Of forking paths and tied hands: Selective publication of findings, and what we should do about it

Maximilian Kasy

December 20, 2019

A number of authors have argued that a sizable share of published research in the social sciences and the life sciences is not replicable Camerer et al. (2016); Open Science Collaboration (2015); Camerer et al. (2018), and much of conventional statistical inference in these fields is distorted, so that estimates are biased and tests don't control size Ioannidis (2005); Gelman and Loken (2013). One possible reason for this lack of replicability is selective publication of findings, whether by researchers (p-hacking, specification searching, the "garden of forking paths") or journals (publication bias). Either source of selection implies that the probability that some finding is published depends on the finding itself (such selection distorts inference), and not just on the question, the method, or the sample size of a study (selection based on these does not distort inference).

In this context, applied researchers, as well as policymakers, are confronted with two big sets of questions. First, how can we tell to what extent selective reporting and publication is really taking place in a given literature? How much are published estimates inflated as a consequence of this? This is crucial for informed decision-making based on published research in the social and life sciences. Second, having established that selective reporting and publication is widespread, how should we reform statistics teaching as well as the publication system in order to reduce these problems? What do different objectives imply for the optimal system; objectives such as replicability, relevance for decision makers, or plausibility of published findings? What is the role of pre-commitment ("tying researchers' hands")?

In this paper, I will discuss these questions, starting with evidence on selective publication. Several methods have been used in the literature to provide evidence for selective reporting and publication: Plotting the distribution of published p-values, regressing published estimates on reported standard errors (or their inverse), and considering the "rate of replication" in replicated experiments (that is, the share of significant findings which are also significant when replicated). While intuitive, these three methods rely on problematic assumptions, and do not allow to estimate the magnitude and form of selection; I will explain why that is the case. In Andrews and Kasy (2019), we propose two alternative methods which allow us to estimate the full extent of selective reporting by researchers and selective publication by journals. The first of these methods uses systematic replication experiments and builds on the intuition that, absent selection, original and replication estimates should be distributed symmetrically. The second of these methods uses meta-studies and builds on the intuition that absent selection the distribution of estimates should be more dispersed for larger standard errors. I will provide visual arguments for identification using either of these approaches.

These approaches allow us to establish that published research in many fields is highly selected. Does that matter, and what should we do about that? It has become broadly recognized that selective reporting and publishing invalidates standard statistical inference and is at the heart of the "replication crisis" of a number of disciplines. This has motivated reform efforts aimed at changes in norms (don't put "stars" based on significance in your tables), changes in journal policies (requiring pre-analysis plans for experimental research to "tie researchers' hands," accepting papers based on registered reports), and changes in the institutional infrastructure for academic research (journals for null-results, journals for replication studies).

Are these efforts going in the right direction? Drawing on Frankel and Kasy (2018) I will argue that different justifiable objectives for scientific institutions lead to contradictory recommendations. We need to be explicit about our objectives in order to discuss the tradeoffs between them. Replicability and the validity of conventional statistical inference constitute one such objective, and they indeed require that publication decisions do not depend on findings. This is what motivates much of current reform efforts. Validity of inference is presumably not the only objective, however – it could easily be achieved by estimates derived from a random number generator. Relevance of findings might be another objective. If our goal is to inform decision makers or to maximize social learning, there is a strong rationale to selectively publish surprising findings. A third objective could be the plausibility of published findings. If there is some uncertainty about the quality of studies and we want to avoid publishing incorrect results, we might want to selectively publish unsurprising findings. How can we resolve the tension between these contradictory recommendations? I will argue that a functionally differentiated publication system can do so, with different outlets focusing on different objectives.

