

Incentive-Compatible Critical Values

Adam McCloskey, Pascal Michailat

May 2020

Statistical hypothesis tests are a cornerstone of scientific research. The tests are informative when their size is properly controlled, so the frequency of rejecting true null hypotheses (type I error) stays below a prespecified nominal level. Publication bias exaggerates test sizes, however. Since scientists can typically only publish results that reject the null hypothesis, they have the incentive to continue conducting studies until attaining rejection. Such p -hacking takes many forms: from collecting additional data to examining multiple regression specifications, all in the search of statistical significance. The process inflates test sizes above their nominal levels because the critical values used to determine rejection assume that test statistics are constructed from a single study—abstracting from p -hacking. This paper addresses the problem by constructing critical values that are compatible with scientists' behavior given their incentives. We assume that researchers conduct studies until finding a test statistic that exceeds the critical value, or until the benefit from conducting an extra study falls below the cost. We then solve for the incentive-compatible critical value (ICCV). When the ICCV is used to determine rejection, readers can be confident that size is controlled at the desired significance level, and that the researcher's response to the incentives delineated by the critical value is accounted for. Since they allow researchers to search for significance among multiple studies, ICCVs are larger than classical critical values. Yet, for a broad range of researcher behaviors and beliefs, ICCVs lie in a fairly narrow range.

McCloskey: University of Colorado–Boulder. Michailat: Brown University. We thank Brian Cadena, Kenneth Chay, Pedro Dal Bo, Stefano DellaVigna, Miles Kimball, Emily Oster, Andriy Norets, Bobak Pakzad-Hurson, and Jesse Shapiro for helpful discussions and comments. This work was supported by the Institute for Advanced Study.

1. Introduction

Statistical hypothesis testing is a key tool for scientific investigation and discovery. It is used to evaluate existing theories and paradigms. In particular it allows for the identification of anomalies: instances where the theory does not accord well with empirical observations. It is also used to assess the effectiveness of new medical treatments, public policies, designs, processes, and other potential remedies to real-world problems. In that context it allows for the detection of novel statistically significant effects.

Scientists conduct many hypothesis tests but they are typically only able to publish those that reject the null hypothesis under investigation. This preference for “significant” results is clearly visible in scientific journals. First identified in psychology journals (Sterling 1959; Bozarth and Roberts 1972), it has since been observed across the social sciences (Christensen, Freese, and Miguel 2019, chap. 3), medical sciences (Begg and Berlin 1988; Song et al. 2000; Ioannidis and Trikalinos 2007; Dwan et al. 2008), biological sciences (Csada, James, and Espie 1996; Jennions and Moeller 2002), and many other disciplines (Head et al. 2015; Fanelli, Costas, and Ioannidis 2017). This preference may reflect the presumption that “significant” results are more valuable to scientific progress than “insignificant” results.

Publications, however, are the main output of scientific research, and the main marker of scientific productivity. A scientist’s publications determine her career path: whether she will receive tenure, whether she will be promoted further, and at each step what salary she will command (Gibson, Anderson, and Tressler 2014). Publications also determine a scientist’s recognition and status in the profession and the resources made available to her such as research funds, grants and fellowships. Such resources are integral to pursuing new projects and obtaining additional publications.

Since publishing yields significant rewards but requires statistically significant results, scientists are given the incentive to collect and analyze data until they are able to reject the null hypothesis they are investigating (Nosek, Spies, and Motyl 2012). Such behavior is well-documented: researchers inflate reported t -statistics by choosing amongst specifications (Brodeur et al. 2016), they fail to write up experimental results that are not statistically significant (Franco, Malhotra, and Simonovits 2014), and they favor reporting statistically significant outcomes at the expense of insignificant outcomes (Dwan et al. 2008). And flexibility in data collection and analysis affords researchers many opportunities to obtain statistically significant results, even when the null hypothesis is true (Lovell 1983; Simmons, Nelson, and Simonsohn 2011).

Given this incentive, scientists’ behavior invalidates the assumptions of classical statistics upon which standard hypothesis testing is based. One key assumption is that scientists report

whatever they observe from a single data sample that is representative of an underlying population of interest. Thus, the use of standard critical values (CVs) based upon this assumption to determine statistical significance in scientific publications leads to the over-rejection of true null hypotheses. For instance, when using a standard CV of 1.96 for a two-sided t -test of nominal level 5%, a true null hypothesis can be rejected much more often than 5% of the time. This lack of size control is one of the most troubling consequences of publication bias, “the systematic error induced in a statistical inference by conditioning on the achievement of publication status” (Begg and Berlin 1988, p. 422).

This consequence of publication bias is problematic because the informativeness of hypothesis tests relies upon properly controlling the size of the test: to limit spurious findings of statistical significance, one must be confident that the frequency of rejecting true null hypotheses does not exceed a pre-specified nominal level. When using data to test theory, if the size of a test is too large, too many apparent anomalies will be “discovered” and perfectly sound paradigms may become disfigured or abandoned. In applied science, if the size is too large, ineffective methods, remedies, or policies will be implemented in the real world, and the problems they were designed to alleviate may persist or even worsen.

To address the size distortions induced by researchers’ incentives, we construct CVs that are compatible with these incentives. Our incentive-compatible CVs (ICCVs) therefore bound test size by the nominal significance level. To construct ICCVs, we model the strategic behavior of researchers. Researchers face costs and benefits from conducting studies, and these incentives determine how many studies researchers conduct.

Imagine for example that a researcher conducts two-sided t -tests at a nominal significance level of 5%. The usual CV for such tests based upon the large sample standard normal distribution of a t -statistic under the null hypothesis is 1.96. Now imagine that the research collects a random sample of data, performs the test, and fails to reject the null hypothesis because the observed absolute t -statistic is below 1.96. Given that such a result is unlikely to be published, if the relative cost is low enough, the researcher may collect another sample of data and analyze it in a further attempt to reject the null hypothesis. This implies that the CV of 1.96, based upon the large sample distribution of a t -statistic from a single study in isolation, is generally not incentive-compatible: it is built upon the assumption that researchers collect one data sample, although it generally incentivizes them to collect more than one (so long as the benefits from publication are large enough).

Indeed, for a given CV, researchers have the incentive to continue collecting data until either their test statistic exceeds the CV, or the expected benefit from data collection falls below the collection cost. A CV that is incentive-compatible needs to take this behavior into account to

bound test size by the nominal significance level. When the ICCV is used, researchers may still conduct several studies. Nevertheless, using information about the cost of conducting research, the rewards to publication, and data-collection process, we can deduce the maximum number of studies that the researcher has the incentive to conduct at any given CV. From this, we can compute an upper bound on the distribution of the reported test statistic, and therefore make sure that the CV achieves the desired significance level. When using our ICCVs, readers can be confident that a published rejection of a true null hypothesis occurs no more frequently than the nominal level of the test.

Because researchers often conduct more than one study before submitting their results to a journal, our ICCVs are necessarily larger than standard CVs. In particular, when scientific rewards are higher or data-collection costs are lower, the ICCVs are larger. Using both theoretical and numerical results, we obtain ICCVs for a broad range of researcher behavior, research costs, publication rewards, and researcher prior knowledge; for two-sided tests with 5% nominal significance level, we find ICCVs between 1.96 and 3 instead of the standard 1.96. For fields with better estimates of the cost and benefits of research and of prior knowledge, it is possible to obtain more precise ICCVs; however, our results indicate that ICCVs are fairly insensitive to these inputs, yielding convenient rules-of-thumb for practical application.

Other methods to control test size: modeling researcher behavior. The issue of publication bias is of course well-known, and several methods have been developed to address various consequences of it (Christensen, Freese, and Miguel 2019; Wasserstein, Schirm, and Lazar 2019). In this paper we focus on one important consequence of publication bias: test size distortions. Such distortions mean that nominal significance levels (say, 5%) understate the actual probability of rejecting a true null hypothesis, such that positive results may be due much more to chance—and much less to the null hypothesis being false—than nominally stated.

To address size distortions, we opt to model researcher behavior and take this behavior into account to undo the distortions. Our approach is related to other approaches in the literature that recognize that researcher behavior may invalidate classical assumptions underlying standard statistical analyses. In particular, Glaeser (2006) assumes that researchers do not report the result of one study, but the most extreme result from $n > 1$ studies, and subsequently proposes corrections to debias estimates arising from researchers' behavior. Another example is Andrews and Kasy (2019) who provide a means of adjusting estimates and confidence intervals to correct for the effects of researchers' behavior in the presence of meta-analyses or replication studies. However in practice, these techniques are likely to unravel: when corrections based upon assumed fixed behavior are implemented, researchers may then conduct additional studies in search of

results that overwhelm the initial corrections. Because our method takes researchers' response to CVs into account, it is not subject to this limitation and does indeed properly control the size of the published test results. In this sense, the CVs we propose are incentive compatible.

Other seemingly related methods are lowering the nominal significance level (Benjamin et al. 2018) or deriving CVs from truncated normal distributions (McCrary, Christensen, and Fanelli 2016). Yet, although these approaches make it harder to reject null hypotheses, they may not control the size of hypothesis tests because researchers may respond to these new significance levels and CVs by conducting even more studies to attain rejection. Finally, although related to ICCVs, multiple-testing corrections are not useful in our setting because readers cannot observe the number of studies conducted by the researcher. Multiple-testing corrections can only be used when the number of studies is fixed and known. Similarly, "always-valid" sequential confidence intervals or p -values could be applied in principle to control size in some contexts that we consider, namely "pooling" data; but these techniques also require that the number of studies conducted by the researcher is known (see Johari, Pekelis, and Walsh 2019; Howard et al. 2019).

Other methods to control test size: constraining researcher behavior. A different approach to controlling test size consists of constraining researchers' behavior through a pre-analysis plan (PAP), so they indeed only conduct one study and report the results from that study (Miguel et al. 2014; Olken 2015; Coffman and Niederle 2015). With such constraints, researchers' behavior conforms to the assumptions from classical statistics and publication bias disappears. In exchange for abiding by a PAP, journals could promise to publish results irrespective of significance as long as the research design is of sufficient quality (Christensen, Freese, and Miguel 2019, pp. 110–112). Results-blind review would eliminate journals' current preference for significance and thus researchers' incentive to work within the boundaries of the PAP to obtain significance.

Christensen, Freese, and Miguel (2019, pp. 107–117) discuss the strengths and limitations of PAPs. An important limitation is that PAPs prevent scientists from engaging in exploratory analysis, although it is often a source of new ideas and scientific discoveries. With observational data, an additional problem is that it would difficult to ensure that the PAP is written before the data are observed.¹

¹Gelman and Loken (2014, p. 464) concur: "For most of our own research projects [a PAP] hardly seems possible: In our many applied research projects, we have learned so much by looking at the data. Our most important hypotheses could never have been formulated ahead of time. . . . In any case, as applied social science researchers we are often analyzing public data on education trends, elections, the economy, and public opinion that have already been studied by others many times before, and it would be close to meaningless to consider preregistration for data with which we are already so familiar. . . . The most valuable statistical analyses often arise only after an iterative process involving the data. Preregistration may be practical in some fields and for some types of problems, but it cannot realistically be a general solution."

The ICCV and PAP approaches therefore appear complementary. Relying on ICCVs could be particularly helpful in two settings. First, because the ICCV approach is much more flexible than the PAP approach, it could be useful for more innovative, less explored research questions, for which it is surely fruitful to let researchers explore the data without any constraints. Second, because the ICCV approach does not require one to have never seen the data, it could be useful with observational data, which are often familiar to researchers (for instance, the time series of unemployment and nominal interest rates in macroeconomics). Finally, PAPs are not entirely enforceable and still allow researchers some latitude to deviate from a strict plan of execution. ICCVs based upon limited forms of p -hacking could be used in conjunction with PAPs to further ensure correct size.

Methods to correct other facets of publication bias. The focus of this paper is on controlling the size of hypothesis tests, which is inflated by publication bias. Publication bias also takes other forms, which our methodology cannot address, and which require additional corrections. For example, our methodology does not address the fact that only significant results are published, and almost all non-significant results remain unpublished, or even unwritten. Such selectivity is a problem for meta-analyses of literatures: a large subset of all studies may never appear in print, biasing the evidence included in meta-analyses (Rosenthal 1979). This bias is different from ours; it would occur even if scientists were not strategic and did not respond to incentives: even then the sample of published studies would be truncated at the significance level. Numerous methods have been developed to address this bias, and they would continue to be useful in meta-analyses even if ICCVs replaced standard CVs.²

2. Research and publication process

We start by modeling the process of research and publication. We assume that researchers in a scientific community are interested in inferring the true value of a parameter $\beta \in \mathbb{R}$. In particular, we focus on problems for which scientists wish to test a particular null hypothesis for the value that β takes. In economics, the parameter β is often a causal or treatment effect, a regression coefficient, or a parameter in a structural model.

For concreteness, we focus on the most common hypotheses used in economics: two-sided

²See Begg and Berlin (1988, sec. 5.3), Lau, Ioannidis, and Schmid (1998), and Stanley (2005) for surveys of publications bias in meta-analysis and solutions to debias the meta-estimates. See Simonsohn, Nelson, and Simmons (2014) and Andrews and Kasy (2019) for recent procedures to tackle publication bias in meta-analysis.

hypotheses of the form

$$(1) \quad H_0 : \beta = \beta_0 \quad \text{versus} \quad H_1 : \beta \neq \beta_0,$$

where H_0 and H_1 constitute the null and alternative hypotheses of interest. In the common application of statistical significance testing, β_0 takes the value of zero. The analysis here extends straightforwardly to one-sided hypotheses (appendix B).

After performing statistical tests of (1), researchers communicate their findings in journals. We assume that journal editors only wish to publish papers that reject the prevailing null hypothesis characterized by H_0 . We characterize journals as wishing to publish “significant” or “non-null” results, as is common practice in economics and many other disciplines. Researchers gather data to form a t -statistic for testing (1) and report the value of this statistic to a journal. Common practice in empirical work is to then deem H_0 as rejected or the result of the research as “statistically significant” if the absolute value of the statistic exceeds an appropriate CV. Researchers receive an expected payoff $v > 0$ for producing a study that rejects H_0 . The value that expected payoff v takes depends upon the set of journals the researcher intends to submit to and the research question itself. For example, a researcher receives a higher expected payoff for submitting their work to a more prestigious set of journals but a lower expected payoff if those journals do not find the research question interesting.

It is common practice to derive the CV that determines rejection of H_0 from the upper quantiles of a standard normal distribution as this is the approximate large-sample distribution of a t -statistic under standard assumptions when H_0 holds. One of these standard assumptions is that only a single model and dataset is used to form the t -statistic. However, this assumption is violated in most applied research. Typically, a researcher looks through multiple model specifications or sets of data to form statistics. Given a researcher’s positive payoff from publishing their work, we work under the assumption that the researcher actually reports the maximum absolute value of a set of latent t -statistics, generated across different sets of data or model specifications, to maximize his chances of rejecting H_0 . Formally, we model the researcher as constructing a sequence of n t -statistics X_1^*, \dots, X_n^* from n “studies” that are unobservable to the editor. As usual, we assume the latent t -statistics take the form

$$(2) \quad X_i^* = \frac{\hat{\beta}_i - \beta_0}{\text{se}(\hat{\beta}_i)},$$

where $\hat{\beta}_i$ denotes the estimator of β in study i and $\text{se}(\hat{\beta}_i)$ denotes a consistent standard error estimator for $\hat{\beta}_i$. After conducting n studies, the researcher reports $X_n = \max\{|X_1^*|, \dots, |X_n^*|\}$

to the journal upon submission of his research article. By standard central limit theorem and standard error consistency arguments, each of the latent t -statistics X_i^* is approximately normally distributed with unit variance when H_0 holds. In particular under H_0 , $X_i^* \sim \mathcal{N}(0, 1)$. Unless $n = 1$, the use of a standard CV equal to the $1 - \alpha/2$ quantile of a standard normal distribution results in over-rejection of a true null hypothesis as the reported statistic is distributed as the maximum absolute value of a sequence of normal random variables. That is, the probability of rejecting H_0 when it is true is greater than the nominal level α . This over-rejection is the type of publication bias we aim to correct.

