Journal of Economic Perspectives—Volume 35, Number 2—Spring 2021—Pages 1–18

Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It

Maximilian Kasy

The author Jorge Luis Borges wrote a short story in 1941 called "The Garden of Forking Paths." The plot involves (among other elements) a journey in which the road keeps forking; a novel in which when two alternative outcomes arise, both of them happen; and a labyrinth that may or may not have been built. Statisticians have used the metaphor from Borges to convey how empirical research also involves a garden of forking paths: how data is chosen and prepared for use, what variables are the focus of inquiry, what statistical methods are used, what results are emphasized in writing up the study, and what decisions are made by journal editors about publication. If the published results are the outcome of many unobserved forking paths, then conventional estimators, hypothesis tests, and confidence sets in published studies in the social and life sciences may convey a distorted impression (Ioannidis 2005; Gelman and Loken 2013). A possible response to this issue is to "tie researchers' hands," to use another metaphor. By requiring researchers to pick beforehand which of the forking paths they will take, we might be able to restore the validity and replicability of research. Put differently, with their hands tied, researchers are prevented from cherry picking.

Faced with such concerns, applied researchers in the social and life sciences—as well as policymakers—are confronted with two sets of questions that I will address in this paper. 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

■ *Maximilian Kasy is Associate Professor of Economics, University of Oxford, Oxford, United Kingdom. His email address is maximilian.kasy@economics.ox.ac.uk.*

For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at https://doi.org/10.1257/jep.35.2.1.

affected as a consequence? Second, how should we reform the practice and teaching of statistics, as well as the academic publication system, to reduce these problems?

I begin by discussing several methods which have been used in the literature to provide evidence for selective reporting and publication. These methods are based on 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 these three methods can be useful for demonstrating the existence of selective reporting and publication, they do depend on problematic assumptions, and they allow neither estimates of the magnitude nor the form of selection. Thus, I will review two alternative methods proposed by Andrews and Kasy (2019), which allow us to estimate the extent of selective reporting by researchers and selective publication by journals. One of these approaches uses systematic replication experiments and builds on the intuition that, absent selection, original and replication estimates should be distributed symmetrically. The other approach uses meta-studies and builds on the intuition that, absent selection, the distribution of estimates should be more dispersed for findings with larger standard errors. Taken together, these approaches establish that published research in many fields is highly selected.

I will next turn to the debates about how to reform the practice of statistics and the academic publication system. As a starting point, I will argue that there are different justifiable objectives for scientific studies (Frankel and Kasy forthcoming), and that 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. 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 put some emphasis on publishing surprising findings. Yet another 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 put some emphasis on publishing unsurprising findings.

Against the backdrop of these different objectives, I will then discuss some current reform efforts and proposals in greater detail: for example, the push to report estimates and standard errors while de-emphasizing statistical significance, as promoted by the American Economic Association policy of banning "stars" in estimation tables, and the increasingly common requirement of pre-analysis plans which involve tying the hands of researchers in how they will analyze the data, especially in experimental research. There are also new initiatives to launch journals for null results and journals for replication studies that 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.

In conclusion, I will argue that these debates raise some fundamental questions for statistical theory. In order to discuss these issues coherently, statistical theory should seek to understand 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, along with a recognition that these actors engage in strategic behavior.

Is Published Research Selected?