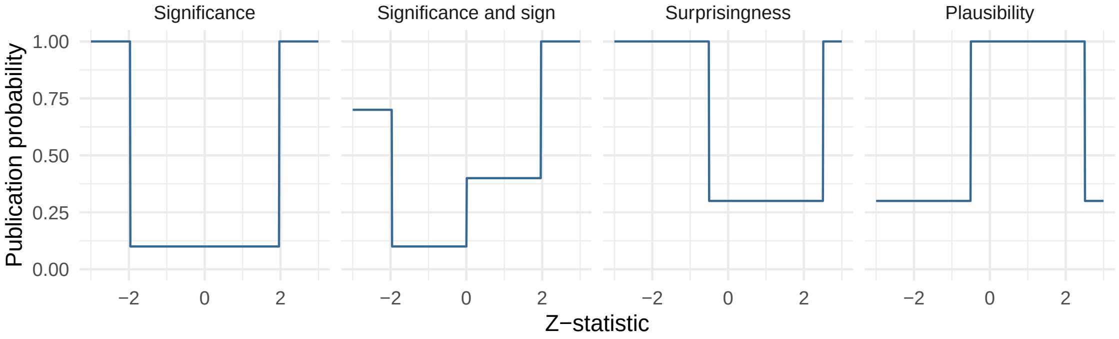
Against the backdrop of these different objectives, I will then discuss some current reform efforts and proposals in greater detail. Emphasizing the reporting of estimates and standard errors rather than statistical significance is beneficial because it discourages binary interpretations of data. This also discourages the selection of findings based on the arbitrary points of 0 effects and 5% percent significance, which seem hardly ever optimal. There are uses of theory in empirical research other than testing; for instance in the construction of shrinkage estimators (Fessler and Kasy, 2019). Requiring pre-analysis plans involves more complex tradeoffs. In game theoretic terms, we can think of pre-analysis plans as a commitment device for researchers, chosen before they see any data. Such commitment devices can be rationalized iff there is a conflict of interest between researchers and journals or readers, and if there are costs to simply reporting all the data. But pre-analysis plans also come with a cost. They prevent the optimal selection of findings to report. Similar arguments apply to pre-results review, where not only researchers commit but journals commit as well, to publish a study independent of its findings. Lastly, new initiatives to launch journals for null results and journals for replication studies could fulfill an important role in a functionally differentiated publication system. They could allow for the existence of a vetted public record of findings that would be an input to meta-studies, while allowing for the existence of selective outlets with a higher profile.

Concluding the present paper, I will briefly take a step back and argue that these debates raise some fundamental questions for statistical theory. In order to coherently discuss these issues, statistical theory needs a model of the work of empirical research that goes beyond the *single-agent* model of statistical decision theory. We should understand statistics (quantitative empirical research) as a *social* process of communication and collective learning that involves many different actors with differences in knowledge and expertise, different objectives, and constraints on their attention and time, who engage in strategic behavior.

Is published research selected?

Forms of selection Let us begin by sketching some forms that selection based on findings might take. Findings might be selected by researchers – which specifications are included in a paper, which outcome variables or controls are considered, etc. This has been described as a "garden of forking paths" of contingent analysis decisions. Findings might also be selected by journals – are null results published, or results that contradict conventional wisdom, etc.

Figure 1: Some possible forms of selection



The most commonly discussed and criticized form of selection is based on significance. Studies might be more likely to be published if their headline finding corresponds to a test-statistic exceeding the 5% critical value, or some other conventional value. The leftmost plot in Figure 1 illustrates. The function $p(.)$ shown is the probability that a finding is written up and published, depending on its z-statistic. In the example shown, the publication probability jumps up once the z-statistic exceeds the critical value of 1.96.

But there are many other forms of selection that might be operational in various fields, which are less often discussed in debates about replicability. In addition to significance, referees might for instance care about whether the estimate has the "right sign," according to theory or conventional beliefs; this might lead to functions $p(.)$ as in the second plot of Figure 1. Researchers or referees might also compare findings to a reference point other than zero. They might for instance value surprisingness relative to some prior mean, as in the third example shown in Figure 1. Or they might do the opposite, and consider findings implausible when they deviate a lot from prior believes, which might lead to selection as in the last example shown. The examples in Figure 1 involve step-functions for illustration; in practice publication probabilities might of course also vary continuously.

Detecting selection In order to discover the presence of p-hacking or publication bias, three methods are commonly used. The first method is based on the p-values corresponding to the headline findings of a set of publications (Brodeur et al., 2016). If the distribution of these p-values across publications shows a discrete jump at values such as 5%, that provides evidence of selection. The second method is based on meta-studies, regressing point-estimates on standard errors (or their inverse) across a set of publications (Card and Krueger, 1995; Egger et al., 1997). If the slope of this regression is different from zero, this again provides evidence of selection. The third method looks at the "rate of replication" for experiments that are repeated with the same protocol (Open Science Collaboration, 2015). The "rate of replication" is defined as the share of published significant estimates for which the replication estimates exceed the significance threshold as well. A low rate of replication is taken as evidence of selection or some other problems.