The researcher determines when to stop conducting additional studies based upon a marginal cost-benefit analysis. After conducting $n - 1$ studies, the researcher receives the expected payoff of $v > 0$ if and only if X_n exceeds the CV z used by the editor to determine a statistically significant rejection of H_0 . If $X_{n-1} < z$, the researcher chooses to conduct the n th study if the expected marginal payoff of conducting this additional study exceeds the expected cost of doing so. Formally, we model the expected benefit of conducting the n th study as

$$v \mathbb{P}(X_n > z \mid \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z),$$

where \mathbb{P} is the researcher's subjective probability measure that incorporates his beliefs on the true value of β and $\mathbf{X}_n^* = (X_1^*, \dots, X_n^*)$.³ We may wish to allow the expected cost of a study to depend upon the specific nature of that study. The costs of studies can vary due factors such as different costs of acquiring data or running experiments or differences in the opportunity costs of different studies. In the most general version of our model, we assume the researcher incurs an expected marginal cost of $c(n) \geq 0$ for conducting the n th study. After $n - 1$ studies, the researcher's expected marginal profit from conducting the n th study is thus

$$\begin{aligned} \mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid \mathbf{X}_{n-1}^*) &= v \mathbb{P}(X_n > z \mid \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n) \\ &= v \mathbb{P}(\max\{|X_n^*|, X_{n-1}\} > z \mid \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n) \\ (3) \qquad \qquad \qquad &= v \mathbb{P}(|X_n^*| > z \mid \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n). \end{aligned}$$

However, the researcher has the option to conduct further studies after the n th study if and only if he chooses to conduct the n th study. Thus, conducting the n th study not only yields the direct payoff in (3) but also the indirect payoff of the option value to conduct further studies, leading to a dynamic decision problem. Nevertheless, in the following proposition we show that if the researcher pursues studies in the order of their direct expected profitability, pursuing

³Conditioning on \mathbf{X}_{n-1}^* does not make sense when $n = 1$. As a notational convention, we set $X_0^* = \mathbf{X}_0^* = 0$.

more profitable studies before less profitable ones, the decision problem is static, involving only whether the marginal profit in (3) is positive.

PROPOSITION 1. *Suppose $v \mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*) - c(n) \geq v \mathbb{P}(|X_{n+j}^*| > z | \mathbf{X}_{n-1}^*) - c(n+j)$ for all $j > 0$. Then the researcher chooses to conduct the next study if and only if $X_{n-1} < z$ and*

$$v \mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*) - c(n) \geq 0.$$

Since it is reasonable to believe that researchers pursue studies in order of their profitability, we model the researcher as conducting the n th study according to whether (3) is positive or negative. Taking the reader or editor's perspective, for a given CV z , we seek to characterize the maximum number of studies that are profitable for the researcher to pursue. That is we seek to characterize the maximum $n = N(z)$ for which $\mathbb{E}(\pi(X_{n-1}, X_n; z, v, c) | \mathbf{X}_{n-1}^*) > 0$ because according to his incentives, the researcher reports $X_{N(z)}$ when faced with the CV z . If the editor chooses the CV threshold by looking at the upper α -quantile of $X_{N(z)}$ under H_0 , she can therefore be confident that when H_0 holds she will receive a submission with probability α . Since $N(z)$ endogenously depends upon z , the editor should then wish to find the value of z at which the probability that $X_{N(z)}$ exceeds z is no larger than α when H_0 is true. This guarantees that over repeated submissions of results in favor of H_1 , less than $100 \cdot \alpha\%$ of these results are false when H_0 is the truth.

Our goal is to characterize the value of the CV $z = z^*$ that, based upon the researcher's incentives, does not lead to over-rejection of a true H_0 . That is, we seek to find the smallest value z^* such that

$$(4) \quad P_{H_0}(X_{N(z^*)} \geq z^*) \leq \alpha,$$

where P_{H_0} denotes the objective probability measure given that the null hypothesis holds.⁴ We define this z^* as the *incentive-compatible critical value* since the researcher has no incentive to conduct additional studies that would result in rejecting a true H_0 with probability larger than the nominal level of α . From the point of view of the editor, the maximum $N(z)$ is a random variable since it depends upon the latent t -statistics X_i^* 's. Nevertheless, if we know the joint distribution of the sequence of latent t -statistics X_i^* 's, we may obtain the distribution of $N(z)$. We now state a general result ensuring that the ICCV z^* is well-defined and providing a sufficient condition under which it yields tests with non-trivial power.

⁴We seek to find the smallest z^* satisfying (4) in order to maximize power subject to controlling size.

TABLE 1. Parameter values in simulations

	Description	Source
1.99	Mean of t -statistic prior	Roth and Murnighan (1978); Murnighan and Roth (1983)
0.40	Standard deviation of t -statistic prior	Roth and Murnighan (1978); Murnighan and Roth (1983)
$v = 5000$	Expected value of a publication	Gibson, Anderson, and Tressler (2014)
$c = 933$	Cost of a study	Dal Bo (2005)
$\epsilon = 1$	Cost elasticity	–

PROPOSITION 2. *Suppose $v \mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*) - c(n) \geq v \mathbb{P}(|X_{n+j}^*| > z | \mathbf{X}_{n-1}^*) - c(n+j)$ for all $j > 0$ and $c(1) > 0$. The ICCV z^* defined in (4), (i) exists and (ii) yields a test with non-zero power if and only if $\mathbb{P}(|X_1^*| > z^*) \geq c(1)/v$.*

We now study different settings that yield different cost functions, subjective probability measures and joint distributions of latent studies, leading us to different conclusions on how z^* ought to be chosen by the journal editor so (4) holds.

3. Illustrative calibration

In the next sections we construct ICCVs under different assumptions about researchers and the research process. In these different cases we derive theoretical results and compute the ICCVs in Monte Carlo simulations. Here, as a preamble, we calibrate the parameters of our model (table 1). We use these calibrated values in the subsequent simulations. The calibration also illustrates how the model can be mapped to the actual research process.

This illustrative calibration is based on Dal Bo (2005). There are several advantages to using this paper for calibration. First, it is based on independent and (nearly) identically distributed laboratory experiments, making it straightforward to model the joint distribution of \mathbf{X}_n^* . Second, everything in the paper is well documented, making calibration of most model parameters straightforward. Third, the paper is based on a specific question, within a well delineated literature, making the calibration of reasonable subjective beliefs of the researcher straightforward.

Conceptually, the null hypothesis explored by Dal Bo is H_0 : people do not account for the future in strategic decisions; the alternative hypothesis is H_1 : people do account for the future in strategic decisions. The hypothesis is explored in eight experiments. Dal Bo compares the level of cooperation in repeated prisoner’s dilemma games that have different probabilities of continuation. In some treatments, the prisoner’s dilemma games are one-shot. The only equilibrium is no cooperation, so players are not expected to cooperate. In other treatments, the probability of continuation—which governs the probability of future interaction between players—is 1/2 or

3/4. In these treatments, while no cooperation remains an equilibrium outcome, cooperation equilibria also appear. Then, the likelihood of cooperation between players is expected to increase. Formally, the statistical parameter of interest β is the difference in the cooperation probability between games in which theory predicts no cooperation and games in which theory predicts possible cooperation. The null and alternative hypotheses become

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0.$$

Here, we model the alternative hypothesis as two-sided, in agreement with convention for empirical work in economics. To test the alternative that the possibility of future interaction increases cooperation, it might be more appropriate to use the upper one-sided alternative $H_1 : \beta > 0$ —a test covered in appendix B.

The first parameter to calibrate is Dal Bo’s subjective prior belief about the mean of the t -statistic, which is equal to $\beta/\text{sd}(\hat{\beta})$. Here, we take an empirical Bayes approach, eliciting Dal Bo’s prior from previous estimates. Before Dal Bo’s article, there were two articles on the same topic: Roth and Murnighan (1978) and Murnighan and Roth (1983). Dal Bo presumably used these two precursors to form his prior beliefs about the true value of β . In fact, Dal Bo (2005, table 1) reports the results from these studies. Although these studies report estimates of β , they do not report the associated t -statistics or standard errors $\text{se}(\hat{\beta})$, which are estimates of $\text{sd}(\hat{\beta})$. However, the matched pairs design of the experiments in these studies, allows us to bound $\text{se}(\hat{\beta})$ from above and below in each study. We then simply take the midpoint of these bounds to equal the estimate of $\text{sd}(\hat{\beta})$ used to calibrate the researcher’s prior for $\beta/\text{sd}(\hat{\beta})$. The details are provided in appendix C.

We obtained four previous values of t -statistics for testing H_0 , $\beta/\text{sd}(\hat{\beta})$, from the four treatments in the two articles. Taking the sample size-weighted average of these values, we obtain the mean of the prior belief on the mean of the t -statistic: $\mathbb{E}(\beta/\text{sd}(\hat{\beta})) = 1.99$.⁵ Taking the (sample size-weighted) sample variance of these values, we obtain the variance of the prior belief on the mean of the t -statistic: $\text{var}(\beta/\text{sd}(\hat{\beta})) = 0.4^2$. Finally, we assume that Dal Bo’s subjective prior follows a normal distribution.⁶

The laboratory where the experiments were conducted was available at no cost, so the marginal cost of an experiment is the cost to pay the participants and the research assistant monitoring them. On average, there are 48.75 subjects in each experiment, and each subject earns \$18.94, so paying the participants costs $48.75 \times \$18.94 = \923.3 . In addition each experiment lasts

⁵The weights account for the fact that one study has 121 subjects while the other has 252 subjects.

⁶From an empirical Bayes perspective, a normal prior can be justified when averaging results of previous studies via a central limit theorem.

one hour, so the research assistant must be paid for that time, which we assume costs \$10. The average cost of an experiment therefore is $c = \$923.30 + \$10 = \$933.30$.

Dal Bo’s objective is to reject the null hypothesis that people are not forward-looking. Such result opens the door to a publication in a top economics journal—and indeed Dal Bo’s work was published in the *American Economic Review* (AER). What is the value of such publication? Using data from the University of California system, Gibson, Anderson, and Tressler (2014) find that the publication of 10 AER-quality pages increases academic earnings by 1.3%, which translates to a net present value over an entire career of about \$28,466. Since the length of Dal Bo’s article is 14 pages, its value is $\$28,466 \times 1.4 = \$39,853$. The same would be true of articles published in the *Quarterly Journal of Economics*, *Econometrica*, and *Journal of Political Economy*. The value of publishing in other journals can be inferred from their relative ranking compared to these top journals (see Gibson, Anderson, and Tressler 2017, table A1). Of course obtaining a significant result is not sufficient to publish in the top journals: each of them publishes less than 5% of the articles they receive (Card and DellaVigna 2013; Zheng and Kaiser 2016). Depending on the correlation structure between the decisions of these top journals, as well as the correlations with decisions of lower-ranked journals, we could compute the expected net present value of submitting a significant result. We approximate this value as $v = \$5,000$. We will see that the cost-benefit ratio, c/v , actually determines the ICCV.

In appendix D, we provide an alternative indirect method for eliciting the implied value of this ratio from the number of studies a researcher would typically conduct without rejecting H_0 before moving on to a different research question. Again using data from Dal Bo (2005), we find a range of implied values consistent with the calibrated value of $c/v = 933/5,000 \approx 0.19$.

4. Baseline: an impossibility result

To build intuition, we begin with the simplest version of our model of research and publication. Consider a researcher who may conduct a sequence of independent studies to determine the value of β , where the estimator of β in study i , $\hat{\beta}_i$, is constructed from the data gathered in study i only. These estimators are statistically identical and, appealing to a standard central limit theorem, approximately distributed as $\mathcal{N}(\beta, \text{var}(\hat{\beta}))$ in large samples. This case would arise for example when the studies are experiments, the sample size in each experiment is the same and the observations in each experiment are drawn from the same underlying population. Since the researcher does not know the true value β , we treat it as a random variable from his point of view and write $\hat{\beta}_i | \beta \stackrel{iid}{\sim} \mathcal{N}(\beta, \text{var}(\hat{\beta}))$. In this setting, the latent t -statistics are independent and

identically distributed approximately following a normal distribution:

$$X_i^* | \theta \stackrel{iid}{\sim} \mathcal{N}(\theta, 1), \quad \text{where} \quad \theta = \frac{\beta - \beta_0}{\text{sd}(\hat{\beta})}$$

is a scaled version of the difference between the true value β and its null version hypothesized β_0 with $\text{sd}(\hat{\beta})$ being the true population standard deviation of $\hat{\beta}$. Note that in this setting, the null and alternative hypotheses (1) are equivalent to

$$(5) \quad H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0.$$

Suppose the researcher incurs a constant cost of $c(n + 1) = c \geq 0$ for conducting each additional study. The researcher has a prior distribution on θ of $F(\theta)$.⁷ This prior distribution may arise, for example, as the result of previous studies conducted by other researchers or as the biases of the researcher. For now we assume that the researcher does not learn about θ in the process of conducting his studies so his subjective distribution on θ does not change as he conducts more studies. Specializing (3) to the current context, the expected marginal profit from conducting an additional study after conducting $n - 1$ studies already is equal to

$$(6) \quad \begin{aligned} \mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) | \mathbf{X}_{n-1}^*) &= v \int \mathbb{P}(|X_n|^* > z | \mathbf{X}_{n-1}^*, \theta) dF(\theta) \mathbb{1}(X_{n-1} < z) - c \\ &= v \int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta) \mathbb{1}(X_{n-1} < z) - c. \end{aligned}$$

At stage $n-1$, the researcher engages in another study if and only if $\mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) | \mathbf{X}_{n-1}^*) > 0$. That is, he will conduct another study if and only if $X_{n-1} < z$ and

$$(7) \quad \int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta) \geq \frac{c}{v}.$$

Now we can make two observations:

REMARK 1. *Depending upon the values c, v, z and θ , either the researcher conducts zero studies or continues to conduct studies indefinitely until $X_n^* > z$.*

REMARK 2. *A low marginal cost c , high payoff v , low CV z or a prior distribution placing large weight on large values of θ lead to engagement in studies indefinitely searching for significance.*

Given remark 1, it is actually impossible for the journal editor to set the CV z to a value so the probability that the observed X_n exceeds z is between zero and one under H_0 . If z is set low

⁷Since $\theta = \frac{\beta - \beta_0}{\text{sd}(\hat{\beta})}$, this prior can be elicited directly from a prior on the original parameter of interest β .

enough for (7) to hold, the researcher continues to conduct studies until he rejects H_0 . Conversely, if the editor sets z high enough so (7) is violated, the researcher never conducts a single study. In other words, depending upon the editor's choice of z , the resulting size of the test is either zero (and so is the power) or one. In both cases, the test becomes uninformative. Therefore, it is impossible to find an ICCV in this setting that yields an informative test. The threshold choice of CV z that dictates which of these two cases pertains is equal to about 3 for the calibrated model inputs from the previous section.