Forms of Selection

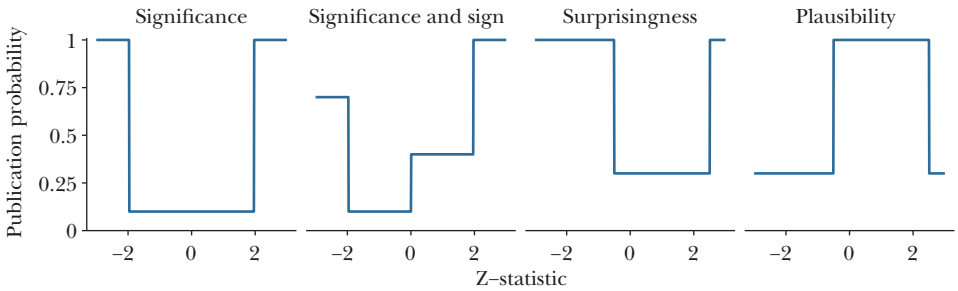
Let us begin by sketching some forms that selection based on findings might take. As noted earlier, findings might be selected by researchers as they navigate the forking paths of a research effort: which specifications are included in a paper, which outcome variables or controls are considered, and so on. Findings might also be selected by journals—for example, are null results published, or results that contradict conventional beliefs? Perhaps the most commonly discussed and criticized form of selection is based on significance. For instance, studies might be more likely to be published if their headline finding corresponds to a test-statistic exceeding the 5 percent critical value or some other conventional value.

Figure 1 illustrates different patterns of selection that might exist in the published literature. Each of the panels in this figure plots a possible dependence of the probability of publication on the z-statistic corresponding to an empirical finding, where the z-statistic is given by the estimate divided by its standard error. The relationship between the z-statistic and the probability of publication can be viewed as a reduced form summary of possible mechanisms driving selection, which might be due to various researcher or journal preferences.

For example, the left-hand panel in Figure 1 shows that if statistical significance at the 5 percent level is the key driver of what is published, then a paper is more likely to be written up if the absolute value of its z-statistic exceeds the critical value of 1.96 (for standard normal estimates): otherwise, the paper is quite unlikely to be written up and/or published. This is the pattern we found in Andrews and Kasy (2019) when analyzing data on lab experiments in economics from Camerer et al. (2016); results significant at the 5 percent level are over 30 times more likely to be published than are insignificant results in this field.

As an alternative, assume that selection occurs both on the basis of statistical significance and also based on whether an estimate has the "right sign," according to theory or conventional beliefs. In this case, as shown in the second panel, statistically significant results with the "right" sign are more likely to be published than significant results of the "wrong" sign, and in addition, statistically insignificant results with the "right" sign have some chance of being published as well. This is the pattern we found in Andrews and Kasy (2019), when analyzing data from Wolfson and Belman (2015). Studies finding a negative and statistically significant effect of minimum wage increases on employment are more likely to be published than either studies finding an insignificant effect or studies finding a positive and significant effect.

Figure 1
Some Possible Forms of Selection



Note: The first two plots show the effect of only publishing significant estimates (with a *z*-statistic above 1.96) on the bias of point estimates (average estimate minus truth) and the coverage of confidence intervals (probability of containing the truth) conditional on publication. The third plot shows the effect on the posterior absent publication.

Researchers or referees might also compare findings to a reference point other than zero. For instance, they might value surprisingness relative to some prior mean. The third panel of Figure 1 shows such a pattern in which "surprising" results are more likely to be published. As argued below, this type of pattern could be optimal when the goal of publication is to inform policy decisions. Or journal editors and referees might do the opposite, and may be disinclined to publish findings that deviate a lot from prior beliefs, because such findings are considered implausible, which might lead to selection as in the last example shown. The examples in Figure 1 are shown as step-functions for illustration only; in practice, publication probabilities might, of course, also vary continuously.

Detecting Selection

To discover the presence of selection—whether it is due to "*p*-hacking" by researchers, or due to 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 percent, that provides evidence of selection. However, this method cannot spot all forms of selection, nor can it 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. For instance, a large number of small *p*-values, suggesting a high degree of statistical significance in the results, could be due to either a large number of null hypotheses that are indeed false, or to strong selection on the basis of significance. Observing a certain distribution of *p*-values in the published literature does not allow one to distinguish between these

two explanations. That said, without selection and for continuously distributed test-statistics such as the *t*-test, one would never expect to find a discontinuity in the density of *p*-values across studies. Such discontinuities thus do provide strong evidence of selection.