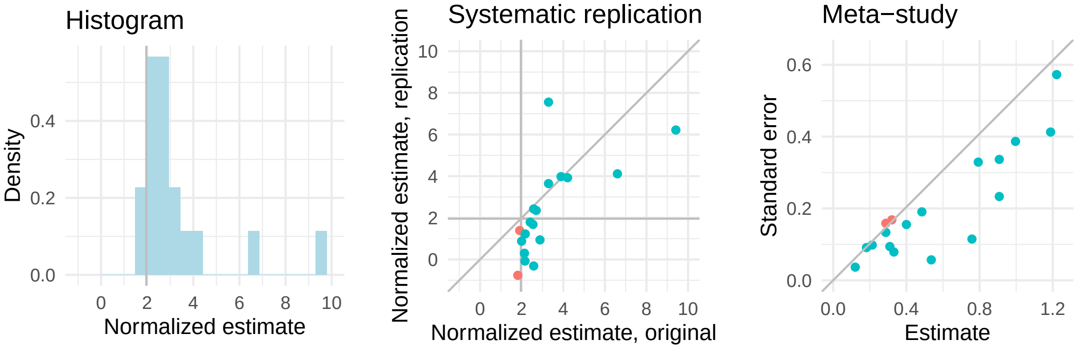
The first two of these methods provide valid tests of the null hypothesis of no selection. They do have some limitations, however. There are many forms of selection that neither of these methods would discover. And these methods cannot recover the form and magnitude of selection. To see why, note that the distribution of published p-values depends not only on selection, but also on the underlying distribution of true effects. A large number of small p-values, for instance, could be due to a large number of null-hypotheses which are indeed false. It could alternatively be due to strong selection on significance. The distribution of p-values does not allow to distinguish between these two explanations. That said, without selection and for continuously distributed test-statistics such as the t-test we could never get a discontinuity in the density of p-values across studies. Such discontinuities thus do provide strong evidence of selection.

What about meta-regressions? These implicitly rely on the assumption that there is no systematic relationship between true effect size and sample size across studies. Even under this assumption, however, there are many forms of selection which do not create a systematic dependence between mean estimates and standard errors, and which can thus not be detected by meta-regressions. A systematic dependence does, however, provide evidence of selection. Additionally, meta-regressions are often used to extrapolate to the hypothetical mean estimate for a standard error of zero (infinite sample size). This extrapolated value is then interpreted as an estimate of the true average effect across published and unpublished studies. The intuition for this interpretation is that it is implicitly assumed that all studies with sufficiently large t-statistics are published, and for small enough standard error that means all studies are published. The problem with this interpretation is that the relationship between average estimates and standard errors is never linear, but extrapolation to zero requires such a functional form restriction.

The "rate of replication" of significant findings, lastly, taken by itself does not tell us much about selection. To see why, consider again the distribution of true effects across studies. Suppose, first, that all true effects are zero. In that case, even without any selective publication or manipulation of findings, only 5% of significant findings would replicate. Suppose, alternatively, that all true effects are very large. In that case, almost all replications of significant findings would turn out significant again, no matter how selective the publication process is.

Estimating selection In Andrews and Kasy (2019), we develop two alternative methods for identifying and estimating the form and magnitude of selection in the publication process. This is a more ambitious goal than just testing for the presence of some form of selection. Identifying the form and magnitude of selection allows us to then assess the magnitude of implied biases, and to correct for them in the interpretation of published findings.

Figure 2: Evidence for selective publication in economics lab experiments



I will use the data provided by Camerer et al. (2016) to provide some intuition for our methods. Camerer et al. (2016) replicated 18 laboratory experiments published in top economics journals in the years 2011 to 2014. Figure 2 plots data from this systematic replication in different ways. The left figure shows the distribution of z-statistics based on the original studies. The distribution of z-statistics exhibits a jump at the cutoff of 1.96, suggesting the presence of selection on significance at the 5% level. The middle figure shows (normalized) original and replication estimates. The right figure shows original estimates and their standard errors.

Our first proposed method is based on systematic replication estimates, as shown in the middle figure. In the absence of selective publication, there should be no systematic difference between originally published estimates and replication estimates, so that flipping the axes in the figure should not systematically change the picture (leaving differences in sample size aside). We should in particular find that the points plotted are equally likely to lie above the 45 degree line or below. Selective publication, however, breaks this symmetry. Suppose for instance that significant findings are 10 times more likely to be published than insignificant findings. Then it will be 10 times more likely to observe studies with the combination [original is significant, replication is insignificant] than with the combination [original is insignificant, replication is significant]. More generally, by comparing the density of observations between points with flipped coordinates we can fully pin down (i.e., estimate) the relative probability of publication between any pair of values of the estimate.

Our second proposed method is based on only the estimates and standard errors from the original studies, and is illustrated in the right figure. This method relies on slightly stronger assumptions and builds on the idea of meta-regressions. In the absence of selective publication, estimates for studies with higher standard errors (smaller sample sizes) should be more dispersed. More specifically, if we take estimates from studies with smaller standard errors and add normal noise of the appropriate magnitude, we should recover the distribution of estimates for studies with larger standard errors. Deviations from this prediction again allow to fully pin down (estimate) the mapping from estimates to publication probabilities.[1]

Estimates of selective publication based on systematic replication studies are valid under very weak assumptions. The estimates based on meta-studies, while relying on stronger assumptions, are much more widely applicable. In settings where we could apply both, we found that both methods yield almost identical estimates.

Which findings should be published?