In the next three sections, we show that this impossibility result is a special case that relies on (i) a constant marginal cost, (ii) a researcher that does not learn and (iii) studies that are independent of each other.

5. Sampling increasingly costly data

We now modify the analysis of section 4 by assuming that the expected cost of conducting an additional study is increasing in the number of studies already run. This modification can incorporate situations in which it becomes increasingly costly to collect new data or run new experiments or could simply capture increasing marginal opportunity costs on the part of the researcher. In terms of the marginal cost function introduced in section 2, we assume that $c : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is continuous, strictly increasing, $c(0) = 0$ and $\lim_{n \rightarrow \infty} c(n) = \infty$. These assumptions ensure that new studies eventually become overwhelmingly costly so a researcher eventually stops conducting new studies pertaining to (1) and that it is costless not to conduct a study. The continuity assumption is without loss of generality since c will only be evaluated at integer values.

We maintain all of the other assumptions of section 4 so an analogous analysis leads us to the conclusion that at stage $n - 1$, the researcher engages in the n th study if and only if $X_{n-1} < z$ and

$$(8) \quad \int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta) \geq \frac{c(n)}{v}.$$

For this problem, $N(z)$ is thus equal to the maximum positive integer n for which $X_{n-1} < z$ and the above condition holds under H_0 . If

$$(9) \quad \int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta) < \frac{c(1)}{v},$$

$N(z) = 0$ since it is not profitable to engage in any studies of (1). This occurs when the expected returns to examining the research question are too low because, given the researcher's beliefs characterized by $F(\theta)$, the significance threshold z is too large relative to $c(1)/v$. Since the marginal cost function is strictly increasing in this setting, we obtain the following result that breaks the

impossibility result of the previous section.

PROPOSITION 3. *For any CV z , the researcher eventually stops conducting studies even if he never attains rejection.*

Since X_n is the maximum of the absolute value of n mean-zero iid normal variables under H_0 , $P_{H_0}(X_n < z) = [\Phi(z) - \Phi(-z)]^n$ for $z \geq 0$ so we can rewrite the (implicit) definition of the ICCV in this context as the smallest value z^* such that

$$1 - [\Phi(z^*) - \Phi(-z^*)]^{N(z^*)} \leq \alpha.$$

If the researcher's prior distribution $F(\theta)$ is conjugate with the normal distribution, the computation of $N(z)$ simplifies because the integral in the continuation condition (8) can be expressed in closed form. The proposition below provides three examples of such computational simplification: a mass point prior, a uniform prior, and a normal prior.

PROPOSITION 4. *In the increasing cost setting of this section, $N(z)$ is equal to the maximum positive integer n for which $X_{n-1} < z$ and*

1. *if $dF(\theta) = \delta(\tilde{\theta})d\theta$ for some $\tilde{\theta} \in \mathbb{R}$, where δ denotes the Dirac delta function,*

$$n \leq c^{-1} \left(v[1 - \Phi(z - \tilde{\theta}) + \Phi(-z - \tilde{\theta})] \right);$$

2. *if $dF(\theta) = (b - a)^{-1} \mathbb{1}(a \leq \theta \leq b)d\theta$ for some $a < b$,*

$$n \leq c^{-1} \left(v[1 - (b - a)^{-1} \{ \phi(z - b) - \phi(z - a) + (b - z)\Phi(z - b) + (z - a)\Phi(z - a) - \phi(-z - b) + \phi(-z - a) - (b + z)\Phi(-z - b) + (a + z)\Phi(-z - a) \}] \right);$$

3. *if $dF(\theta) = \sigma^{-1} \phi((\theta - \mu)/\sigma)d\theta$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$,*

$$n \leq c^{-1} \left(v[1 - \Phi((z - \mu)/\sqrt{1 + \sigma^2}) + \Phi((-z - \mu)/\sqrt{1 + \sigma^2})] \right).$$

For each of these special cases, it would be reasonable for the editor to take an empirical Bayes approach and estimate the parameters entering the prior distributions using data from previous studies.

The analysis yields the following remarks:

REMARK 3. *If the expected payoff from rejecting H_0 , v , is larger, the ICCV z^* is larger. This means that editors evaluating research about “important” questions should impose higher CVs. All else*

fixed, researchers have more incentive to search across many studies when the publication payoff is higher. A larger CV can be used by the editor to counteract the higher rate of false rejections induced by these high payoffs by reducing the incentive to conduct many studies.

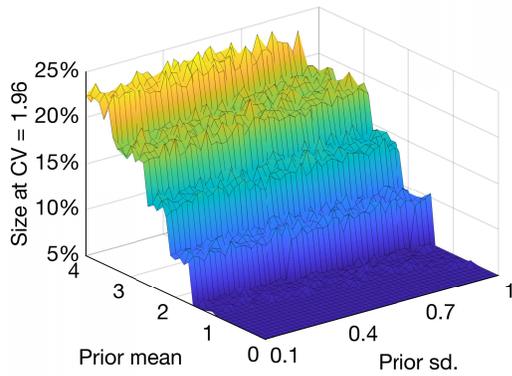
REMARK 4. *If the cost of running experiments, $c(n)$, is larger, the ICCV z^* is smaller. This means that editors evaluating research with high costs should impose smaller CVs (and vice versa). All else fixed, researchers have less incentive to conduct studies when they are costly. A smaller CV can be used by the editor to account for high research costs by increasing the incentive to conduct studies.*

REMARK 5. *The more probability mass a researcher's prior places on large values of θ , the larger the ICCV z^* is. This means that editors evaluating studies for which researchers are more confident that H_0 is violated should impose a larger CVs. When they are more confident that H_0 is false, researchers expect larger payoffs from continuing to conduct studies. All else equal, this gives researchers more incentive to continue conducting studies on (1). Hence, a larger CV can be used by the editor to counteract the corresponding higher rate of false rejections.*

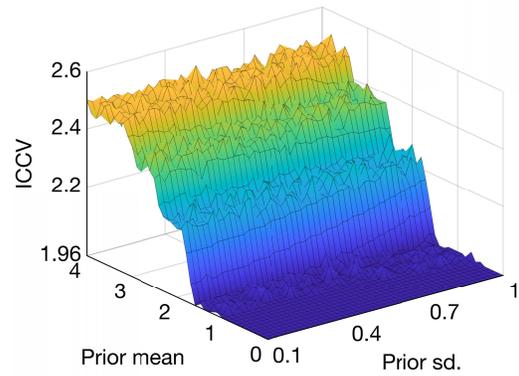
REMARK 6. *In practice, the journal editor may not know the prior of the researcher. However, it is reasonable to estimate this prior from the outcomes of past studies, either parametrically or nonparametrically. If there are no past studies to gather data from, it is reasonable to simply use a diffuse prior.*

The ICCVs with increasingly costly studies, obtained by Monte-Carlo simulations, are illustrated in figure 1. A first insight from the figure is that the publication bias that we aim to correct—the size distortion of hypothesis tests—may be fairly large. The figure displays the size of a two-sided hypothesis test at the standard 5% CV of 1.96, for various priors (panel A), various cost-benefit ratios (panel C), and various cost elasticities (panel E). If researchers conducted only one study and performed the hypothesis test on it, as is assumed in classical statistics, the size would be 5%. But when researchers sample datasets, the size is much higher than 5%: as high as 24% in panel A, 23% in panel C, and 30% in panel E. So a significant result is much more likely to be due to luck than advertised by the nominal significance level of 5%.

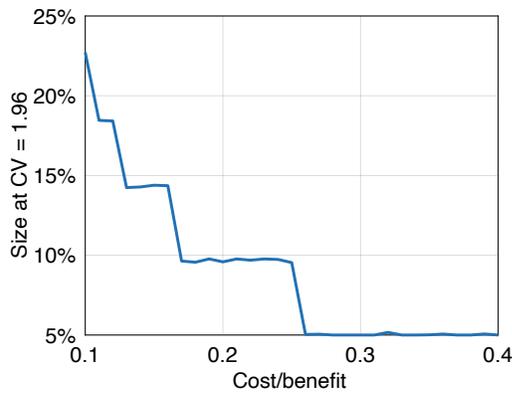
A second insight from the figure is that the ICCVs required to bring back the actual test size to 5% are above the standard value of 1.96 but not by a tremendous amount. The figure displays the ICCV that delivers a size of 5% for two-sided tests, for various priors (panel B), various cost-benefit ratios (panel D), and various cost elasticities (panel F). Across parameterizations, the ICCV is above 1.96 but never above 2.5. Moreover, prior beliefs about the estimated parameter and incentives faced by researchers do not have a strong impact on the ICCV. Across a broad range of priors and cost-benefit ratios from publication, the ICCV lies between 1.96 and 2.5.



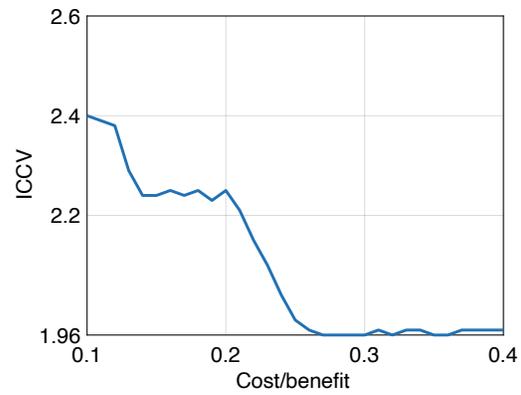
A. Size distortions for various priors



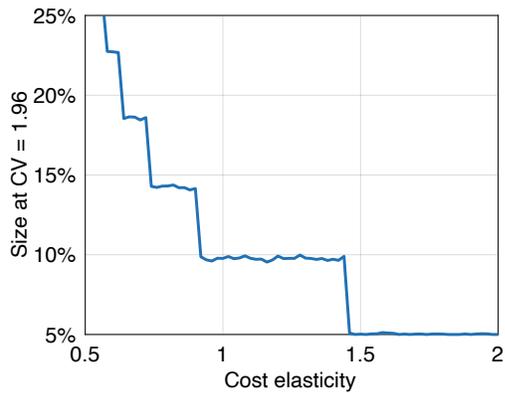
B. ICCVs for various priors



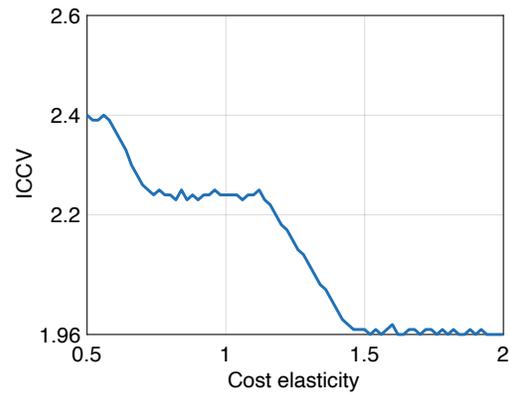
C. Size distortions for various incentives



D. ICCVs for various incentives



E. Size distortions for various cost elasticities



F. ICCVs for various cost elasticities

FIGURE 1. ICCVs at 5% significance level when researchers sample increasingly costly data

This figure considers researchers who attempt to obtain significance in two-sided hypothesis tests by sampling iid datasets that are increasingly costly to collect: collecting the n th dataset costs $c(n) = c_0 \times n^\epsilon$, where ϵ is the cost elasticity. All results are obtained from Monte-Carlo simulations with the parameter values in table 1, except for those parameter values specified on the axes. Panels A, C, and E display the size of the hypothesis test at the standard CV of 1.96, for various priors (panel A), various cost-benefit ratios (panel C), and various cost elasticities (panel E). If researchers collected only one dataset and conducted the test on it, as is assumed in classical statistics, the size would be 5%. But when researchers sample datasets, the size is much higher than 5%, sometimes as high as 25%. Panels B, D, and F display the ICCV that delivers a size of 5%, for various priors (panel B), various cost-benefit ratios (panel D), and various cost elasticities (panel F). Across parameterizations, the ICCV falls between 1.96 and 2.6.

6. Learning while sampling data

We now depart from section 4 by allowing the researcher to learn about the true value of θ . To keep the analysis straightforward we assume that as he conducts successive studies, the researcher updates his prior distribution $F(\theta)$ according to Bayes' rule but maintain all of the other assumptions of section 4. Learning about the true value of θ can induce the researcher to stop conducting additional studies, breaking the negative result of section 4.

After $n - 1$ studies, instead of using his prior distribution $F(\theta)$ to compute the expected marginal profit from conducting an additional study according to (6), the researcher incorporates all of the information contained in the $n - 1$ studies by using the posterior distribution $F(\theta|X_{n-1}^*)$. Assuming that the researcher's prior distribution $F(\theta)$ admits a pdf $f(\theta)$, for $n \geq 1$, the posterior pdf for θ takes the form

$$(10) \quad f(\theta|X_n^*) = \frac{f(X_n^*|\theta)f(\theta)}{\int f(X_n^*|\theta)f(\theta)d\theta},$$

where the likelihood $f(X_n^*|\theta)$ is the pdf of a $\mathcal{N}(\theta\iota_n, I_n)$ distribution evaluated at X_n^* with ι_n denoting an n -vector of one's. This is the joint pdf corresponding to the random vector X_n^* when the value of θ is known. As a notational convention, for $n = 0$, $F(\theta|X_n^*)$ is equal to $F(\theta)$. Maintaining all of the other assumptions in section 4, we arrive at the conclusion that at stage $n - 1$, the researcher engages in the n th study if and only if $X_{n-1} < z$ and

$$(11) \quad \int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta|X_{n-1}^*) \geq \frac{c}{v}.$$

In contrast to the previous section, even if we know $X_{n-1} < z$, whether or not the researcher chooses to engage in another study depends upon the realization of the random vector X_{n-1}^* . For example, if the researcher obtains a large draw for $|X_{n-1}^*|$ that is not quite large enough to cross the threshold z , his posterior for θ will be updated to shift probability mass toward larger values and (11) can hold. On the other hand, a very small draw for $|X_{n-1}^*|$ can induce his posterior to shift probability mass toward very small values so (11) can be violated. These facts break the impossibility result of section 4 so we can define $N(z)$ to be the largest n for which $X_{n-1} < z$ and (11) holds when the data are generated under H_0 . As in proposition 5, we obtain a result that breaks the impossibility result of section 4.

PROPOSITION 5. *Let θ^* denote the true value of θ . For any CV z large enough such that $1 - \Phi(z - \theta^*) + \Phi(z - \theta^*) < c/v$, the researcher eventually stops conducting studies even if he never attains rejection.*

The computation of z^* in this context is complicated by the fact that $F(\theta|X_{n-1}^*)$ depends upon the realizations of prior studies, which are not observed by the journal editor. Nevertheless, since we know the distribution of X_{n-1}^* under H_0 , namely $\mathcal{N}(0, I_{n-1})$, z^* remains feasible to compute in the presence of learning by the researcher. In the following proposition, we examine the special case for which the researcher's prior on θ follows a $\mathcal{N}(\mu, \sigma^2)$ distribution since this distribution results naturally from an empirical Bayes approach to prior construction (via a central limit theorem approximation) and because it provides analytically tractable results that are useful for conveying intuition.