The second method for detecting selection is based on meta-studies, which regress point-estimates on standard errors (or their inverse) across a set of publications (Card and Krueger 1995; Egger et al. 1997). The meta-regression approach relies on the assumption that there is no systematic relationship between true effect size and sample size (where sample size will affect standard errors) across studies. Even under this assumption, however, many forms of selection do not create a systematic dependence between mean estimates and standard errors, and can thus not be detected in this approach. A systematic dependence between standard errors and point estimates 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 (corresponding to a hypothetical study with an infinite sample size). This extrapolated value is then interpreted as an estimate of the true average effect across published and unpublished studies. This interpretation is based on the implicit assumption that all studies with sufficiently large *t*-statistics are published, which implies that for small enough standard errors, 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.

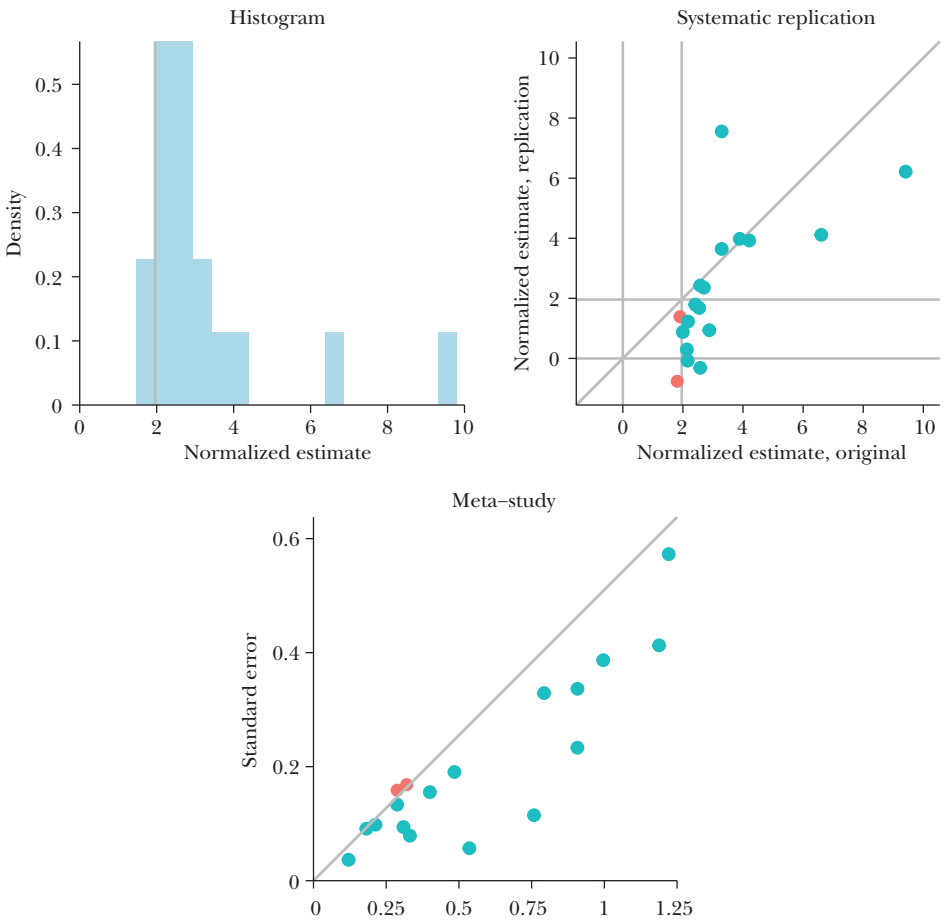
The third method of detecting selection looks at the "rate of replication" for experiments that are repeated with the same protocol, but using different subjects (Open Science Collaboration 2015). The "rate of replication" is defined as the share of published significant estimates for which the replication estimates exceed the significance threshold as well. A low rate of replication is taken as evidence of selection or some other problems. However, the "rate of replication" of significant findings, taken by itself, does not tell us much about selection. To see why, suppose first that all true effects are zero. In that case, even without any selective publication or manipulation of findings, only 5 percent 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 the Form and Magnitude of Selection

In Andrews and Kasy (2019), we develop two alternative methods for identifying and estimating the form and the magnitude of selection in the publication process. Identifying the form and magnitude of selection allows us to assess the magnitude of implied biases and to correct for them in the interpretation of published findings.