The debates around the perceived replication crisis in various disciplines have led to the the creation of a number of projects, initiatives, and centers that aim to improve the transparency and reproducibility of research. These initiatives include, among others, the project on *Reproducibility and Replicability in Science* by the National Academy of Science, the *Berkeley Initiative for Transparency in the Social Sciences*, the *Institute for Quantitative Social Science* at Harvard, the *Meta-Research Innovation Center at Stanford*, as well as *Teaching Integrity in Empirical Research*, spanning several institutions. The reforms that have been promoted by these initiatives, and by others, include changes in norms (don't put "stars" based on significance in your tables), changes in journal policies (requiring pre-analysis plans for experimental research, accepting papers based on registered reports), and changes in the institutional infrastructure for academic research (journals for null-results, journals for replication studies). How should we assess these proposals?

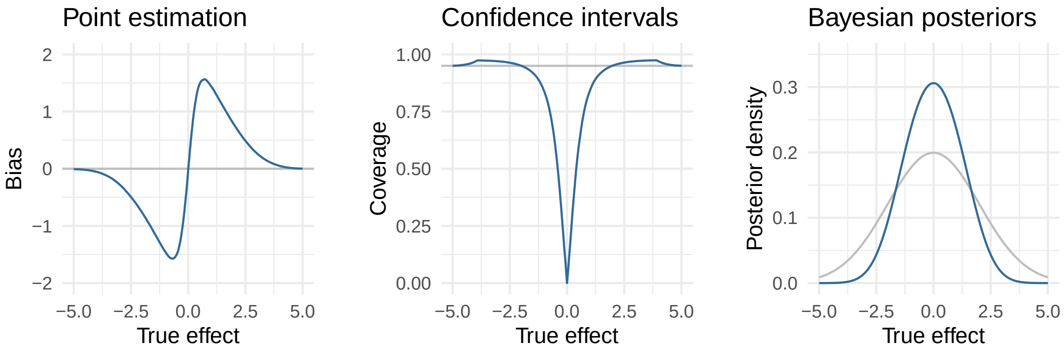
In order to provide a systematic evaluation of these reform attempts, I will first take a step back and discuss several alternative objectives that we might wish to pursue in reforming statistics education and the publication system. I will then argue that these alternative objectives and their contradictory implications might be reconciled in a functionally differentiated publication system. Thereafter, I will consider several more specific reform attempts in turn and will try to assess their merits and potential costs.

[1] An app implementing this method, which allows you to estimate selection based on a meta-study, can be found at `https://maxkasy.github.io/home/metastudy/`

Possible objectives

Validity of inference Why is selection of findings for publication, whether by researchers or by journals, a problem? Because it makes all our standard inferences invalid. As we show in Frankel and Kasy (2018), standard inference methods are valid if and only if publication probabilities do not depend on findings in any way (dependence on standard errors is allowed), in canonical settings. Any form of selection leads to biased estimates, distortions of size for tests and confidence sets, and incorrect Bayesian posteriors, if not properly accounted for. As an illustration, consider the extreme case where only findings exceeding the 5% significance threshold of 1.96 (for standard normal estimates) are published. Figure 3 illustrates this case.

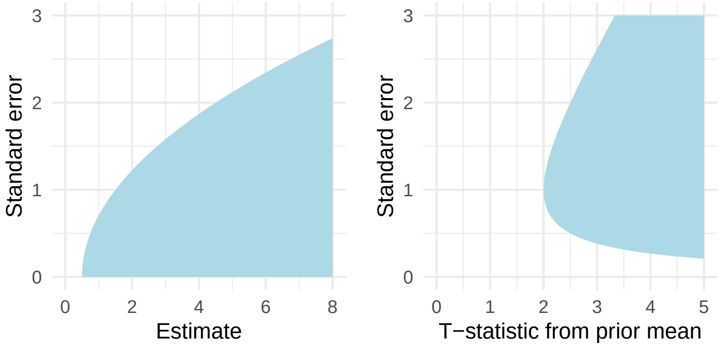
Figure 3: Distortions induced by selective publication



The figure on the left shows the resulting bias of point estimates as a function of the true effect, conditional on publication. For very large true effects no bias occurs, since such studies are published with very high probability. For a true effect of zero, no bias occurs either, by symmetry of selection. For intermediate effect sizes of around 1 standard error, however, point estimates are biased upward by up to 1.5 standard errors, conditional on publication. The middle figure similarly plots the probability of containing the true effect (coverage) of a nominal 95% confidence interval, conditional on the true effect and conditional on publication. For large true effects, again no size distortions happen. For small effects, however, coverage conditional on publication can go all the way down to 0. Consider, lastly, a Bayesian reader of the literature. How should she update her beliefs? When observing a published finding, she need not take into account selection based on that finding, but she needs to update her beliefs in the *absence* of a publication. *Not* observing a publication makes it more likely that the true effect is close to zero. The figure on the right contrasts the Bayesian posterior (for a normal prior) in the absence of publication to the corresponding naive posterior, which ignores selection. Note in particular that the distortions induced by selective publication affect both frequentist and Bayesian inference.

