult statistics and scientific crisis. *Significance*, *15*(4), 40–43. https://doi.org/10.1111/j.1740-9713.2018.01174.x

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A., Saykin, J., LiShen, Foroud, T., Pankratz, N., Huentelman, M. J., Craig, D. W., Gerber, J. D., Allen, A. N., Corneveaux, J. J., De Chairo, B. M., Potkin,S. G., Weiner, M. W., ... & Thompson, P. M. (2010). Voxelwise genome-wide association study (vGWAS). *Neuroimage*, *53*(3), 1160–1174. https://doi.org/10.1016/j.neuroimage.2010.02.032

Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society: Series B, 66*(1), 187–205. https://doi.org/10.1111/j.1467-9868.2004.00439.x

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and applied social psychology, 37*(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

Tukey, J. W. (1953). *The problem of multiple comparisons*. Unpublished manuscript. In H. I. Braun (Ed.) (1994), *The collected works of John W. Tukey* (Volume VIII-Multiple Comparisons: 1948–1983; pp. 1–300). Chapman and Hall/CRC.

US Food and Drug Administration. (2017). Multiple endpoints in clinical trials; Draft guidance for industry. FDA-2016-D-4460.

Wasserman, L., & Roeder K. (2009). High-dimensional variable selection. *Annals of Statistics, 37*(5A), 2178–2201. https://doi.org/10.1214/08-AOS646

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p-Values: Context, process, and purpose*. *The American Statistician, 70*(2), 129–131. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$." *The American Statistician, 73*(Suppl. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Weinstein, A., Fithian, W., & Benjamini, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association, 108*(501), 165–176. https://doi.org/10.1080/01621459.2012.737740

Yekutieli, D. (2008). Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association, 103*(481), 309–316. https://doi.org/10.1198/016214507000001373

Yekutieli, D. (2012). Adjusted Bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B, 74*(3), 4255–4271. https://doi.org/10.1111/j.1467-9868.2011.01016.x

Zeevi, Y., Estashenko, S., Meir, A., & Benjamini, Y. (2020). *Improving the replicability of results from a single psychological experiment*. [stat.ME] [Manuscript submitted for publication].

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., & Barrett, J. C. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science, 316*(5829), 1336–1341. https://doi.org/10.1126/science.1142364

Zerbo, C., Qian, Y., Yoshida, C., Fireman, B. H., Klein, N. P., & Croen, L. A. (2017). Association between influenza infection and vaccination during pregnancy and risk of autism spectrum disorder. *JAMA Pediatrics*,

171(1), Article e163609. https://doi.org/10.1001/jamapediatrics.2016.3609

Zhong, H., & Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics, 9*(4), 621–634. https://doi.org/10.1093/biostatistics/kxn001