I will use the data of 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 study in different ways. The left figure shows that the

Figure 2
Evidence for Selective Publication in Economics Lab Experiments



Note: Based on data of Camerer et al. (2016), as explained in the text.

distribution of *z*-statistics in the original studies exhibits a jump at the cutoff of 1.96, suggesting the presence of selection based on significance at the 5 percent level.

The second panel in Figure 2 shows (normalized) original and replication estimates. 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). In particular, we should 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 ten times more likely to be published than insignificant findings. Then it will be ten 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]. This type of pattern is exactly what we find to be the case for the data of Camerer et al. (2016); lab experiments are much more likely to be published if they find significant effects.

In Andrews and Kasy (2019), we propose a model that allows for an arbitrary distribution of true effects across studies and for an arbitrary function mapping *z*-statistics into publication probabilities (as in Figure 1). This model can be nonparametrically identified and estimated using replication data such as those of Camerer et al. (2016). We can therefore learn from the data how much selection there is and what form it takes. To implement this idea in practice, we propose to assume parametric models: for instance, a step function with jumps at conventional significance levels for publication probabilities, and a *t*-distribution, recentered and scaled with unknown degrees of freedom, for the distribution of true effects across studies. The parameters of such a model can be estimated using maximum likelihood.

The second method proposed in Andrews and Kasy (2019) only relies on the original estimates and their standard errors and does not need replication studies. This method is illustrated in the last panel of Figure 2. 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 (and thus 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 us to pin down fully (estimate) the mapping from estimates to publication probabilities. We propose a model that again allows for an arbitrary distribution of true effects across studies and for an arbitrary function mapping *z*-statistics into publication probabilities, but now assume additionally that standard errors are independent of true effects across studies. This model, or a parametric specification thereof, can be estimated using the data of any meta-study which records estimates and standard errors for different studies. Using this approach, we can again learn how much selection there is, and what form it takes. That is, we can learn what the function mapping z-statistics into publication probabilities looks like.[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 approaches, we found that both methods yield almost identical estimates.

[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/. The source code for this app is available at https://github.com/maxkasy/MetaStudiesApp.

Possible Objectives for Reforms of the Publication System

Motivated by concerns about publication bias and replicability, a number of current projects, initiatives, and centers are seeking to improve the transparency and reproducibility of research. These initiatives include 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, and Teaching Integrity in Empirical Research, spanning several institutions. The reforms that have been promoted by these initiatives and others include changes in norms (don't put "stars" based on statistical 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 and journals for replication studies). We will assess these proposals in the next section. But before doing so, it is useful to take a step back and discuss several alternative objectives that we might wish to pursue in reforming statistics education and the academic publication system: validity, relevance, and plausibility. These alternative objectives can have contradictory implications, which complicates the task of evaluating reforms.

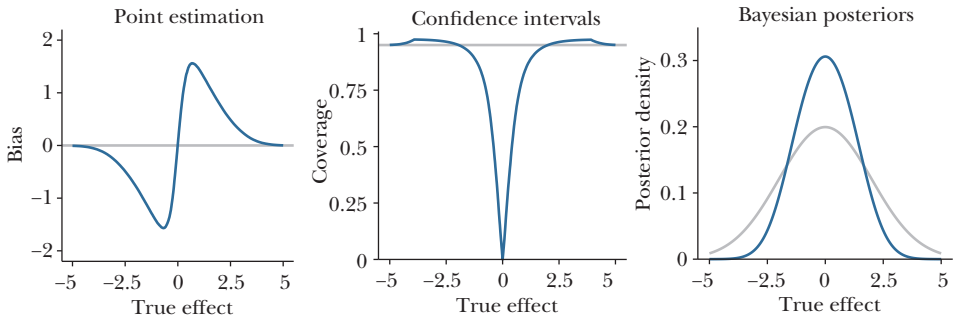
Validity

Why is selection of findings for publication, whether by researchers or by journals, a problem? In canonical settings, standard inference methods are valid if and only if publication probabilities do not depend on findings in any way, although dependence on standard errors is allowed (Frankel and Kasy forthcoming). Any form of selection leads to biased estimates, distortions of size for tests and of coverage for confidence sets, and incorrect Bayesian posteriors—if not properly accounted for.