To summarize, selective publication can heavily distort statistical inference, whether frequentist or Bayesian. And there is ample evidence that publication is selective, albeit to different degrees and in different ways across different empirical fields. Recognition of these facts has motivated reform initiatives aimed at increasing the replicability and credibility of published research by reducing selection in the publication process. These are valuable initiatives that are likely to improve the standards of scientific evidence. They do raise, however, the question of what it is that reforms of academic research institutions and norms wish to ultimately achieve? So let us take a step back and consider this question. Clearly, validity of inference should not be our only goal. Presumably we also care about ultimate objectives such as scientific progress, social learning, or helping decision makers in medicine, public policy, and technology. To put it starkly, publishing only estimates calculated based on a random number generator will yield statistical inference that is valid, but presumably not very interesting.

Figure 4: Example of optimal publication regions for binary decisions



Relevance for decision making Consider, as an example, clinical research on new therapies. Suppose that in some hypothetical area of medicine many new therapies, say drugs or surgical methods, are tested in clinical studies. Suppose that most of these trials don't work out - the new therapies just don't deliver. Absent a publication of successful clinical research, no doctor would implement these new therapies. And doctors have limited time - they are not going to read hundreds of studies every month. But they might read some. In this hypothetical scenario, which findings should be published? That is, which subset of studies should doctors read? In order to improve medical practice, it would arguably be best to tell doctors about the small subset of new therapies which were successful in clinical trials. Those are the ones they should incorporate into their practice. Figure 4 illustrates optimal publication regions for an example of this form; see Frankel and Kasy (2018) for details.

If this is the selection rule used for publication, however, published findings are biased upward. Replications of the published clinical trials will systematically find smaller positive effects or even sometimes negative effects. This reasoning suggests that there is a deep tension between relevance (for decision making) and replicability in the design of optimal publication rules. In Frankel and Kasy (2018), we argue that this type of logic holds more generally, in any setting where published research informs decision makers and there is some cost which prevents us from communicating all the data. Such a cost clearly must be present; otherwise it would be optimal to simply publish all data, without any role for statistical inference, researchers, or journals. Given such a cost, it is not worth it to publish "null-results," understood as results that do not change decisions relative to the default absent publication. Surprising results, on the other hand, especially results that lead to large changes of optimal decisions, are of great value to decision makers, and should thus be preferred for publication. Remarkably this conclusion holds whether or not readers are sophisticated in their interpretation of selectively published findings. Furthermore, some notions of social learning, such as reducing the variance of posterior beliefs, are isomorphic to informing decision makers. Therefore similar conclusions go through when our goal is to maximize social learning, subject to attention constraints.

Plausibility We have argued that validity of standard inference requires that we eliminate selection on findings, while (policy) relevance compels us to publish surprising findings. But what about the plausibility of findings? Often we would think that extreme (surprising) findings indicate that there might be some problem with the study design. For instance, if a study reports that a very minor intervention has major health benefits, it might be more likely that the reported findings are biased than that the authors stumbled upon a miracle cure.

To formalize this intuition, suppose that the referees and readers of a study are uncertain about the bias of a study, where the latter might arise for all kinds of reasons. Correct updating of beliefs will imply that the posterior expected bias is increasing in the surprisingness of a finding. Suppose we are again interested in the relevance of findings for decision makers. As before, unsurprising findings are not relevant for decisionmakers and should not be published. But very surprising findings are implausible, suggesting issues with the study, and should also not be published. Only intermediate findings satisfy the requirements of relevance and plausibility.

Achieving multiple objectives in a functionally differentiated publication system

These considerations leave us with the practical question of what to do about the publication system. How shall we trade off these conflicting objectives? Can we have validity, relevance, and plausibility at the same time? A possible solution might be based on a functional differentiation of publication outlets, which could build on the present landscape, while making the differences of objectives and implied publication policies across outlets more explicit. Such a differentiation avoids having to sacrifice one of these objectives (e.g. relevance) for the sake of another (e.g., validity and replicability). The following provides a sketch.

There might be a set of top outlets focused on publishing surprising ("relevant") findings, subject to careful quality vetting by referees. These outlets would have the role of communicating relevant findings to attention-constrained readers (researchers and decision makers). A key feature of these outlets would be that their results are biased, by virtue of being selected based on surprisingness. In fact, this is likely to be true for prominent outlets today, as well. Readers should be aware that this is the case: "Don't take findings published in top outlets at face value."

There might then be another, wider set of outlets that are not supposed to select on findings, but have similar quality-vetting as the top-outlets, thus focusing on validity and replicability. For experimental studies, pre-analysis plans and registered reports (results-blind review) might serve as institutional safe-guards to ensure the absence of selectivity by both researchers and journals. Journals that explicitly invite submission of "null results" might be an important part of this tier of outlets. This wider set of outlets would serve as a repository of available vetted research, and would not be subject to the biases induced by the selectivity of top-outlets. Hiring and promotion decisions should take care to give similar weight to this wider set of publications as to top publications, so as to minimize the incentives for researchers to distort findings, whether by p-hacking or other means.