PROPOSITION 6. *In the learning while sampling setting of this section, if $dF(\theta) = \sigma^{-1}\phi((\theta - \mu)/\sigma)d\theta$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$, then $N(z)$ is the largest positive integer n for which $X_{n-1} < z$ and*

$$1 - \Phi\left(\frac{z - \mu_{n|n-1}}{\sqrt{1 + \sigma_{n|n-1}^2}}\right) + \Phi\left(\frac{-z - \mu_{n|n-1}}{\sqrt{1 + \sigma_{n|n-1}^2}}\right) \geq c/v,$$

where

$$\mu_{n|n-1} = \frac{\sigma^2 \sum_{i=1}^{n-1} X_i^* + \mu}{(n-1)\sigma^2 + 1} \quad \text{and} \quad \sigma_{n|n-1}^2 = \frac{\sigma^2}{(n-1)\sigma^2 + 1}.$$

This proposition allows us to make a few observations:

REMARK 7. *The mean of the posterior distribution $\mu_{n|n-1} = \frac{(n-1)\sigma^2}{(n-1)\sigma^2 + 1} \bar{X}_{n-1}^* + \frac{1}{(n-1)\sigma^2 + 1} \mu$ is a weighted average of the sample mean of previous latent studies $\bar{X}_{n-1}^* \equiv (n-1)^{-1} \sum_{i=1}^{n-1} X_i^*$ and the prior mean μ , where the former receives relatively more weight as the number of previous studies $n-1$ increases. For example, in the simplest case of only one previous study and a prior variance equal to the variance of that previous study ($\sigma^2 = 1$), the posterior mean is equal to the arithmetic average of the realization of the previous study and the prior mean.*

REMARK 8. *As the number of previous studies n grows large, the posterior mean $\mu_{n+1|n}$ converges to the sample mean of the previous studies.*

REMARK 9. *As the number of previous studies n grows large, the posterior variance $\sigma_{n+1|n}^2 = \sigma^2/(\sigma^2 + 1)$ shrinks toward zero. In conjunction with remark 8, this implies that for very large n the posterior distribution concentrates heavily around the sample mean of the previous studies, which by the law of large numbers converges to the true value θ . This is simply a consequence of a more general Bernstein-von Mises theorem or standard posterior contraction result from Bayesian statistics. Letting the true value of θ be denoted by θ^* this implies that the researcher must eventually stop conducting studies if $1 - \Phi(z - \theta^*) + \Phi(-z - \theta^*) < c/v$. Since Φ is strictly increasing and $\lim_{x \rightarrow \infty} \Phi(x) = 1$ and $c/v > 0$, this implies that for any θ^* , there is a z large enough that induces the researcher to eventually stop conducting studies.*

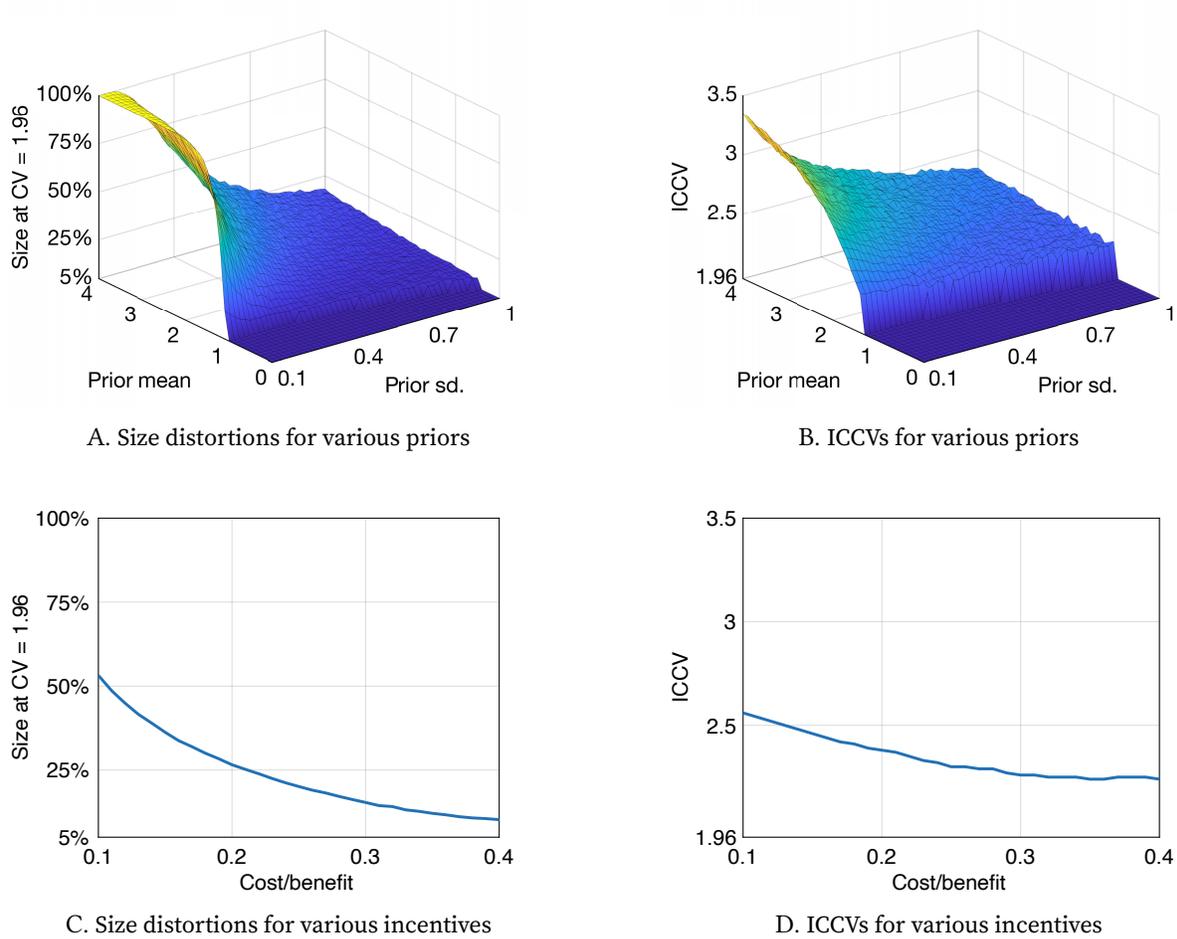


FIGURE 2. ICCVs at 5% significance level when researchers learn while sampling data

The figure considers researchers who attempt to obtain significance in two-sided hypothesis tests by sampling iid datasets, and who update their beliefs about the distribution of the t -statistic after observing the value of the t -statistic in each new dataset. All results are obtained from Monte-Carlo simulations with the parameter values in table 1, except for those parameter values specified on the axes. Panels A and C display the size of the hypothesis test at the standard CV of 1.96, for various priors (panel A) and various cost-benefit ratios (panel C). If researchers collected only one dataset and conducted the test on it, as is assumed in classical statistics, the size would be 5%. But when researchers sample datasets, the size is much higher than 5%, sometimes as high as 100%. Panels B and D display the ICCV that delivers a size of 5%, for various priors (panel B) and various cost-benefit ratios (panel D). Across parameterizations, the ICCV falls between 1.96 and 3.4.

The ICCVs when researchers learn while sampling data, obtained by Monte Carlo simulations, are illustrated in figure 2. As in figure 1, the publication bias that we aim to correct—the size distortion of hypothesis tests—may be fairly large. The figure displays the size of a two-sided hypothesis test at the standard CV of 1.96, for various priors (panel A) and various cost-benefit ratios (panel C). When researchers learn while sampling datasets, the size is much higher than 5%: as high as 100% in panel A and 53% in panel C. So in this context, for a broad range of parameter values, a null hypothesis is more likely to be rejected than not even when it is true (as size is above 50% for a broad range of parameter values).

As in figure 1, the ICCVs required to bring back the actual test size to 5% are above 1.96 but not by a tremendous amount. The figure displays the ICCV that delivers a size of 5% for two-sided tests, for various priors (panel B) and various cost-benefit ratios (panel D). Across parameterizations the ICCV is above 1.96 but never above 3.4—and typically well below 3.

Finally, the incentives faced by researchers do not have a strong impact on the ICCV. Across a broad range of cost-benefit ratios from publication, the ICCV remains between 2.2 and 2.6. On the other hand, prior beliefs about the estimate (especially the prior mean) have a stronger impact on the ICCV than in figure 1.

7. Pooling data

Instead of allowing the researcher to learn about the true value of θ , we take a different departure from section 4 here and have the researcher accumulate data as he moves from one study to the next. To keep the analysis simple, suppose that the i th study involves collecting an additional T random variables with associated constant cost $c(i) = c$ and the researcher’s estimate of β in the i th study is equal to the sample mean of the $n \times T$ independent and identically distributed random variables collected thus far. That is,

$$\hat{\beta}_i = \frac{1}{iT} \sum_{j=1}^{iT} Y_j$$

for some sequence of independent and identically distributed random variables Y_1, \dots, Y_{iT} . In this setting, a standard large sample approximation yields $\hat{\beta}_i | \beta \sim \mathcal{N}(\beta, \text{var}(Y_j)/(iT))$. However, since they use some of the same underlying random variables in their construction, the different $\hat{\beta}_i$ ’s are no longer independent of one another. Instead, they are jointly normally distributed with

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_k) = \frac{1}{kT} \text{var}(Y_j)$$

for $i \leq k$. In turn, the latent t -statistics are jointly normally distributed according to

$$X_i^* | \theta \sim \mathcal{N}(\sqrt{i}\theta, 1), \quad \text{where} \quad \theta = \frac{\sqrt{T}(\beta - \beta_0)}{\text{sd}(Y_j)}$$

and $\text{cov}(X_i^*, X_k^*) = \sqrt{i/k}$ for $i \leq k$.

Due to the correlation between the latent t -statistics as well as their differing means, specializing (3) to the current context changes the expression for the expected marginal profit from conducting an additional study from (6). In this setting,

$$X_n^* | \mathbf{X}_{n-1}^* = \mathbf{x}_{n-1}^*, \theta \sim \mathcal{N}\left(\sqrt{n}\theta + \sqrt{\frac{n-1}{n}}(x_{n-1}^* - \sqrt{n-1}\theta), \frac{1}{n}\right),$$

so that

$$\begin{aligned} (12) \quad \mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) | \mathbf{X}_{n-1}^*) &= v \int \mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*, \theta) dF(\theta) \mathbb{1}(X_{n-1} < z) - c \\ &= v \int \left[1 - \Phi\left(\sqrt{nz} - \theta - \sqrt{n-1}X_{n-1}^*\right) + \Phi\left(-\sqrt{nz} - \theta - \sqrt{n-1}X_{n-1}^*\right) \right] dF(\theta) \mathbb{1}(X_{n-1} < z) - c. \end{aligned}$$

Hence, the researcher engages in the additional study if and only if $X_{n-1} < z$ and

$$(13) \quad \int \left[1 - \Phi\left(\sqrt{nz} - \theta - \sqrt{n-1}X_{n-1}^*\right) + \Phi\left(-\sqrt{nz} - \theta - \sqrt{n-1}X_{n-1}^*\right) \right] dF(\theta) \geq \frac{c}{v}.$$

Similarly to the case with learning, even if we know $X_{n-1} < z$, the researcher's choice to engage in another study depends upon the realization of the previous latent t -statistic X_{n-1}^* . Though the mechanism through which the previous latent t -statistic determines the researcher's choice here is different from the case with learning, the qualitative effect is similar: large enough draws of X_{n-1}^* cause (13) to hold while small enough draws can cause it to be violated. This effect arises because the researcher is accumulating data, rather than updating his prior, so the previous latent t -statistic contains a lot of information about the subsequent t -statistics. As in propositions 3 and 5, we obtain a result that breaks the impossibility result of section 4, this time in the context of pooling data.

PROPOSITION 7. *Under H_0 , for any CV $z > 0$, the researcher will eventually stop conducting studies even if he never attains rejection.*

Since the impossibility result of section 4 is also broken in this context, we similarly define $N(z)$ to be the largest n for which $X_{n-1} < z$ and (13) holds when the data are generated under H_0 . We make a few additional observations:

REMARK 10. *Results analogous to propositions 4 and 6 also apply here. Certain specifications of the researcher’s prior allow us to analytically evaluate the integral on the left hand side of (13) and simplify the expression for $N(z)$.*

REMARK 11. *Since $\sqrt{n-1}X_{n-1}^* = (n-1)\theta^* + \sqrt{n-1}Z$, where θ^* denotes the true value of θ and $Z \sim \mathcal{N}(0, 1)$, the terms \sqrt{nz} and $\sqrt{n-1}X_{n-1}^*$ dominate any finite value of θ in order of magnitude if the researcher has accumulated enough data (for n large enough). This implies that the effect of any prior distribution $F(\theta)$ that only places positive probability mass over finite values of θ (as is standard) on the decision of the researcher eventually disappears as the researcher continues to accumulate data. This feature of the prior distribution being eventually “washed out” from the decision problem is also apparent in the learning context (see remarks 8 and 9). However, this occurs in the current context because the researcher accumulates data, rather than throwing it away in each successive latent study. As more data accumulate, the researcher has more information on the value of future t -statistics.*

REMARK 12. *If H_1 holds so $\theta^* \neq 0$,*

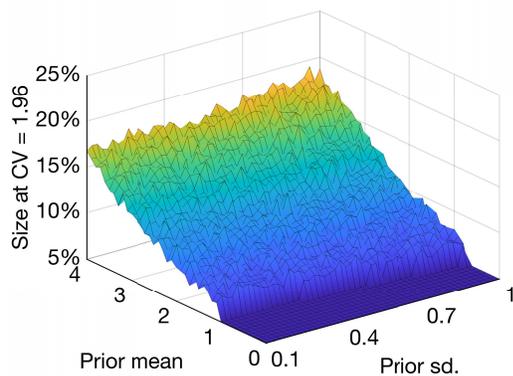
$$\sqrt{nz} - \theta - \sqrt{n-1}X_{n-1}^* = -(n-1)\theta^* + O_p(\sqrt{n})$$

diverges (in probability) for any finite θ as the researcher accumulates more data. This implies that for any standard prior distribution $F(\theta)$, the left hand side of (13) converges to one (in probability) so if $c \leq v$ and the number of studies already conducted is large, there is a high probability that the researcher continues to conduct studies and accumulate data until he can reject H_0 .

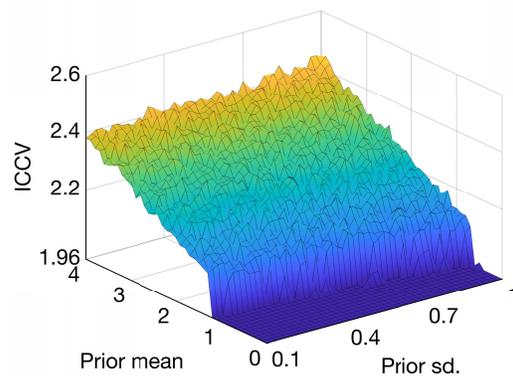
The ICCVs when researchers pool data, obtained by Monte Carlo simulations, are illustrated in figure 3. Here the size distortions we aim to correct are more moderate. Figure 3 displays the size of a two-sided t -test at the standard CV of 1.96 for various priors (panel A) and various cost-benefit ratios (panel C). When researchers pool datasets, the size is higher than 5%, indicating size distortion: as high as 20% in panel A and 10% in panel C. Hence, the size distortions when researchers pool data are less severe than when researchers sample data, as in Figures 1 and 2.