As an illustration, consider the extreme case where only findings exceeding the 5 percent significance threshold of a z-score of 1.96 (for standard normal estimates) are published. Figure 3 illustrates this case. Each panel in this figure shows the baseline absent selection as a light grey line, and the case of selection as a darker blue line. The first panel shows the bias of point estimates as a function of the true effect, conditional on publication. For very large true effects (whether positive or negative), no bias occurs, because such studies are published with very high probability. For a true effect of zero, no bias occurs either, because positive and negative results are equally likely to be selected. For intermediate effect sizes where the true effect is around 1 standard error, however, point estimates are biased upward by up to 1.5 standard errors from the true value, conditional on publication. This is because studies are only published (in this example) when the estimate exceeds the 5 percent significance threshold.

The middle panel similarly plots the probability that a nominal 95 percent confidence interval contains the true effect, conditional again on the size of the true effect and under the assumption that only results significant at the 5 percent

Figure 3
Distortions Induced by Selective Publication Based on Statistical Significance



Note: The first two plots show the effect of only publishing significant estimates (with a *z*-statistic above 1.96) on the bias of point estimates (average estimate minus truth) and the coverage of confidence intervals (probability of containing the truth) conditional on publication. The third plot shows the effect on the posterior absent publication.

level are published. Again, for large true effects, no distortions happen. When the true effect is small, however, the probability that the confidence interval contains the true effect is much smaller than 95 percent.

Finally, consider a Bayesian reader of the published literature. This reader will update prior beliefs based on the published findings. When observing a published finding, the reader actually does not need to take into account selective publication based on findings. But the reader needs to update beliefs in the absence of a publication! Not observing a publication makes it more likely that the true effect is close to zero, in our example. The last panel of Figure 3 shows two posterior distributions for a Bayesian who starts with a normal prior, when no finding is published (the normal prior is chosen purely for illustration; similar arguments hold for any prior distribution). Relative to the naïve posterior, which ignores selection, the correct posterior that takes selection into account will recognize that the presence of unpublished and unobserved research makes it more likely that the true effect is close to zero, because the Bayesian interprets published findings as a censored sample.[2]

To summarize, there is ample evidence that publication is selective, albeit to different degrees and in different ways across various empirical fields. Selective publication can heavily distort statistical inference, whether frequentist or Bayesian. However, validity of inference should not be the only goal of statistical research. Presumably, researchers 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

[2]Alternatively, we could condition on the number of published findings, leading to a truncation-based perspective with very similar implications for Bayesian inference (Andrews and Kasy 2019).

number generator can yield statistical inference that is valid, but completely useless to decision-makers or substantive researchers.

Relevance for Decision-Making

Consider, as an example, that many new therapies in some hypothetical area of medicine—say drugs or surgical methods—are tested in clinical studies. Suppose that most of these trials don't work out and the new therapies just don't deliver. Absent a publication of successful clinical research, no doctor would implement these new therapies. In addition, doctors have limited time: no human can read hundreds of studies every month. But 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 that were successful in clinical trials. In Frankel and Kasy (forthcoming), we derive optimal publication rules when the goal of publication is to inform decision-makers, as in this example. These optimal publication rules confirm the intuition that findings are most useful for decision-makers when they are surprising, and surprising findings should thus have priority in publication.