To make the findings from this wider set of publications available to attention-constrained decision makers, systematic efforts at aggregation of findings in review articles and meta-studies by independent researchers would be of great value (Vivalt, ming; Meager, 2019). Lastly, systematic replication studies can serve as a corrective for the biases of top publications, and as a further safe-guard to check for the presence of selectivity among non-top publications.

Specific reform proposals

De-emphasizing statistical significance Much of traditional statistics teaching, editorial standards, and statistical software focuses on the notion of statistical significance. More recently, however, journals such as those of the American Economic Association have changed their standards to de-emphasize statistical significance, advising authors

> Do not use asterisks to denote significance of estimation results. Report the standard errors in parentheses.

Controversial debates over the notions of statistical testing and statistical significance have a long history, which we will not recapitulate here. A companion paper in this symposium [**INSERT REFERENCE**] reviews some of these debates around statistical significance. For the present article, it is useful to disentangle several distinct aspects of the common emphasis on testing for equality to 0 of some effect or coefficient, controlling size at the 5% level.

First, there is the emphasis on the largely arbitrary value of 0. Arguably, very few effects in the social and life sciences (perhaps in contrast to physics) are exactly equal to 0. Rejecting the null hypothesis of 0 is thus largely a matter of sample size in most applications. Switching the emphasis of teaching and publishing from *tests* to *confidence sets* allows us to move away from the focus on this arbitrary value, while maintaining an easily communicable measure of statistical precision.

Second, there is the emphasis on the arbitrary cutoff of 5%. Established largely by historical accident, there is little reason to assume that this cutoff provides a good tradeoff between size and power in any particular setting. Reporting point estimates and standard errors, as per the AEA guidelines, provides a resolution to this issue. *Point estimates and standard errors* are sufficient statistics for the parameter of interest under conventional normal approximations, so that all the relevant information is communicated. In practice, of course, readers trained to think in terms of significance testing might still calculate a test (in their head), comparing estimates to twice their standard error, thus undoing the effect of the reformed reporting standards.

Third, there is the emphasis on a *binary interpretation* of the data. Following the logic of testing, empirical research is often discussed in terms of whether the authors "found an effect of X on Y" or not. This is a very coarse representations of data that are usually quite complex. Nothing prevents, in principle, less coarse representations, such as point estimates and standard errors, except that the latter are harder to remember. The fact that such coarse representations are popular seems to point to attention constraints, which provide one of the motivations for optimal selection rules as discussed in Frankel and Kasy (2018), and in related work by Andrews and Shapiro (2019). Statistical recommendations should take into account such attention constraints.

Fourth, there is *selective publication* of significant findings. The notion that effects that are significantly different from 0 are more interesting than those that are not appears to drive much of selective publication; see for instance Figure 2 above. Selection on significance bears some resemblance to selection on surprisingness. As we argued above, the latter is optimal for relevance or learning objectives. But neither selection centered at 0 nor selection at the 5% significance cutoff are optimal for relevance, and they lead to distortions of inference.

Motivated by the observation that very few effects in economics are exactly equal to 0, and more generally that few theories can be assumed to hold exactly, Fessler and Kasy (2019) proposed an alternative use of economic theory in empirical research. In this paper, we suggest a framework for the construction of estimators which perform particularly well when the empirical implications of a theory under consideration are approximately correct. Estimators constructed in the proposed way, shrinking toward theoretical restrictions, tend to outperform estimators ignoring the theory, regardless of what the true data generating process is and whether the theory is correct or not. If hypotheses or theories hold approximately, as is often plausible, tests always reject for large samples, while shrinkage estimators perform very well.

Pre-analysis plans Increasingly, pre-analysis plans (PAPs) have become a precondition for the publication of experimental research in economics, for both field experiments and lab experiments. In its adoption of PAPs, the field of economics economics follows the earlier adoption of PAPs for clinical trials, cf. Food and Drug Administration (1998)[2] This change in methodological norms has not gone uncontested; see for instance Coffman and Niederle (2015) and Olken (2015) in this journal for discussions of the costs and benefits of PAPs in experimental economics.

PAPs, in their ideal form, specify a full mapping from data to reported statistics. In practice, they often do not specify a full mapping, but instead constrain the analysis and the results to be reported. The justification of PAPs most commonly invoked is the need for size control in frequentist hypothesis tests. PAPs are considered to be a remedy for the distortions introduced by un-acknowledged multiple hypothesis testing. More generally[3], PAPs are arguably to frequentist notions of bias and size control what randomized controlled trials (RCTs) are to causality – they are necessary for the very definition of these notions.