Since the size distortions are smaller, the ICCVs required to bring back the actual test size to 5% are above the standard value of 1.96, but not by much. The figure displays the ICCV that delivers a size of 5% for two-sided tests, for various priors (panel B) and various cost-benefit ratios (panel D). Across parameterizations, the ICCV is above 1.96, but never above 2.5.

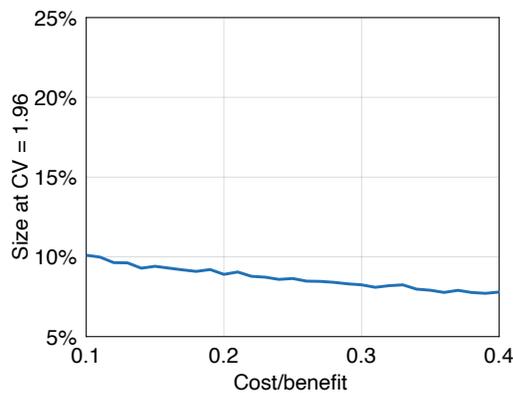
Figure 3 also shows that the incentives faced by researchers barely affect the ICCV. Across a broad range of cost-benefit ratios from publication, the ICCV remains around 2.2 (panel D). Prior beliefs about the estimated parameter (especially the prior mean) have a stronger impact on the



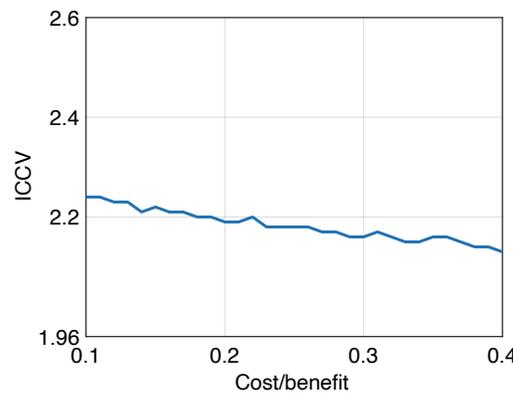
A. Size distortions for various priors



B. ICCVs for various priors



C. Size distortions for various incentives



D. ICCVs for various incentives

FIGURE 3. ICCVs at 5% significance level when researchers pool data

The figure considers researchers who attempt to obtain significance in two-sided hypothesis tests by pooling iid datasets. All results are obtained from Monte-Carlo simulations with the parameter values in table 1, except for those parameter values specified on the axes. Panels A and C display the size of the hypothesis test at the standard CV of 1.96, for various priors (panel A) and various cost-benefit ratios (panel C). If researchers collected only one dataset and conducted the test on it, as is assumed in classical statistics, the size would be 5%. But when researchers sample datasets, the size is higher than 5%, sometimes as high as 20%. Panels B and D display the ICCV that delivers a size of 5%, for various priors (panel B) and various cost-benefit ratios (panel D). Across parameterizations, the ICCV falls between 1.96 and 2.5.

ICCV. However, if researchers were learning while pooling data, the effect of prior beliefs would be dampened.

8. General case

In the general case, we would like to allow for dependence across latent studies and potentially different means of the underlying sequence of latent t -statistics to incorporate observational studies with dependent data, cases for which the researcher adds or subtracts new data to a study, instrument selection when using two-stage least squares, regression specification for ordinary least squares, looking through studies of different precision (resulting in different standard errors), etc. In this more general formulation of the problem, the latent outcome of empirical study $i \geq 1$ is X_i^* , which is approximately distributed according to the distribution of a $\mathcal{N}(\theta_i, 1)$ distribution with $\text{cov}(X_i^*, X_j^*) \equiv \omega_{ij}$ so for any $n \geq 1$, we may write

$$X_n^* \sim \mathcal{N}(\theta_n, \Omega_n) \quad \text{with} \quad \theta_n \equiv [\theta_1, \dots, \theta_n]' \quad \text{and} \quad \Omega_n \equiv [\omega_{ij}]_{i,j=1:n}.$$

In particular, $\theta_n = 0$ under H_0 . We assume that θ_n is unknown to the researcher for all $n \geq 1$ but that Ω_n is known after n studies have been conducted. These assumptions approximate the large sample joint distribution of a set of latent t -statistics for testing (1) in a general framework under which standard errors and correlations between estimators can be consistently estimated. We provide concrete examples of these quantities in different research settings below. With a slight abuse of notation, we denote the researcher's prior distribution on θ_n as $F(\theta_n)$ and also discuss below how a prior distribution on β translates to $F(\theta_n)$.

Sampling data. In the case that each successive study uses an independent sample from the same population to estimate β , from the point of view of the researcher we have the following large-sample distributional approximation: $\hat{\beta}_i | \beta \stackrel{iid}{\sim} \mathcal{N}(\beta, \text{var}(\hat{\beta}_i))$, where $\text{var}(\hat{\beta}_i) = \zeta^2 / T_i$ for some ζ^2 with T_i denoting the sample size of the i th study. In this case, $\text{sd}(\hat{\beta}_i) \approx \zeta / \sqrt{T_i}$ so using the form of the latent t -statistic (2), $\theta_i = \sqrt{T_i}(\beta - \beta_0) / \zeta \geq 0$. Since the studies are independent, $\Omega_n = I_n$. Finally, suppose the researcher has a prior distribution on β . His prior on each individual θ_i is thus the same distribution shifted by β_0 and scaled by $\zeta / \sqrt{T_i}$. From the point of view of the researcher, each θ_i is a shifted and scaled version of the same underlying random variable β , making $F(\theta_n)$ degenerate.

Pooling data. Suppose that each successive study simply adds additional data to the previous study to estimate β , where these additional data are independent and collected from the same

underlying population. All of the analysis of the previous section applies to this case with the exception of $\Omega_n = \mathbf{I}_n$. Instead, ω_{ij} is a decreasing function of $|i - j|$. In particular, assume that we can approximate $\hat{\beta}_i$ by a sample mean:

$$\hat{\beta}_i \approx \frac{1}{T_i} \sum_{i=1}^{T_i} Y_i$$

for some sequence of independent random variables Y_1, \dots, Y_{T_i} . This approximation can be used for instance for standard linear regression estimators. Since the researcher accumulates data when forming each estimate, further assume that $T_i > T_{i-1}$. In this case, we have

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) \approx \frac{1}{T_j} \text{var}(Y_i)$$

for $i \leq j$ so $\omega_{ij} = \sqrt{T_i/T_j}$ for $i \leq j$.

Regression specification. For this example, we assume that the researcher uses ordinary least squares in the standard linear regression model to estimate the effect of interest. In practice, a typical effect of interest corresponds to the population value of a regression coefficient. However, when the researcher uses different regression specifications across different latent studies, the effect of interest is no longer fixed across studies so (1)–(2) do not apply in general. Nevertheless, a simple generalization of (1)–(2) can be used for this example. More specifically, suppose that in the i th study the researcher uses ordinary least squares to estimate a regression coefficient in a regression of y_i on w_i from a set of T independent data points (y_{i1}, \dots, y_{iT}) and (w_{i1}, \dots, w_{iT}) so

$$\hat{\beta}_i = \frac{\sum_{t=1}^T w_{it} y_{it}}{\sum_{t=1}^T w_{it}^2}.$$

Here, w_i represents the regressor of interest after it has been projected off of the space spanned by the covariates included in the i th regression model, allowing for both different specifications of the regressor of interest and covariates across studies. This framework also allows for different specifications of the dependent variable y_i across latent studies. When using different regression specifications across different studies, the researcher implicitly sets the object of interest in the i th study equal to the population regression coefficient

$$\beta_i = \frac{E(w_{it} y_{it})}{E(w_{it}^2)},$$

where E denotes the expectation operator with respect to the true objective probability measure.

From the point of view of the researcher, standard assumptions yield the following large-sample distributional approximation: $\hat{\beta}_i | \beta_i \sim \mathcal{N}(\beta_i, \text{var}(\hat{\beta}_i))$, where $\text{var}(\hat{\beta}_i) = \text{var}(w_{it}y_{it})/TE(w_{it}^2)^2$. When allowing for different specifications across studies (1) must be modified to

$$(14) \quad H_0 : \beta_i = \beta_{i0} \quad \text{versus} \quad H_1 : \beta_i \neq \beta_{i0}$$

and (2) must be correspondingly modified to

$$X_i^* = \frac{\hat{\beta}_i - \beta_{i0}}{\text{se}(\hat{\beta}_i)}$$

so in large samples

$$X_i^* | \theta_i \stackrel{iid}{\sim} \mathcal{N}(\theta_i, 1), \quad \text{where} \quad \theta_i = \frac{\sqrt{T}(\beta_i - \beta_{i0})}{\text{sd}(w_{it}y_{it})/E(w_{it}^2)}.$$

In addition,

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) \approx \frac{\text{cov}(w_{it}y_{it}, w_{jt}y_{jt})}{TE(w_{it}^2)E(w_{jt}^2)}$$

so $\omega_{ij} = \text{cov}(w_{it}y_{it}, w_{jt}y_{jt})/[\text{sd}(w_{it}y_{it}) \text{sd}(w_{jt}y_{jt})]$. In this example, the researcher's prior distribution on $(\beta_1, \dots, \beta_n)$ induces a prior distribution on θ_n via simple shifting and scaling. It may be reasonable in some examples to assume that the researcher has an identical prior distribution for all β_i 's, leading to a degenerate $F(\theta_n)$.

Instrument selection. By modifying some of the definitions in the previous example, we can also cover the case for which the researcher uses two-stage least squares in a standard linear regression model to estimate the effect of interest. Assuming that the instruments are both strong and valid, we can simply modify the definition of w_i to equal the regressor of interest after (i) all regressors have been projected onto the space spanned by the instruments used in the i th study and then (ii) the resulting regressor of interest has been projected off of the space spanned by the covariates included in the i th regression model. In the special case for which the researcher uses the same dependent variable and covariates across all latent studies and only changes the set of instruments used across studies, if the regression model is correctly specified, (1)–(2) continue to hold since each β_i will simply equal the true regression coefficient.

For the general case, we allow the researcher to learn about the true value of the parameter of interest in a potentially limited fashion. The researcher's subjective distribution over θ_n may correspond to his prior $F(\theta_n)$, a fully updated posterior $F(\theta_n | \mathbf{X}_{n-1}^*)$ or some other function of the

data such as a convex combination of his prior and the posterior distribution. We will simply use $F(\theta_n; \mathbf{X}_{n-1}^*)$ to denote the researcher's subjective distribution after $n - 1$ studies.

We may then express the expected marginal profit of the n th study (3) as

$$(15) \quad \begin{aligned} \mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid \mathbf{X}_{n-1}^*) &= v \int \mathbb{P}\left(|X_{n|n-1}^*| > z \mid \theta_n\right) dF(\theta_n; \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n) \\ &= v \int [1 - \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) + \Phi((-z - \theta_{n|n-1})/\omega_{n|n-1})] dF(\theta_n; \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n), \end{aligned}$$

where $X_{n|n-1}^* \sim X_n^* \mid \mathbf{X}_{n-1}^*$ such that $X_{n|n-1}^* \sim \mathcal{N}(\theta_{n|n-1}, \omega_{n|n-1}^2)$ with

$$\begin{aligned} \theta_{n|n-1} &\equiv \theta_n + \Omega_{n,21} \Omega_{n-1}^{-1} [X_{n-1}^* - \theta_{n-1}], \\ \omega_{n|n-1}^2 &\equiv 1 - \Omega_{n,21} \Omega_{n-1}^{-1} \Omega_{n,12}, \\ \Omega_n &\equiv \begin{bmatrix} \Omega_{n-1} & \Omega_{n,12} \\ \Omega_{n,21} & 1 \end{bmatrix}. \end{aligned}$$

Finally, the ICCV z^* is defined in (4) with $N(z)$ being equal to the largest value of n such that $X_{n-1} < z$ and

$$(16) \quad \int [1 - \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) + \Phi((-z - \theta_{n|n-1})/\omega_{n|n-1})] dF(\theta_n; \mathbf{X}_{n-1}^*) \geq \frac{c(n)}{v}.$$

Since we know the distribution of X_{n-1}^* under H_0 , namely $\mathcal{N}(0, \Omega_{n-1})$, when it exists, z^* remains feasible to compute in general: since Ω_n is known for all $n \geq 1$, it is straightforward to compute $P_{H_0}(X_n > z)$ for any given n, z by Monte Carlo simulation.

General results with a normal prior. If the researcher's prior distribution is conjugate with the normal distribution, we can provide an analytical expression for his posterior distribution. In this subsection, we assume that the researcher's prior over θ_n follows a normal distribution. This special case is interesting because it allows us to characterize $N(z)$ and therefore the ICCV defined in (4) more explicitly. Moreover, a normal prior may be natural in many contexts. For example, if the researcher's prior is based upon averaging estimation results from previous studies, a central limit theorem argument leads to an approximate normal distribution for such an average.

PROPOSITION 8. *Suppose that the researcher's prior on θ_n follows a $\mathcal{N}(\mu_n, \Sigma_n)$ distribution for all $n \geq 1$. Then $F(\theta_n \mid \mathbf{X}_{n-1}^*)$ is equal to the distribution function of a multivariate normal distribution with mean*

$$\mu_{n|n-1} = \left(\tilde{\Omega}_n^{-1} + \Sigma_n^{-1} \right)^{-1} \left(\tilde{\Omega}_n^{-1} \mathbf{X}_n^* + \Sigma_n^{-1} \mu_n \right).$$

and covariance matrix $\Sigma_{n|n-1} \equiv (\tilde{\Omega}_n^{-1} + \Sigma_n^{-1})^{-1}$, where

$$\tilde{\Omega}_n^{-1} = \begin{bmatrix} \Omega_{n-1}^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

By the definition of $\tilde{\Omega}_n^{-1}$,

$$\tilde{\Omega}_n^{-1} \mathbf{X}_n^* = \begin{bmatrix} \Omega_{n-1}^{-1} \mathbf{X}_{n-1}^* \\ 0 \end{bmatrix}$$

so this distribution only depends upon the realization of the first $n - 1$ studies.

To accommodate limited or partial learning on behalf of the researcher, we may specify the researcher's subjective distribution after conducting $n - 1$ studies as a convex combination of the his prior and posterior:

$$F(\boldsymbol{\theta}_n; \mathbf{X}_{n-1}^*) = \alpha F(\boldsymbol{\theta}_n) + (1 - \alpha) F(\boldsymbol{\theta}_n | \mathbf{X}_{n-1}^*),$$

where $\alpha \in [0, 1]$ is a parameter that indexes the sophistication of the researcher with $\alpha = 1$ corresponding to an unsophisticated researcher who does not update his beliefs about the parameter of interest, $\alpha = 0$ corresponding to a sophisticated researcher who completely updates his beliefs and $\alpha \in (0, 1)$ corresponding to a partially sophisticated researcher who only partially updates his beliefs. The results of proposition 8 can also be applied to simplify the computation of $N(z)$ according to (16) in this case since, for example,

$$\begin{aligned} \int \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) dF(\boldsymbol{\theta}_n; \mathbf{X}_{n-1}^*) &= \alpha \int \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) dF(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \\ &+ (1 - \alpha) \int \Phi((z - \theta_{n|n-1})/\omega_{n|n-1}) dF(\boldsymbol{\theta}_n | \mathbf{X}_{n-1}^*) d\boldsymbol{\theta}_n. \end{aligned}$$

The integral inside the first term is the expected value of $\Phi((z - \theta_{n|n-1})/\omega_{n|n-1})$ when $\boldsymbol{\theta}_n$ follows a $\mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$ distribution, and the integral inside the second term is the expected value of $\Phi((z - \theta_{n|n-1})/\omega_{n|n-1})$ when $\boldsymbol{\theta}_n$ follows a $\mathcal{N}(\boldsymbol{\mu}_{n|n-1}, \Sigma_{n|n-1})$ distribution.