However, if the selection rule for publication is based on success in a clinical trial, then 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 publication rules. In Frankel and Kasy (forthcoming), 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 worthwhile to publish "null results"—that is results that do not change decisions relative to the default absent publication. Surprising results, on the other hand, especially those that lead to large changes of optimal decisions, are of great value to decision-makers, and should thus be preferred for publication. 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 the goal of informing decision-makers. Therefore, similar conclusions hold when our goal is to maximize social learning, subject to attention constraints.

Plausibility

Validity of standard inference requires that we eliminate selection on findings, while (policy) relevance encourages us to publish surprising findings. But what about the plausibility of findings? After all, extreme or surprising findings may just indicate that there is some problem with the study design. 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.

We can formalize this idea by assuming that readers have some prior distribution over the bias of a study, that is, some prior probability that the study design is flawed. Very surprising findings make it more likely that the bias is large. Very surprising findings therefore lead to less updating of beliefs about the true effect relative to moderate findings.

Suppose now that we are again interested in the relevance of findings for decision-makers. As before, unsurprising findings are not relevant for decision-makers and should not be published. But very surprising findings are implausible, suggesting issues with the study, and should also not be published. Under this model, only intermediate findings satisfy the requirements of both relevance and plausibility.

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? As argued below, 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 (like relevance) for the sake of another (like validity and replicability). But before we get there, let us discuss some specific reform proposals, while keeping in mind the tension between these objectives.

Specific Reform Proposals

Deemphasizing Statistical Significance

Much of traditional statistics—including teaching, editorial guidelines, and statistical software—focuses on the notion of statistical significance. However, a number of academic journals have recently changed their guidance to de-emphasize statistical significance. For example, the American Economic Association advises prospective authors: "Do not use asterisks to denote significance of estimation results. Report the standard errors in parentheses." [3]

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 by Guido Imbens reviews some issues in these debates.) But for present purposes, it is useful to disentangle four distinct aspects of the common emphasis on testing whether some effect or coefficient is different from a null effect of zero at the 5 percent statistical significance level, before returning to the question of selective publication.

First, there is the emphasis on the largely arbitrary null hypothesis that the true value equals zero, when evaluating estimated results. Arguably, very few effects in the social and life sciences (perhaps in contrast to physics) are exactly equal to

[3] At https://www.aeaweb.org/journals/aer/submissions/accepted-articles/styleguide, accessed January 19, 2021.

zero. For this reason, rejecting the null hypothesis of zero is thus largely a matter of sample size in most applications; with large enough samples, the null hypothesis will always be rejected, because it is wrong. Switching the emphasis of teaching and publishing from significance 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, the 5 percent cutoff for statistical significance is arbitrary, and there is little reason to assume that this cutoff provides a good tradeoff between size and power, that is, between type I errors and type II errors. 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, statistical testing imposes a binary interpretation of the data. 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 representation of data that are usually quite complex. Nothing prevents, in principle, less coarse representations, such as point estimates and standard errors, except perhaps that the latter are harder to summarize or remember. However, 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 (forthcoming) and in related work by Andrews and Shapiro (2019). Statistical recommendations should take such attention constraints into account.

Fourth, the focus on statistical significance is a major factor driving selective publication, motivated by the notion that effects that are significantly different from zero are somehow more interesting than those that are not. Selection on significance bears some resemblance to selection on surprisingness, which matters for relevance or learning objectives (as discussed above). But neither selection centered at zero nor selection at the 5 percent significance cutoff are optimal for relevance, and they lead to distortions of inference. Selection based on significance should thus be avoided.

Motivated by the observation that very few effects in economics are exactly equal to zero and more generally that few theories can be assumed to hold exactly, Fessler and Kasy (2019) propose an alternative use of economic theory in empirical research that does not involve conventional statistical testing. Instead, 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. Our proposed estimators "shrink" empirical findings towards the predictions of a theory. As an example, we might shrink estimated demand functions toward the theoretical prediction that compensated own-price elasticities of demand are negative. By choosing the amount of shrinkage in a data-dependent

manner, we can construct estimators that perform uniformly well and have large gains in performance when the theoretical predictions are approximately correct.