I suggest, however, to take a slightly different perspective, moving away from statistical testing when thinking about pre-analysis plans. Rather than imposing frequentist size control (or unbiased estimation) as primitive requirements, one might analyze PAPs in the context of a micro-economic model of researchers and journal editors or readers as agents with potentially divergent interests. These agents maximize expected utility, and thus act as Bayesian decision makers. One might model the statistical analysis of data by researchers as a decision tree (or garden of forking paths, in Borges' metaphor). Researchers can sequentially conduct various analyses, and condition their subsequent analyses on the outcomes of the preceding ones. They might pre-specify part or all of this decision tree in a PAP, thus tying their hands. Findings which are reported but stem from analyses that are not pre-specified go under the header of "exploratory analysis." Researchers potentially incur a cost of pre-specifying more complex trees, as well as a cost of conducting actual analyses, and a cost of reporting more findings. Once a report of findings is generated, an editor (reader) gets to see the reported findings. The editor updates her beliefs about some underlying state of the world based on the reported findings. Given her updated believes, she makes a decision (e.g., a publication decision). Then researcher utility and editor utility are realized. Utility for both of them might depend on the editor's decision, as well as on the state of the world.[4]

In this setting, PAPs are effectively a commitment device available to researchers. Researchers can choose to "tie their hands" ex-ante, if doing so is beneficial to them. The potential benefits of a commitment device for researchers are mediated by its impact on the updating of beliefs by editors. In this setting, several conditions are necessary in order for PAPs to be used by researchers in equilibrium. First, there need to be divergent interests between researchers and editors. Absent that, there would be no dynamic inconsistency that would motivate a commitment device. Second, there needs to be a cost of reporting additional findings, for instance an attention cost. Otherwise, journals might simply constrain authors to report all data with no role for statistical analysis. Third, there needs to be a lack of sufficiently constraining social norms, in terms of the analyses to be reported. If such norms were present, they could substitute for a PAP. Eliminating any of these elements from the model sketched above implies that no PAPs will be used in equilibrium.

Pre-results review Pre-analysis plans, at least in theory, eliminate selective reporting of findings by researchers themselves. They do not, however, eliminate selective publication of findings by journals. In an attempt at eliminating the latter, some outlets such as the Journal of Development Economics (JDE) now allow for submission of "registered reports," where studies are approved for publication based on a pre-results review. As described in the author guidelines of the JDE,[5]

[2] There is an interesting parallel to the adoption of RCTs as a method of choice a few years earlier. RCTs were similarly imported from clinical research to economics.

[3] As Andrew Gelman succinctly put it in `https://statmodeling.stat.columbia.edu/2017/03/09/preregistration-like-random-sampling-controlled-experimentation/`

[4] We analyze a model of this form in work in progress with Alex Frankel and Jann Spiess.

[5] See `https://www.bitss.org/wp-content/uploads/2018/03/JDE_RR_Author_Guidelines.pdf`, accessed Dec 20, 2019.

> Stage 1 submissions for pre-results review typically include key background literature and motivation for the study, hypotheses, study procedures, proposed statistical analysis plan, a statistical power analysis, and pilot data (wherever applicable). Following peer review, high-quality Stage 1 submissions are accepted based on pre-results review, after which authors implement the pre-specified research design and submit a full manuscript, including results and discussion sections (Stage 2). This final manuscript is appraised by reviewers for quality assurance and then published, provided that the implementation of the data collection and analysis maintains high standards of quality.

Pre-results review is the policy that most fully implements publication independent of findings, but possibly dependent on sample size, question, method, etc. Such independence of findings is required if our goal is validity of conventional inference. Such independence is not necessarily desirable if our objective also includes other criteria, as discussed above. Pre-results review is a valuable addition, however, to a functionally differentiated publication system, as discussed above.

Journals for null results and replication studies Another recent set of innovations in the publication system are journals dedicated explicitly to null results, or to replication studies. In economics, there are for instance the *Series of Unsurprising Results in Economics (SURE)*, who's guidelines state

> We accept papers from all fields of Economics which have been rejected at a journal indexed in EconLit with the ONLY important reason being that their results are statistically insignificant or otherwise "unsurprising." To document that your paper meets the above eligibility criteria, please send us all referee reports and letters from the editor from the journal where your paper has been rejected.

Such an outlet, focused on unsurprising or insignificant findings, again has a useful role to play in a functionally differentiated publication system. It provides a completion of the record of findings for meta-studies and related exercises. Note that such a role is distinct from a general selection on unsurprisingness. The latter might be a problem for some parts of structural econometrics, in particular, where anecdotally models are often revised until the estimates imply the expected signs and magnitudes for key behavioral elasticities.

There is also the *International Journal for Re-Views in Empirical Economics (IREE)*, with guidelines stating

> The replicated original studies should already have been published in a peer-reviewed journal (listed in Scopus or Web of Science). IREE publishes replication studies independent of their result (successful and failed replications). [...] Replication studies submitted to IREE are subject to a double-blind peer-review.