9. Conclusion

Statistical hypothesis testing is a key tool for scientific investigation and discovery. It is used to evaluate existing paradigms, and to assess the effectiveness of new medical treatments, public policies, and other potential remedies to real-world problems. To be informative, hypothesis test-

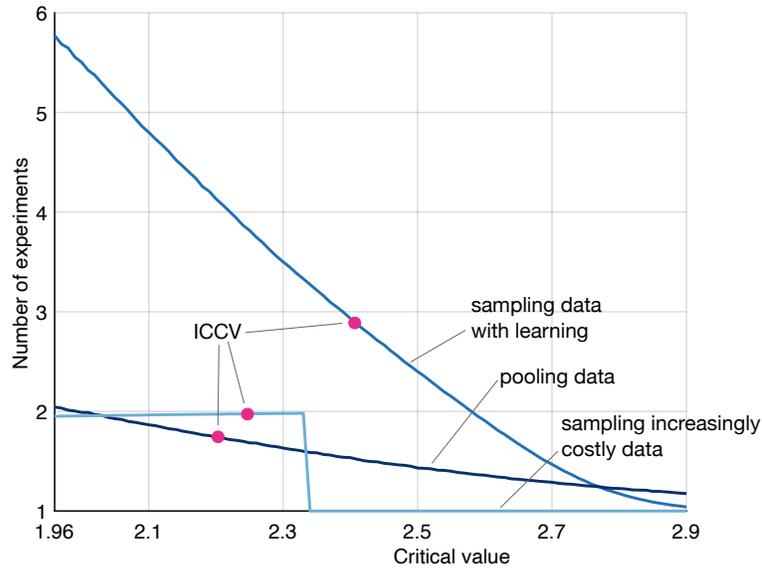


FIGURE 4. Number of experiments for different CVs

The figure displays the average number of experiments conducted by researchers for different CVs (and in particular at the ICCV). The results for the case “sampling increasingly costly data” come from the simulations in figure 1. The results for the case “sampling data with learning” come from the simulations in figure 2. The results for the case “pooling data” come from the simulations in figure 3. In all the simulations, we set parameter values as in table 1.

ing relies on properly controlling size—the probability of rejecting a true null hypothesis. A major issue in modern science, however, is that test sizes in published test results are systematically distorted because scientists are given the incentive to continue to conduct studies until they are able to reject the null hypothesis they are investigating. This is because rejecting a null hypothesis is often considered more interesting from a scientific perspective, and therefore is often required for publication. As a result, a true null hypothesis is much more likely to be rejected than what is advertised in scientific publications.

To correct these size distortions, we construct CVs that are compatible with the incentives faced by scientists and therefore deliver the promised nominal test size. To construct ICCVs, we model the strategic behavior of researchers. Researchers face costs and benefits from collecting data, and these incentives determine how many studies are conducted. Once an ICCV is in place, researchers may conduct several studies to be able to publish their work; nevertheless, readers can be confident that true null hypotheses are not rejected more often than a pre-specified nominal level. Our CVs are larger than standard ones. For example, for two-sided t -tests with a 5% significance level, we find that ICCVs are not 1.96 but between 1.96 and 3 in experimental settings across a wide range of researcher behavior. The exact values of the ICCVs depend on the costs of research, rewards from rejecting the null hypothesis, and the researcher’s prior beliefs. Imposing

the upper bound of 3.4 in this range will control size across all configurations we examined.

Our approach to correcting size distortions in hypothesis tests is to construct CVs that take into account researchers' behavior. Another approach is to constrain researchers' behavior by asking researchers to register their experiments and analysis plans in advance (Christensen, Freese, and Miguel 2019, part 3). These two strategies result in a very different research process. Under ICCVs, researchers pool about 2 studies on average, or sample between 2 and 3 studies, depending on the learning process and cost structure (figure 4). On the other hand, with preregistration, researchers should only conduct one study and report those results. Each approach may be more appropriate in different settings. Ours could be more appropriate for observational data and in the early stage of a research question, when scientific exploration plays a key role. The preregistration approach could be better suited to experimental data and later stages of a research question, when the research question is well understood and delineated, and it is important to obtain precise estimates of the parameters of interest. Moving forward, it would be useful to obtain more information on the parameters of our model to improve the calibration of the ICCVs, and possibly adapt them to specific research methodologies. For example, it would be useful to elicit costs of research methodologies and researchers' prior beliefs. Indeed, ongoing work on researcher prior elicitation such as that summarized in DellaVigna, Pope, and Vivaldi (2019) could prove useful for computing more refined ICCVs than the 1.96–3 range found here.

References

- Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109 (8): 2766–2794.
- Begg, Colin B., and Jesse A. Berlin. 1988. "Publication Bias: a Problem in Interpreting Medical Data." *Journal of the Royal Statistical Society (Series A)* 151 (3): 419–445.
- Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Bjorn Brembs, Lawrence Brown, Colin Camerer et al. 2018. "Redefine Statistical Significance." *Nature Human Behaviour* 2 (1): 6.
- Bozarth, Jerold D., and Ralph R. Roberts. 1972. "Signifying Significant Significance." *American Psychologist* 27 (8): 774–775.
- Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: the Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.
- Card, David, and Stefano DellaVigna. 2013. "Nine Facts about Top Journals in Economics." *Journal of Economic Literature* 51 (1): 144–161.
- Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. Oakland, CA: University of California Press.
- 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.

- Csada, Ryan D., Paul C. James, and Richard H. M. Espie. 1996. "The 'File Drawer Problem' of Non-Significant Results: Does It Apply to Biological Research?" *Oikos* 76 (3): 591–593.
- Dal Bo, Pedro. 2005. "Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games." *American Economic Review* 95 (5): 1591–1604.
- DellaVigna, Stefano, Devin Pope, and Eva Vivaldi. 2019. "Predict Science to Improve Science." *Science* 366 (6464): 428–429.
- Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, Carrol Gamble et al. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLoS ONE* 3 (8): e3081.
- Fanelli, Daniele, Rodrigo Costas, and John P.A. Ioannidis. 2017. "Meta-Assessment of Bias in Science." *Proceedings of the National Academy of Sciences* 114 (14): 3714–3719.
- Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–1505.
- Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460–465.
- Gibson, John, David L. Anderson, and John Tressler. 2014. "Which Journal Rankings Best Explain Academic Salaries? Evidence from the University of California." *Economic Inquiry* 52 (4): 1322–1340.
- Gibson, John, David L. Anderson, and John Tressler. 2017. "Citations or Journal Quality: Which Is Rewarded More in the Academic Labor Market?" *Economic Inquiry* 55 (4): 1945–1965.
- Glaeser, Edward L. 2006. "Researcher Incentives and Empirical Methods." NBER Technical Working Paper 329.
- Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLoS Biology* 13 (3): e1002106.
- Howard, Steven R., Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. 2019. "Uniform, Nonparametric, Non-Asymptotic Confidence Sequences." arXiv:1810.08240.
- Ioannidis, John P.A., and Thomas A. Trikalinos. 2007. "An Exploratory Test for an Excess of Significant Findings." *Clinical Trials* 4 (3): 245–253.
- Jennions, Michael D., and Anders P. Moeller. 2002. "Publication Bias in Ecology and Evolution: An Empirical Assessment Using the 'Trim and Fill' Method." *Biological Reviews* 77 (2): 211–222.
- Johari, Ramesh, Leo Pekelis, and David J. Walsh. 2019. "Always Valid Inference: Continuous Monitoring of A/B Tests." arXiv:1512.04922.
- Lau, Joseph, John P.A. Ioannidis, and Christopher H. Schmid. 1998. "Summing Up Evidence: One Answer Is Not Always Enough." *Lancet* 351 (9096): 123–127.
- Lovell, Michael C. 1983. "Data Mining." *Review of Economics and Statistics* 65 (1): 1–12.
- McCrary, Justin, Garret Christensen, and Daniele Fanelli. 2016. "Conservative Tests under Satisficing Models of Publication Bias." *PLoS ONE* 11 (2): e0149590.
- Miguel, Edward, Colin Camerer, Katherine Casey, Joshua Cohen, Kevin M. Esterling, Alan Gerber, Rachel Glennerster, Don P. Green, Macartan Humphreys, Guido Imbens et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31.
- Murnighan, J. Keith, and Alvin E. Roth. 1983. "Expecting Continued Play in Prisoner's Dilemma Games: A Test of Several Models." *Journal of Conflict Resolution* 27 (2): 279–300.
- Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and

- Practices to Promote Truth Over Publishability.” *Perspectives on Psychological Science* 7 (6): 615–631.
- Olken, Benjamin A. 2015. “Promises and Perils of Pre-Analysis Plans.” *Journal of Economic Perspectives* 29 (3): 61–80.
- Rosenthal, Robert. 1979. “The ‘File Drawer Problem’ and Tolerance for Null Results.” *Psychological Bulletin* 86 (3): 638–641.
- Roth, Alvin E., and J. Keith Murnighan. 1978. “Equilibrium Behavior and Repeated Play of the Prisoner’s Dilemma.” *Journal of Mathematical Psychology* 17 (2): 189–198.
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–1366.
- Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. “P-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143 (2): 534.
- Song, F., A. J. Eastwood, S. Gilbody, L. Duley, and A. J. Sutton. 2000. “Publication and Related Biases: A Review.” *Health Technology Assessment* 4 (10): 1–115.
- Stanley, T. D. 2005. “Beyond Publication Bias.” *Journal of Economic Surveys* 19 (3): 309–345.
- Sterling, Theodore D. 1959. “Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa.” *Journal of the American Statistical Association* 54 (285): 30–34.
- Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. “Moving to a World Beyond ‘ $p < 0.05$ ’” *American Statistician* 73 (S1): 1–19.
- Zheng, Yuqing, and Harry M. Kaiser. 2016. “Submission Demand In Core Economics Journals: A Panel Study.” *Economic Inquiry* 54 (2): 1319–1338.

Appendix A. Proofs

A.1. Proof of proposition 1

After conducting $n - 1$ studies, if $X_{n-1} < z$, the researcher chooses to conduct the next study if and only if (i) the expected direct profit from the next study is positive or (ii) the expected direct profit from the next study is negative but the expected direct profit from later studies offset the expected loss from conducting the next study. That is, if and only if (i)

$$v \mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*) - c(n) \geq 0$$

or (ii)

$$v \mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*) - c(n) < 0$$

and

$$\sum_{j \in \mathcal{J}} \{v \mathbb{P}(|X_{n+j}^*| > z | \mathbf{X}_{n-1}^*) - c(n+j)\} \geq 0$$

for some $\mathcal{J} \subset \{1, 2, 3, \dots\}$. However, since $v \mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*) - c(n) \geq v \mathbb{P}(|X_{n+j}^*| > z | \mathbf{X}_{n-1}^*) - c(n+j)$ for all $j > 0$, if $v \mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*) - c(n) < 0$, then $v \mathbb{P}(|X_{n+j}^*| > z | \mathbf{X}_{n-1}^*) - c(n+j) < 0$ for all $j > 0$ so $\sum_{j \in \mathcal{J}} \{v \mathbb{P}(|X_{n+j}^*| > z | \mathbf{X}_{n-1}^*) - c(n+j)\} < 0$ —making (ii) impossible. The result of the proposition immediately follows.

A.2. Proof of proposition 2

To show (i), note that for any $v, c(1)$, there exists $\bar{z} > 0$ such that $\mathbb{P}(|X_n^*| > z | \mathbf{X}_{n-1}^*) < c(1)/v$ for all $z \geq \bar{z}$. Thus by proposition 1, $N(z) = 0$ for all $z \geq \bar{z}$ and therefore $P_{H_0}(X_{N(z)} > z) = P_{H_0}(0 > z) = 0$ for all $z \geq \bar{z}$.

To show (ii), note that the test using z^* yields non-zero power if and only if $P_{H_1}(X_{\tilde{N}(z^*)} \geq z^*) > 0$, where P_{H_1} denotes the objective probability measure for some true value of θ under the alternative hypothesis. Since X_n is a continuous random variable with support on the entire positive real line for all $n > 0$ under any value true of θ consistent H_1 (as well as H_0), this implies that the test using z^* yields non-zero power if and only if $N(z^*) > 0$. Finally, $N(z^*) > 0$ if and only if $\mathbb{P}(|X_1^*| > z^*) \geq c(1)/v$.

A.3. Proof of proposition 4

For each case, $N(z)$ is equal to the largest positive integer n such that $X_{n-1} < z$ and (8) holds. In case (i),

$$\int [\Phi(z - \theta) - \Phi(-z - \theta)] dF(\theta) = \Phi(z - \tilde{\theta}) - \Phi(-z - \tilde{\theta})$$

so (8) is equal to the condition given in (i). In case (ii),

$$\begin{aligned} \int [\Phi(z - \theta) - \Phi(-z - \theta)] dF(\theta) &= (b - a)^{-1} \int_a^b [\Phi(z - \theta) - \Phi(-z - \theta)] d\theta \\ &= (b - a)^{-1} [\phi(z - b) - \phi(z - a) + (b - z)\Phi(z - b) + (z - a)\Phi(z - a) \\ &\quad - \phi(-z - b) + \phi(-z - a) - (b + z)\Phi(-z - b) + (a + z)\Phi(-z - a)] \end{aligned}$$

so (8) is equal to the condition given in (ii). In case (iii),

$$\begin{aligned} \int [\Phi(z - \theta) - \Phi(-z - \theta)] dF(\theta) &= \sigma^{-1} \int_{-\infty}^{\infty} [\Phi(z - \theta) - \Phi(-z - \theta)] \phi((\theta - \mu)/\sigma) d\theta \\ &= \int_{-\infty}^{\infty} [\Phi(z - \mu - \sigma x) - \Phi(-z - \mu - \sigma x)] \phi(x) dx \\ &= \Phi((z - \mu)/\sqrt{1 + \sigma^2}) - \Phi((-z - \mu)/\sqrt{1 + \sigma^2}) \end{aligned}$$

so (8) is equal to the condition given in (iii).

A.4. Proof of proposition 5

Standard posterior contraction results imply that

$$\int [1 - \Phi(z - \theta) + \Phi(-z - \theta)] dF(\theta | X_{n-1}^*) \xrightarrow{P} 1 - \Phi(z - \theta^*) + \Phi(-z - \theta^*)$$

as $n \rightarrow \infty$, so the researcher must eventually stop conducting studies even if he never attains rejection if $1 - \Phi(z - \theta^*) + \Phi(-z - \theta^*) < c/v$. Since Φ is strictly increasing, $\lim_{x \rightarrow -\infty} \Phi(x) = 0$ and $\lim_{x \rightarrow \infty} \Phi(x) = 1$, this implies the statement of the proposition.