Pre-analysis Plans

Pre-analysis plans have increasingly become a precondition for the publication of experimental research in economics, for both field experiments and lab experiments. Historically, economics first imported randomized controlled trials as a method of choice from clinical research, and then a few years later again followed clinical research (for comparison, see the guidelines from the Food and Drug Administration 1998) in an emphasis on pre-analysis plans. This change in methodological norms has not gone uncontested; for a discussion of the costs and benefits of pre-analysis plans in experimental economics, see Coffman and Niederle (2015) and Olken (2015) in this journal, as well as Banerjee et al. (2020).

In their ideal form, pre-analysis plans specify a full mapping from data to what statistics will be reported. In practice, however, pre-analysis plans often do not specify a full mapping from data to reported results, but instead constrain the analysis and the results to be reported. By tying the researcher's hands, pre-analysis plans prevent the researcher from cherry-picking which results to report. They might thus provide a remedy for the distortions introduced by unacknowledged multiple hypothesis testing. Pre-analysis plans arguably play the same role to frequentist notions of bias and size control as randomized controlled trials play to causality—they are necessary for the very definition of these notions.[4]

In ongoing research (Kasy and Spiess 2021), we take a slightly different perspective. Rather than motivating pre-analysis plans in terms of frequentist hypothesis testing, we propose to model statistical inference as a mechanism design problem. To motivate this approach, note that in pure statistical decision theory there is no need for pre-analysis plans. Rational decision-makers have consistent preferences over time, and thus, no need for the commitment device that is provided by a pre-analysis plan. The situation is different, however, when there are multiple agents with conflicting interests. As an example, consider the conflict of interest between pharmaceutical companies that want to sell drugs and the Food and Drug Administration that wants to protect patient health. Another example would be researchers who want to get published (in order to get tenure) and readers of research who want to learn the truth about economic phenomena.

The mechanism design approach proposed in Kasy and Spiess (2021) takes the perspective of a reader of empirical research who wants to implement a statistical decision rule. Not all rules are implementable, however, when researchers have divergent interests and private information. We characterize implementable rules under these constraints and consider the problem of finding optimal statistical decision rules subject to implementability. In such models, there is a role for

[4]Andrew Gelman makes this point succinctly in https://statmodeling.stat.columbia.edu/2017/03/09/preregistration-like-random-sampling-controlled-experimentation/.

pre-analysis plans under some conditions. In particular, if researchers have many choices (degrees of freedom) for their analysis—there are many forking paths—and if communication costs are high (there is a lot of private information), then pre-analysis plans can improve the welfare (statistical risk) of readers. If, on the other hand, researchers face a smaller number of choices and private information is limited, the reader might be better off without requiring a pre-analysis plan.

Pre-results Journal Review

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

Pre-results review is the policy that most fully implements publication decisions that do not depend on findings but possibly depend on the sample size, question, method, and so on. Such independence of publication from findings is required if our goal is validity of conventional inference. However, such independence is not necessarily desirable if our objective also includes other criteria, such as relevance and plausibility.

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. Such journals are made possible, in particular, by the reductions in publication cost that come with online-only publication. Economics, for instance, has the *Series of Unsurprising Results in Economics*. Such an outlet, focused on unsurprising or insignificant findings, has a useful role to play in a functionally differentiated publication system. It provides a completion of the record of published findings that can serve as an input for meta-studies and related exercises. There is also the *International Journal for ReViews in Empirical Economics* (IREE), a journal focused on replication studies.[6] 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, such replications allow for a credible assessment of the selectivity of published findings in some subfield, using for instance 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 us 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.

[5]For instance, see https://www.elsevier.com/__data/promis_misc/JDE_RR_Author_Guidelines.pdf, accessed January 19, 2021.
[6]For instance, see https://www.iree.eu/aims-and-scope, accessed January 19, 2021.