Again, replications – with the key caveat of being published independent of findings – can provide a useful addition to a differentiated publication system. Among other roles, they allow for a credible assessment of the selectivity of published findings in some subfield, using the methods of Andrews and Kasy (2019). Extrapolation of estimated selectivity to other findings in the same field then allows for bias corrections in the interpretation of these findings. In addition to allowing to assess selectivity, replications might also shed light on effect heterogeneity not captured by standard errors, thus providing insight into the external validity of published estimates.

Conclusion

Summary The research and debates reviewed in this article warrant the following main conclusions. (1) Published research is selected, and we can not, in general, assume that reported estimates are unbiased and tests control size. (2) Conventional methods to detect publication bias have their limitations. But we can identify and estimate the form and magnitude of selection, using either replication studies or meta-studies. (3) Replicability and validity of inference should not be our only goal, and reform efforts focused on this goal alone are misguided. There is a fundamental tension between alternative objectives such as (policy) relevance, plausibility, and replicability. (4) We can resolve this tension by building a functionally differentiated publication system.

Statistics as decision making versus statistics as communication Let us conclude by taking a step back to consider what the debates around replicability and selective publication imply for the foundations of statistics. One of the leading textbook models of statistics is statistical decision theory. The activity of statistics as conceived by decision theory is a rather solitary affair. It's just you and the data, and you have to make some decision based on the data - estimate a parameter, test a hypothesis, etc. This perspective can be extremely useful. It forces us to be explicit about what our objective is, what the action space is, and what prior information we wish to incorporate (in terms of the statistical model, in terms of a Bayesian prior, or in terms of a set of parameters for which we wish to control worst-case risk). And it makes explicit the tradeoffs involved in the choice of any statistical procedure. But this perspective also has severe limitations, as evidenced by the discussions around p-hacking, publication bias, and pre-analysis plans. It is hard to make sense of these discussions from the vantage point of decision theory.

For instance, why don't we simply communicate all the data to the readers of research? If we took decision theory literally, that would indeed be optimal. Just communicating all the data avoids any issues of selection as well as any waste of information. In practice, as consumers of research we of course do prefer to read concise summaries of findings ("X has a big effect on Y, when W holds."), rather than staring at large unprocessed datasets. There is a role for researchers who carefully construct such summaries for readers. But it is hard to make sense of such a role for researchers unless we think of statistics as communication, and unless there is some constraint on the attention or time or information-processing capacity of readers.

Or, to take another example, what is the point of pre-analysis plans? Their purpose is often discussed in terms of multiple hypothesis testing (where reported p-values should account for selective reporting of tested hypotheses), or more generally in terms of the "garden of forking paths" of specification searching. But, taking the perspective of decision theory literally again, there is no obvious role for publicly committing to a pre-analysis plan in order to resolve this issue. Frequentist inference indeed requires knowledge of the mapping from data to reported statistics, for all counterfactual realizations of the data. But researchers might just communicate this mapping at the time of publication. To rationalize publicly registered pre-analysis plans, we again need to consider the social dimension of research, and assume (1) that there is some conflict of interest between researchers and readers, and (2) that there is some (attention) constraint that prevents the reporting of all data.

What these examples illustrate is that statistics (and empirical research more generally) is a social endeavor, involving different researchers, journal editors and referees, readers, policymakers, and others. Taking the social dimension seriously suggests a perspective on statistics where the task of empirical researchers is to provide useful summaries of complex data to their readers in order to promote some form of collective learning. This task is subject to costs of time and attention of researchers, referees, and readers, as well as constraints on social learning in terms of limited information, strategic behavior, the sociology of research, etc. Elaborating this perspective, where statistics gives normative recommendations for empirical practice while taking into account these social constraints, is an exciting task for the years ahead. This endeavor will have to draw, in particular, on microeconomic theory, psychology, and the sociology and history of science.

References

Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94.

Andrews, I. and Shapiro, J. (2019). Statistical reports for remote agents. *Working Paper.*

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, 8(1):1–32.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Card, D. and Krueger, A. B. (1995). Time-series minimum-wage studies: A meta-analysis. *American Economic Review*, 85(2):238–243.

Coffman, L. C. and Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, 29(3):81–98.

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109):629–634.

Fessler, P. and Kasy, M. (2019). How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions. *The Review of Economics and Statistics*, 101(4):681–698.

Food and Drug Administration (1998). Guidance for industry: Statistical principles for clinical trials. *US Department of Health and Human Services.*

Frankel, A. and Kasy, M. (2018). Which findings should be published? *Working Paper.*

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University.*

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8).

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Vivalt, E. (2019, forthcoming). How much can we generalize from impact evaluations? *Journal of the European Economic Association.*