A.5. Proof of proposition 6

Using our knowledge of the distribution of X_n^* , we have

$$f(X_n^* | \theta) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2} \sum_{i=1}^n (X_i^* - \theta)^2\right)$$

$$f(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\theta - \mu)^2\right)$$

so that

$$\begin{aligned} \int f(X_n^* | \theta) f(\theta) d\theta &= \frac{1}{\sqrt{(2\pi)^{n+1}\sigma^2}} \int \exp\left(-\frac{1}{2} \left[\sum_{i=1}^n (\theta - X_i^*)^2 + \frac{1}{\sigma^2}(\theta - \mu)^2 \right]\right) d\theta \\ &= \frac{1}{\sqrt{(2\pi)^{n+1}\sigma^2}} \int \exp\left(-\frac{1}{2} \left(n + \frac{1}{\sigma^2}\right) (\theta - h)^2\right) d\theta \exp(k) \\ &= \frac{\sqrt{\left(n + \frac{1}{\sigma^2}\right)^{-1}}}{\sqrt{(2\pi)^n \sigma^2}} \int \frac{1}{\sqrt{2\pi \left(n + \frac{1}{\sigma^2}\right)^{-1}}} \exp\left(-\frac{1}{2} \left(n + \frac{1}{\sigma^2}\right) (\theta - h)^2\right) d\theta \exp(k) \\ &= \frac{1}{\sqrt{(2\pi)^n (n\sigma^2 + 1)}} \exp(k), \end{aligned}$$

where

$$h = \frac{\sigma^2 \sum_{i=1}^n X_i^* + \mu}{n\sigma^2 + 1},$$

$$k = -\frac{1}{2} \left(\sum_{i=1}^n X_i^{*2} + \frac{\mu^2}{\sigma^2} \right) + \frac{\left(\sum_{i=1}^n X_i^* + \frac{\mu}{\sigma^2} \right)^2}{2 \left(n + \frac{1}{\sigma^2} \right)}.$$

Thus,

$$\begin{aligned} f(\theta | \mathbf{X}_n^*) &= \frac{f(\mathbf{X}_n^* | \theta) f(\theta)}{\int f(\mathbf{X}_n^* | \theta) f(\theta) d\theta} \\ &= \frac{\sqrt{(2\pi)^n (n\sigma^2 + 1)}}{\sqrt{(2\pi)^{n+1} \sigma^2}} \exp \left(-\frac{1}{2} \left[\sum_{i=1}^n (\theta - X_i^*)^2 + \frac{1}{\sigma^2} (\theta - \mu)^2 + \frac{\left(\sum_{i=1}^n X_i^* + \frac{\mu}{\sigma^2} \right)^2}{n + \frac{1}{\sigma^2}} - \left(\sum_{i=1}^n X_i^{*2} + \frac{\mu^2}{\sigma^2} \right) \right] \right) \\ &= \frac{1}{\sqrt{2\pi \sigma_{n+1|n}^2}} \exp \left(-\frac{(\theta - \mu_{n+1|n})^2}{2\sigma_{n+1|n}^2} \right). \end{aligned}$$

This means that $\theta | \mathbf{X}_n^*$ has a normal distribution with mean $\mu_{n+1|n}$ and variance $\sigma_{n+1|n}^2$. The statement of the proposition then follows from analogous arguments to those made in the proof of proposition 4(iii).

A.6. Proof of proposition 7

If H_0 holds so $\theta^* = 0$,

$$P_{H_0} \left(1 - \Phi \left(\sqrt{nz} - \theta - \sqrt{n-1} X_{n-1}^* \right) + \Phi \left(-\sqrt{nz} - \theta - \sqrt{n-1} X_{n-1}^* \right) < \frac{c}{v} \right)$$

is positive for all n and any finite θ and $c \leq v$. This implies that for any standard prior distribution $F(\theta)$, if H_0 is true, the probability that the researcher stops conducting additional studies is positive at every stage. Moreover,

$$\begin{aligned} &P_{H_0} \left(1 - \Phi \left(\sqrt{nz} - \theta - \sqrt{n-1} X_{n-1}^* \right) + \Phi \left(-\sqrt{nz} - \theta - \sqrt{n-1} X_{n-1}^* \right) < \frac{c}{v} \right) \\ &\geq P_{H_0} \left(1 - \Phi \left(\sqrt{nz} - \theta - \sqrt{n-1} X_{n-1}^* \right) < \frac{c}{2v}, \Phi \left(-\sqrt{nz} - \theta - \sqrt{n-1} X_{n-1}^* \right) < \frac{c}{2v} \right) \\ &= P_{H_0} \left(\frac{-\sqrt{nz} - \theta - \Phi^{-1}(c/2v)}{\sqrt{n-1}} < X_{n-1}^* < \frac{\sqrt{nz} - \theta - \Phi^{-1}(1-c/2v)}{\sqrt{n-1}} \right) \\ &= \Phi \left(\frac{\sqrt{nz} - \theta - \Phi^{-1}(1-c/2v)}{\sqrt{n-1}} \right) - \Phi \left(\frac{-\sqrt{nz} - \theta - \Phi^{-1}(c/2v)}{\sqrt{n-1}} \right) \rightarrow \Phi(z) - \Phi(-z) > 0 \end{aligned}$$

as $n \rightarrow \infty$ so even as the researcher accumulates more data, the probability that the researcher continues to conduct more studies remains bounded away from one. Thus, the statement of the

proposition holds.

A.7. Proof of proposition 8

If the researcher's prior distribution $F(\boldsymbol{\theta}_n)$ admits a pdf $f(\boldsymbol{\theta}_n)$, for $n \geq 2$, the posterior pdf for $\boldsymbol{\theta}_n$ takes the form

$$(A1) \quad f(\boldsymbol{\theta}_n | \mathbf{X}_{n-1}^*) = \frac{f(\mathbf{X}_{n-1}^* | \boldsymbol{\theta}_n) f(\boldsymbol{\theta}_n)}{\int f(\mathbf{X}_{n-1}^* | \boldsymbol{\theta}_n) f(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n},$$

where the likelihood $f(\mathbf{X}_{n-1}^* | \boldsymbol{\theta}_n)$ is the pdf of a $\mathcal{N}(\boldsymbol{\theta}_{n-1}, \boldsymbol{\Omega}_{n-1})$ distribution evaluated at \mathbf{X}_{n-1}^* . As a notational convention, for $n = 1$, $f(\boldsymbol{\theta}_n | \mathbf{X}_{n-1}^*)$ is equal to $f(\boldsymbol{\theta}_n)$. Given that the researcher's prior on $\boldsymbol{\theta}_n$ follows a $\mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ distribution for all $n \geq 1$,

$$(A2) \quad f(\mathbf{X}_{n-1}^* | \boldsymbol{\theta}_n) = \frac{1}{\sqrt{(2\pi)^{n-1} |\boldsymbol{\Omega}_{n-1}|}} \exp\left(-\frac{1}{2}(\mathbf{X}_{n-1}^* - \boldsymbol{\theta}_{n-1})' \boldsymbol{\Omega}_{n-1}^{-1} (\mathbf{X}_{n-1}^* - \boldsymbol{\theta}_{n-1})\right)$$

$$(A3) \quad f(\boldsymbol{\theta}_n) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_n|}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_n - \boldsymbol{\mu}_n)' \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\theta}_n - \boldsymbol{\mu}_n)\right)$$

so for

$$\begin{aligned} \tilde{\boldsymbol{\Omega}}_n^{-1} &= \begin{bmatrix} \boldsymbol{\Omega}_{n-1}^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \\ \mathbf{A} &= -\frac{1}{2} \left(\tilde{\boldsymbol{\Omega}}_n^{-1} + \boldsymbol{\Sigma}_n^{-1} \right), \\ \mathbf{b} &= \tilde{\boldsymbol{\Omega}}_n^{-1} \mathbf{X}_n^* + \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_n, \\ \mathbf{c} &= -\frac{1}{2} \left(\mathbf{X}_n^{*'} \tilde{\boldsymbol{\Omega}}_n^{-1} \mathbf{X}_n^* + \boldsymbol{\mu}_n' \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_n \right), \end{aligned}$$

$h = -\mathbf{A}^{-1} \mathbf{b} / 2$ and $k = \mathbf{c} - \mathbf{b}' \mathbf{A}^{-1} \mathbf{b} / 4$, we have

$$\begin{aligned} & \int f(\mathbf{X}_{n-1}^* | \boldsymbol{\theta}_n) f(\boldsymbol{\theta}_n) d\boldsymbol{\theta}_n \\ &= \frac{1}{\sqrt{(2\pi)^{2n-1} |\boldsymbol{\Omega}_{n-1}| |\boldsymbol{\Sigma}_n|}} \\ & \times \int \exp\left(-\frac{1}{2} \left[(\boldsymbol{\theta}_n - \mathbf{X}_n^*)' \tilde{\boldsymbol{\Omega}}_n^{-1} (\boldsymbol{\theta}_n - \mathbf{X}_n^*) + (\boldsymbol{\theta}_n - \boldsymbol{\mu}_n)' \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\theta}_n - \boldsymbol{\mu}_n) \right]\right) d\boldsymbol{\theta}_n \\ &= \frac{1}{\sqrt{(2\pi)^{2n-1} |\boldsymbol{\Omega}_{n-1}| |\boldsymbol{\Sigma}_n|}} \int \exp([\boldsymbol{\theta}_n - h]' \mathbf{A} [\boldsymbol{\theta}_n - h]) d\boldsymbol{\theta}_n \exp(k) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\sqrt{(2\pi)^{n-1}2|A||\boldsymbol{\Omega}_{n-1}||\boldsymbol{\Sigma}_n|}} \int \frac{1}{\sqrt{(2\pi)^{n\frac{1}{2}}|A^{-1}|}} \exp([\boldsymbol{\theta}_n - h]'A[\boldsymbol{\theta}_n - h])d\boldsymbol{\theta}_n \exp(k) \\
&= \frac{1}{\sqrt{(2\pi)^{n-1}2|A||\boldsymbol{\Omega}_{n-1}||\boldsymbol{\Sigma}_n|}} \exp(k) \\
&= \frac{1}{\sqrt{(2\pi)^{n-1}|\tilde{\boldsymbol{\Omega}}_n^{-1}\boldsymbol{\Sigma}_n + \mathbf{I}_n||\boldsymbol{\Omega}_{n-1}|}} \exp(k).
\end{aligned}$$

Thus, (A1)–(A3) imply

$$\begin{aligned}
f(\boldsymbol{\theta}_n | \mathbf{X}_{n-1}^*) &= \frac{\sqrt{(2\pi)^{n-1}|\tilde{\boldsymbol{\Omega}}_n^{-1}\boldsymbol{\Sigma}_n + \mathbf{I}_n||\boldsymbol{\Omega}_{n-1}|}}{\sqrt{(2\pi)^{n-1}|\boldsymbol{\Omega}_{n-1}|(2\pi)^n|\boldsymbol{\Sigma}_n|}} \exp(-(1/2)(\mathbf{X}_{n-1}^* - \boldsymbol{\theta}_{n-1})'\boldsymbol{\Omega}_{n-1}^{-1}(\mathbf{X}_{n-1}^* - \boldsymbol{\theta}_{n-1}) \\
&\quad - (1/2)(\boldsymbol{\theta}_n - \boldsymbol{\mu}_n)'\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\theta}_n - \boldsymbol{\mu}_n) - k) \\
&= \frac{\sqrt{|\tilde{\boldsymbol{\Omega}}_n^{-1} + \boldsymbol{\Sigma}_n^{-1}|}}{\sqrt{(2\pi)^n}} \exp(-(1/2)\boldsymbol{\theta}_n'[\tilde{\boldsymbol{\Omega}}_n^{-1} + \boldsymbol{\Sigma}_n^{-1}]\boldsymbol{\theta}_n + \boldsymbol{\theta}_n'[\tilde{\boldsymbol{\Omega}}_n^{-1}\mathbf{X}_n^* + \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n] \\
&\quad + (1/2)[\mathbf{X}_n^*\tilde{\boldsymbol{\Omega}}_n^{-1} + \boldsymbol{\mu}_n'\boldsymbol{\Sigma}_n^{-1}][\tilde{\boldsymbol{\Omega}}_n^{-1} + \boldsymbol{\Sigma}_n^{-1}]^{-1}[\tilde{\boldsymbol{\Omega}}_n^{-1}\mathbf{X}_n^* + \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n]) \\
&= \frac{1}{\sqrt{(2\pi)^n|[\tilde{\boldsymbol{\Omega}}_n^{-1} + \boldsymbol{\Sigma}_n^{-1}]^{-1}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta}_n - \boldsymbol{\mu}_{n|n-1})'[\tilde{\boldsymbol{\Omega}}_n^{-1} + \boldsymbol{\Sigma}_n^{-1}](\boldsymbol{\theta}_n - \boldsymbol{\mu}_{n|n-1})\right).
\end{aligned}$$

This implies that the statement of the proposition holds.

Appendix B. One-sided hypothesis testing

The main text focuses on two-sided hypothesis tests. This appendix presents analogous results for upper one-sided hypotheses

$$(A4) \quad H_0 : \beta = \beta_0 \quad \text{versus} \quad H_1 : \beta > \beta_0.$$

This analysis also covers lower one-sided hypotheses for which the alternative is instead $H_1 : \beta < \beta_0$ by simply redefining the parameter of interest β to be equal to its negative value. We state results briefly, omitting discussions and derivations since they are analogous to those for the two-sided problem.

For the one-sided hypotheses (A4), redefine $X_n = \max\{X_1^*, \dots, X_n^*\}$. Replace (3) with

$$\begin{aligned}
\mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid \mathbf{X}_{n-1}^*) &= v \mathbb{P}(X_n > z \mid \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n) \\
&= v \mathbb{P}(\max\{X_n^*, X_{n-1}\} > z \mid \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n)
\end{aligned}$$

$$= v \mathbb{P}(X_n^* > z \mid \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n).$$

Proposition 1 should be modified to the following.

PROPOSITION A1. *Suppose $v \mathbb{P}(X_n^* > z \mid \mathbf{X}_{n-1}^*) - c(n) \geq v \mathbb{P}(X_{n+j}^* > z \mid \mathbf{X}_{n-1}^*) - c(n+j)$ for all $j > 0$. Then the researcher chooses to conduct the next study if and only if $X_{n-1} < z$ and*

$$v \mathbb{P}(X_n^* > z \mid \mathbf{X}_{n-1}^*) - c(n) \geq 0.$$

Proposition 2 should be modified to the following.