Achieving Multiple Objectives in a Functionally Differentiated Publication System

Above, we have argued that alternative objectives—relevance for decision-makers, statistical validity, plausibility of published findings—can lead to conflicting recommendations for reforms of the publication system. However, we might reconcile these objectives by striving for a functional differentiation of publication outlets. The following provides a sketch of such a landscape.

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 safeguards 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 aggregating findings in review articles and meta-studies by independent researchers would be of great value (Vivalt 2019; Meager 2019). Lastly, systematic replication studies can serve as a corrective for the biases of top publications and as a further safeguard to check for the presence of selectivity among non-top publications.

Summary and Conclusion

Published research is selected through a process that includes both researchers and journals, so that consumers of such research cannot, in general, assume that reported estimates are unbiased, either in their point estimates or their confidence intervals. In this essay, I have argued that 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. I have further argued that replicability and validity of inference should not be our only goal and reform efforts focused on this goal alone are misguided. However, there is a fundamental tension between alternative objectives such as validity, relevance, plausibility, and

replicability. One approach to resolving this tension, at least partially, is to build a functionally differentiated publication system.

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 main foundations of statistics is statistical decision theory. The activity of statistics as conceived by decision theory is a rather solitary affair. There is just the researcher and the data, and the researcher has to make some decision based on the data: estimate a parameter, test a hypothesis, and so on. This perspective can be extremely useful. It forces us to be explicit about our objective, the action space, and what prior information we wish to incorporate (for example, in terms of the statistical model chosen, or in terms of a Bayesian prior, or in terms of a set of parameters for which we wish to control worst-case risk). The decision-theory perspective makes explicit the tradeoffs involved in the choice of any statistical procedure.

But this decision-theory 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 be optimal. After all, 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 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.

Relatedly, what is the point of pre-analysis plans? Their purpose is often discussed 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. Researchers might just communicate how they mapped data to statistics at the time of publication. To rationalize publicly registered pre-analysis plans, we again need to consider the social dimension of research; in ongoing work (Kasy and Spiess 2021) we do so through the lens of mechanism design.

These examples illustrate that statistics (and empirical research more generally) is a social endeavor, involving different researchers, journal editors and referees, readers, policymakers, and others. Taking this 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 social norms of research, and other factors. Elaborating this perspective in which statistics gives normative recommendations for empirical practice, while taking social constraints into account, is an exciting

task for the years ahead. This endeavor will have to draw on a combination of microeconomic theory, psychology, and the sociology and history of science.

■ *This research was funded in part by the Alfred P. Sloan Foundation (under the grant "Publication bias and specification searching. Identification, correction, and reform proposals").*

References

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–94.

Andrews, Isaiah, and Jesse M. Shapiro. 2019. "A Model of Scientific Communication." https://scholar.harvard.edu/files/iandrews/files/audience.pdf.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-analysis Plans for RCTs in Economics." NBER Working Paper 26993.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36.

Card, David, and Alan B. Krueger. 1995. "Time-Series Minimum-Wage Studies: A Meta-analysis." *American Economic Review* 85 (2): 238–43.

Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (3): 81–98.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder. 1997. "Bias in Meta-analysis Detected by a Simple, Graphical Test." *British Medical Journal* 315 (7109): 629–34.

Fessler, Pirmin, and Maximilian Kasy. 2019. "How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions." *The Review of Economics and Statistics* 101 (4): 681–98.

Food and Drug Administration. 1998. *Guidance for Industry: Statistical Principles for Clinical Trials.* Rockville, MD: US Department of Health and Human Services.

Frankel, Alexander, and Maximilian Kasy. Forthcoming. "Which Findings Should Be Published?" *American Economic Journal: Microeconomics.*

Gelman, Andrew, and Eric Loken. 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." http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8).

Kasy, Maximilian, and Jann Spiess. 2021. "Pre-analysis Plans and Mechanism Design." Unpublished.

Meager, Rachael. 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, Benjamin 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).

Vivalt, Eva. 2019. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18 (6): 3045–89.