PROPOSITION A2. *Suppose $v \mathbb{P}(X_n^* > z \mid \mathbf{X}_{n-1}^*) - c(n) \geq v \mathbb{P}(X_{n+j}^* > z \mid \mathbf{X}_{n-1}^*) - c(n+j)$ for all $j > 0$ and $c(1) > 0$. The ICCV z^* defined in (4), (i) exists and (ii) yields a test with non-zero power if and only if $\mathbb{P}(X_1^* > z^*) \geq c(1)/v$.*

Replace display (5) with

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta > 0,$$

display (6) with

$$\begin{aligned} \mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid \mathbf{X}_{n-1}^*) &= v \int \mathbb{P}(X_n^* > z \mid \mathbf{X}_{n-1}^*, \theta) dF(\theta) \mathbb{1}(X_{n-1} < z) - c \\ &= v \int [1 - \Phi(z - \theta)] dF(\theta) \mathbb{1}(X_{n-1} < z) - c, \end{aligned}$$

display (7) with

$$\int [1 - \Phi(z - \theta)] dF(\theta) \geq \frac{c}{v},$$

display (8) with

$$\int [1 - \Phi(z - \theta)] dF(\theta) \geq \frac{c(n)}{v},$$

and display (9) with

$$\int [1 - \Phi(z - \theta)] dF(\theta) < \frac{c(1)}{v}.$$

The statement immediately following proposition 3 should be replaced by the following: “Since X_n is the maximum of n mean-zero iid normal variables under H_0 , $P_{H_0}(X_n < z) = \Phi(z)^n$ for $z \geq 0$ so we can rewrite the (implicit) definition of the ICCV in this context as the smallest value z^* such that

$$1 - \Phi(z^*)^{N(z^*)} \leq \alpha.”$$

Proposition 4 should be modified to the following:

PROPOSITION A3. *In the increasing cost setting of this section, $N(z)$ is equal to the maximum positive integer n for which $X_{n-1} < z$ and*

1. *if $dF(\theta) = \delta(\tilde{\theta})d\theta$ for some $\tilde{\theta} \in \mathbb{R}$, where δ denotes the Dirac delta function,*

$$n \leq c^{-1} \left(v[1 - \Phi(z - \tilde{\theta})] \right);$$

2. *if $dF(\theta) = (b - a)^{-1} \mathbb{1}(a \leq \theta \leq b)d\theta$ for some $a < b$,*

$$n \leq c^{-1} \left(v[1 - (b - a)^{-1} \{ \phi(z - b) - \phi(z - a) + (b - z)\Phi(z - b) + (z - a)\Phi(z - a) \}] \right);$$

3. *if $dF(\theta) = \sigma^{-1} \phi((\theta - \mu)/\sigma)d\theta$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$,*

$$n \leq c^{-1} \left(v[1 - \Phi((z - \mu)/\sqrt{1 + \sigma^2})] \right).$$

Replace display (11) with

$$(A5) \quad \int [1 - \Phi(z - \theta)] dF(\theta | X_{n-1}^*) \geq \frac{c}{v}.$$

In the paragraph preceding proposition 5, $|X_{n-1}^*|$ should be replaced with X_{n-1}^* . Proposition 5 should be modified to the following:

PROPOSITION A4. *Let θ^* denote the true value of θ . For any CV z large enough such that $1 - \Phi(z - \theta^*) < c/v$, the researcher will eventually stop conducting studies even if he never attains rejection.*

Proposition 6 should be modified to the following:

PROPOSITION A5. *In the learning while sampling setting of this section, if $dF(\theta) = \sigma^{-1} \phi((\theta - \mu)/\sigma)d\theta$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$, then $N(z)$ is the largest positive integer n for which $X_{n-1} < z$ and*

$$1 - \Phi \left(\frac{z - \mu_{n|n-1}}{\sqrt{1 + \sigma_{n|n-1}^2}} \right) \geq c/v,$$

where

$$\mu_{n|n-1} = \frac{\sigma^2 \sum_{i=1}^{n-1} X_i^* + \mu}{(n-1)\sigma^2 + 1} \quad \text{and} \quad \sigma_{n|n-1}^2 = \frac{\sigma^2}{(n-1)\sigma^2 + 1}.$$

In remark 9, “ $1 - \Phi(z - \theta^*) + \Phi(-z - \theta^*) < c/v$ ” should be replaced by “ $1 - \Phi(z - \theta^*) < c/v$ ”.
 Replace display (12) with

$$\begin{aligned}\mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid \mathbf{X}_{n-1}^*) &= v \int \mathbb{P}(X_n^* > z \mid \mathbf{X}_{n-1}^*, \theta) dF(\theta) \mathbb{1}(X_{n-1} < z) - c \\ &= v \int \left[1 - \Phi\left(\sqrt{nz} - \theta - \sqrt{n-1}X_{n-1}^*\right) \right] dF(\theta) \mathbb{1}(X_{n-1} < z) - c,\end{aligned}$$

display (13) with

$$\int \left[1 - \Phi\left(\sqrt{nz} - \theta - \sqrt{n-1}X_{n-1}^*\right) \right] dF(\theta) \geq \frac{c}{v},$$

display (14) with

$$H_0 : \beta_i = \beta_{i0} \quad \text{versus} \quad H_1 : \beta_i > \beta_{i0},$$

display (15) with

$$\begin{aligned}\mathbb{E}(\pi(X_n, X_{n-1}; z, v, c) \mid \mathbf{X}_{n-1}^*) &= v \int \mathbb{P}(X_{n|n-1}^* > z \mid \boldsymbol{\theta}_n) dF(\boldsymbol{\theta}_n; \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n) \\ &= v \int \left[1 - \Phi\left((z - \theta_{n|n-1})/\omega_{n|n-1}\right) \right] dF(\boldsymbol{\theta}_n; \mathbf{X}_{n-1}^*) \mathbb{1}(X_{n-1} < z) - c(n)\end{aligned}$$

and display (16) with

$$\int \left[1 - \Phi\left((z - \theta_{n|n-1})/\omega_{n|n-1}\right) \right] dF(\boldsymbol{\theta}_n; \mathbf{X}_{n-1}^*) \geq \frac{c(n)}{v}.$$

Appendix C. Calibration of researcher’s prior

As mentioned in section 3, the two previous articles to which we calibrate Dal Bo’s prior, Roth and Murnighan (1978) and Murnighan and Roth (1983), do not provide standard error or t -statistic values, but only estimates of β . Nevertheless, the structure of the matched pairs experimental design in each of these two papers allows us to compute bounds for the estimated standard error for $\hat{\beta}$ as follows. Roth and Murnighan (1978) and Murnighan and Roth (1983) report the number of subjects in their experiments who choose to cooperate under high and low probabilities of future interaction and compare the arithmetic averages of these two numbers to assess whether there is statistical evidence that the probability of cooperation is different under high vs low levels of future interaction. Between these two papers, the authors conduct four total experiments.

Formally, in each experiment, the authors report the sample size n , $\sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i$, where $y_i = 1$ if individual i chooses to cooperate under low probability of future interaction and $y_i = 0$ otherwise, and $x_i = 1$ if individual i chooses to cooperate under high probability of future

interaction and $x_i = 0$ otherwise. They then estimate β as the difference in the sample averages of these values between the same set of individuals: $\hat{\beta} = n^{-1} \sum_{i=1}^n x_i - n^{-1} \sum_{i=1}^n y_i = n^{-1} \sum_{i=1}^n (x_i - y_i)$. Though the authors report that $\hat{\beta}$ is statistically different from zero in each experiment, they neither provide t -statistics nor standard errors. However, in a matched pairs design, the standard error is calculated as $\text{se}(\hat{\beta}) = n^{-1/2} \widehat{\text{sd}}(x_i - y_i)$, where

$$\widehat{\text{sd}}(x_i - y_i) = \sqrt{\frac{\sum_{i=1}^n (x_i - y_i - \hat{\beta})^2}{n - 1}}$$

is a consistent estimator of $\text{sd}(x_i - y_i)$. Since

$$\sum_{i=1}^n (x_i - y_i - \hat{\beta})^2 = \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i y_i + \sum_{i=1}^n y_i^2 - n \hat{\beta}^2 = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i - n \hat{\beta}^2 - 2 \sum_{i=1}^n x_i y_i$$

and

$$0 \leq \sum_{i=1}^n x_i y_i \leq \min \left\{ \sum_{i=1}^n x_i, \sum_{i=1}^n y_i \right\},$$

we can bound $\widehat{\text{sd}}(x_i - y_i)$ as follows: $\widehat{\text{sd}}_{lb} \leq \widehat{\text{sd}}(x_i - y_i) \leq \widehat{\text{sd}}_{ub}$, where

$$\widehat{\text{sd}}_{lb} = \sqrt{\frac{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i - n \hat{\beta}^2 - 2 \min \left\{ \sum_{i=1}^n x_i, \sum_{i=1}^n y_i \right\}}{n - 1}}, \quad \widehat{\text{sd}}_{ub} = \sqrt{\frac{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i - n \hat{\beta}^2}{n - 1}}$$

can be computed directly from the values reported in Roth and Murnighan (1978) and Murnighan and Roth (1983). To calibrate Dal Bo's prior, we simply took the midpoint of these two bounds as a reasonable value for $\widehat{\text{sd}}(x_i - y_i)$ that the researcher would infer from these previous studies when forming his prior on the mean of the t -statistic $\beta/\text{sd}(\hat{\beta})$. More specifically, we calibrated the researcher's prior for $\beta/\text{sd}(\hat{\beta})$ as follows:

1. Calculate $\hat{\sigma}_j(x_i - y_i) = 0.5 \widehat{\text{sd}}_{lb,j} + 0.5 \widehat{\text{sd}}_{ub,j}$, where $\widehat{\text{sd}}_{lb,j}$ and $\widehat{\text{sd}}_{ub,j}$ are equal to the values of $\widehat{\text{sd}}_{lb}$ and $\widehat{\text{sd}}_{ub}$ in the j^{th} study for each of the $j = 1, \dots, 4$ studies in Roth and Murnighan (1978) and Murnighan and Roth (1983).
2. Calculate $t_j = \sqrt{n_j} \hat{\beta}_j / \hat{\sigma}_j(x_i - y_i)$, where n_j is the sample size and $\hat{\beta}_j$ is the estimate of β in study j . This serves as an approximation to the t -statistic obtain in each of the $j = 1, \dots, 4$ studies in Roth and Murnighan (1978) and Murnighan and Roth (1983).
3. Calculate the mean of the researcher's prior for $\beta/\text{sd}(\hat{\beta})$ as $\mathbb{E}(\beta/\text{sd}(\hat{\beta})) = \sum_{j=1}^4 w_j \sqrt{48.75/n_j} t_j$, where the weights $w_j = n_j / \sum_{i=1}^4 n_i$ correspond to the relative sample size of study j . The

TABLE A1. Marginal cost-benefit ratio bounds

\bar{n}	1	2	3	4	5	6	7
$E_{H_0}[\mathbb{P}(X_{\bar{n}+1}^* > z \mid \mathbf{X}_{\bar{n}}^*)]$	0.366	0.266	0.212	0.180	0.158	0.142	0.130
$E_{H_0}[\mathbb{P}(X_{\bar{n}}^* > z \mid \mathbf{X}_{\bar{n}-1}^*)]$	0.596	0.366	0.266	0.212	0.180	0.158	0.142

scaling by $\sqrt{48.75/n_j}$ inside of the sum is used to account for the fact that Dal Bo's average sample size is 48.75, rather than n_j .

4. Calculate the variance of the researcher's prior for $\beta/\text{sd}(\hat{\beta})$ as

$$\text{var}(\beta/\text{sd}(\hat{\beta})) = \sum_{j=1}^4 w_j \left(\sqrt{48.75/n_j t_j} - \sum_{i=1}^4 w_i \sqrt{48.75/n_i t_i} \right)^2.$$

Appendix D. Indirect elicitation of cost-benefit ratio

Suppose we knew the maximum number of studies \bar{n} a researcher would typically conduct under H_0 without rejecting H_0 at a given CV z before stopping and moving on to a different research project. Using proposition 1, we could then infer $vE_{H_0}[\mathbb{P}(|X_{\bar{n}}^*| > z \mid \mathbf{X}_{\bar{n}-1}^*)] - c(\bar{n}) \geq 0$ and $vE_{H_0}[\mathbb{P}(|X_{\bar{n}+1}^*| > z \mid \mathbf{X}_{\bar{n}}^*)] - c(\bar{n} + 1) < 0$, where the expectations are taken over the distribution of $\mathbf{X}_{\bar{n}-1}^*$ and $\mathbf{X}_{\bar{n}}^*$ under the objective probability measure when H_0 is true. If marginal costs are constant, we could then bound the marginal cost-to-payoff ratio:

$$E_{H_0}[\mathbb{P}(|X_{\bar{n}+1}^*| > z \mid \mathbf{X}_{\bar{n}}^*)] < \frac{c}{v} \leq E_{H_0}[\mathbb{P}(|X_{\bar{n}}^*| > z \mid \mathbf{X}_{\bar{n}-1}^*)].$$

Moreover, if we were able to obtain this maximum number of studies across a range of CVs z , so \bar{n} is a function of z , we could tighten these bounds:

$$\sup_z E_{H_0} \left[\mathbb{P} \left(|X_{\bar{n}(z)+1}^*| > z \mid \mathbf{X}_{\bar{n}(z)}^* \right) \right] < \frac{c}{v} \leq \inf_z E_{H_0} \left[\mathbb{P} \left(|X_{\bar{n}(z)}^*| > z \mid \mathbf{X}_{\bar{n}(z)-1}^* \right) \right].$$

In principle, this could be achieved by surveying researchers, although a mechanism eliciting truthful reporting would be crucial to such an analysis.

As a practical illustration, consider the pooling data paradigm of section 7 and setting the prior of the researcher according to the calibration in section 3. In this case since $\sqrt{\bar{n}-1}X_{\bar{n}-1}/\sqrt{1+\sigma^2} \sim$

$\mathcal{N}(0, (\bar{n} - 1)/(1 + \sigma^2))$ under H_0 , we obtain

$$\begin{aligned} E_{H_0}[\mathbb{P}(|X_{\bar{n}+1}^*| > z \mid \mathbf{X}_{\bar{n}}^*)] &= E_{H_0} \left[1 - \Phi \left(\frac{\sqrt{\bar{n} + 1}z - \sqrt{\bar{n}}X_{\bar{n}}^* - \mu}{\sqrt{1 + \sigma^2}} \right) + \Phi \left(\frac{-\sqrt{\bar{n} + 1}z - \sqrt{\bar{n}}X_{\bar{n}}^* - \mu}{\sqrt{1 + \sigma^2}} \right) \right] \\ &= 1 - \Phi \left(\frac{\sqrt{\bar{n} + 1}z - \mu}{\sqrt{1 + \sigma^2 + \bar{n}}} \right) + \Phi \left(\frac{-\sqrt{\bar{n} + 1}z - \mu}{\sqrt{1 + \sigma^2 + \bar{n}}} \right), \end{aligned}$$

where μ and σ^2 are the mean and variance of the researcher's prior distribution on $\beta/\text{sd}(\hat{\beta})$.

Similarly,

$$E_{H_0}[\mathbb{P}(|X_{\bar{n}}^*| > z \mid \mathbf{X}_{\bar{n}-1}^*)] = 1 - \Phi \left(\frac{\sqrt{\bar{n}}z - \mu}{\sqrt{\sigma^2 + \bar{n}}} \right) + \Phi \left(\frac{-\sqrt{\bar{n}}z - \mu}{\sqrt{\sigma^2 + \bar{n}}} \right).$$

Table A1 records these bounds for c/v corresponding to different values of \bar{n} for the standard CV of $z = 1.96$. Dal Bo (2005) conducted four studies. If we assume that this is the number of studies he would typically conduct under a true H_0 (so $\bar{n} = 4$), we recover bounds consistent with the calibrated value of $c/v = 933/5000 = 0.187$ from table 1.